Proceedings of CARI 2016

Moussa Lo, Eric Badouel, Nabil Gmati

Proceedings of CARI 2016

Actes du CARI 2016

Mammamet, Tunisia, October 2016

EDITORS : MOUSSA LO - ERIC BADOUEL - NABIL GMATI

AVANT-PROPOS

Le CARI, Colloque Africain sur la Recherche en Informatique, fruit d'une coopération internationale rassemblant universités africaines, centres de recherche français et organismes internationaux, tient sa treizième édition cette année en Tunisie Organisé tous les deux ans en Afrique, ses précédentes éditions se sont tenues à Yaoundé en 1992, à Ouagadougou en 1994, à Libreville en 1996, à Dakar en 1998, à Antananarivo en 2000, à Yaoundé en 2002, à Hammamet en 2004, à Cotonou en 2006, à Rabat en 2008, à Yamoussoukro en 2010, à Alger en 2012 et à Saint-Louis du Sénégal en 2014.

Le colloque est co-organisé par l'Institut National de Recherche en Informatique et en Automatique (Inria), l'Institut de Recherche pour le Développement (IRD), le Centre de coopération Internationale en Recherche Agronomique pour le Développement (Cirad), le Centre International des Mathématiques Pures et Appliquées (Cimpa), et l'Agence Universitaire de la Francophonie (AUF). Cette treizième édition, confiée à l'Ecole Nationale d'Ingénieur de Tunis (ENIT), sous la coordination du professeur Nabil Gmati , a bénéficié d'un large soutien des institutions universitaires suivantes : la Faculté des Sciences de Tunis (FST), l'Ecole Polytechnique de Tunisie (EPT), l'Ecole Nationale Supérieure d'Informatique (ENSI), l'Ecole Supérieure privée d'Ingénierie et de Technologies (ESPRIT) et de l'Ecole Supérieure des Télécommunications (Sup'Com). Quatre universités du grand Tunis se sont également associées pour la réussite de cette nouvelle édition du colloque : l'Université Tunis El Manar (UTM), l'Université de la Manouba (UM), l'Université de Carthage (UC), l'Université Virtuelle de Tunis (UVT). Ces universités regroupent l'essentiel des institutions scientifiques du grand Tunis. Le CARI a également bénéficié du soutient de l'Institut Français de Tunisie (IFT).

Le CARI est devenu un lieu privilégié de rencontre et d'échanges de chercheurs et décideurs africains et internationaux de haut niveau dans les domaines de l'informatique et des mathématiques appliquées. Le programme scientifique, qui reflète la richesse et la diversité de la recherche menée sur le continent africain, met un accent particulier sur les travaux susceptibles de contribuer au développement technologique, à la connaissance de l'environnement et à la gestion des ressources naturelles. Ce programme se décline en 51 communications scientifiques, sélectionnées parmi 130 articles soumis, et des conférences invitées présentées par des spécialistes de renommée internationale.

L'Ecole de recherche Cimpa, qui a porté cette année sur sur les « mathématiques de la biologie », remplace désormais les traditionnels tutoriels organisés en marge du CARI.

Bien plus qu'un simple colloque, le CARI est un cadre dynamique de coopération, visant à rompre l'isolement et à renforcer la communauté scientifique africaine. Toute cette activité repose sur l'action forte et efficace de beaucoup d'acteurs. Nous remercions tous nos collègues qui ont marqué leur intérêt dans le CARI en y soumettant leurs travaux scientifiques, les relecteurs qui ont accepté d'évaluer ces contributions et les membres du Comité de programme qui ont opéré à la sélection des articles. L'ensemble des activités liées au CARI sont répertoriées sur le site officiel du CARI (http://www.cari-info.org/) maintenu par l'équipe du professeur Mokhtar Sellami de l'université d'Annaba. Laura Norcy, d'Inria, a apporté son soutien pour la coordination de cette manifestation. L'organisation du colloque a reposé sur le comité local d'organisation, mis en place par le professeur Nabil Gmati.

Que les différentes institutions, qui, par leur engagement financier et par la participation de leurs membres, apportent leur soutien, soient également remerciées, et, bien sûr, toutes les institutions précédemment citées, qui soutiennent le CARI au fil de ses éditions.

Pour les organisateurs

Moussa Lo, Président du CARI
Eric Badouel, Secrétaire du Comité permanent du CARI
Nabil Gmati, Organisateur du CARI 2016

FOREWORD

CARI, the African Conference on Research in Computer Science, outcome of an international cooperation involving African universities, French research institutes, and international organizations, introduces this year its thirteenth edition in Tunisia. Organized every two years in Africa, its preceding editions were held in Yaoundé in 1992, in Ouagadougou in 1994, Libreville in 1996, Dakar in 1998, Antananarivo in 2000, Yaoundé in 2002, Hammamet in 2004, Cotonou in 2006, Rabat in 2008, Yamoussoukro in 2010, Algers in 2012, and Saint-Louis du Senegal in 2014.

The conference is organized by Institut National de Recherche en Informatique et en Automatique (Inria), the Institut de Recherche pour le Développement (IRD), the Centre de coopération Internationale en Recherche Agronomique pour le Développement (Cirad), the International Center for Pure and Applied Mathematics (ICPAM) and the Agence Universitaire de la Francophonie (AUF). This thirteenth edition, entrusted to ENIT (Ecole Nationale d'Ingénieurs de Tunis), under the coordination of Professor Nabil Gmati, has profited from a generous support of the following academic institutions: *Faculté des Sciences de Tunis* (FST), *Ecole Polytechnique de Tunisie* (EPT), *Ecole Nationale Supérieure d'Informatique* (ENSI), *Ecole Supérieure privée d'Ingénierie et de Technologies* (ESPRIT) and *Ecole Supérieure des Télécommunications* (Sup'Com). Four universities of the greater Tunis, which are the main scientific institutions of Tunis, also joined for the success of CARI: University of Tunis El Manar (UTM), La Manouba University (UM), University of Carthage (UC), *Université Virtuelle de Tunis* (UVT).
CARI also acknowledge supports from the *Institut Français de Tunisie* (IFT).

CARI has evolved into an internationally recognized event in Computer Science and Applied Mathematics. The scientific program, which reflects the richness and the diversity of the research undertaken on the African continent with a special emphasis on works related to the development of new technologies, knowledge in environmental sciences and to the management of natural resources, consists of 51 scientific contributions, selected from 130 submissions, together with invited talks delivered by acknowledged specialists.

From now on, the tutorials that used to precede the CARI are replaced by a CIMPA-ICPAM Research school that puts a focus on some particular topic relevant to CARI's community. For this edition the research school was dedicated to the "Mathematics for Biology".

More than a scientific gathering, CARI is also a dynamic environment for cooperation that brings together African researchers with the end result to break the gap of isolation. The successes of such an initiative rely on the contribution of many actors. We wish first to thank our colleagues who showed their interest in CARI by submitting a paper, the referees who accepted to evaluate these contributions, and the members of the Program Committee who managed the selection of papers. This process rested on the CARI official site (http://www.cari-info.org/) maintained by the team of professor Mokhtar Sellami at the University of Annaba. Laura Norcy, from Inria, was involved in numerous activities for the coordination of the Event. The local organization has been handled by the local organization committee under the supervision of professor Nabil Gmati.

Thanks also for all the institutions that support and provide funding for CARI conferences and related activities, and all the institutions involved in the organization of the conference.

For the organizing committee

Moussa Lo, Chairman of CARI
Eric Badouel, Secretary of CARI Permanent Committee
Nabil Gmati, Chair of CARI 2016

Comité de Programme / Program Committee

Modélisation des systèmes complexes / Complex System Modelling

Arnaud GRIGNARD, *IRD/UPMC*, *France*
Bernard CAZELLES, *UMPC, France*
Christophe LETT, *IRD, France*
Rachid MCHICH, *ENCGT - Morocco*
Tri NGUYEN-HUU, *IRD, France*
Benjamin ROCHE, *IRD, France*

Signal, image et multimédia / Signal, Images and Multimedia

Ezzine ABDELHAk, ENSAT, Morocco
Olivier ALATA, *Univ. Saint-Etienne, France*
Tarik Boujiha, *Univ. ibn tofail-kénitra, Morocco*
Mohamed DAOUDI, *Telecom Lille/LIFL, France*
Denis HAMAD, *Univ. du Littoral Côte d'Opale, France*
Ahmed HAMMOUCH, *CNRST, Morocco*
Lloussaine MASMOUDI, *Univ. Mohammed V, Rabat*, *Morocco*
Ahmed MOUSSA, *LabTIC ENSAT, Morocco*
Rachid OULAD HAJ THAMI, *ENSIAS, Marocco*
Abderrahmane SBIHI, *ENSA Tanger, Morocco*
Raja Touahni, *Université Ibn Tofail, Morocco*
Lynda ZAOUI, *Oran University, Algeria*
Djemel ZIOU, Sherbrooke *University, Canada*

Calcul scientifique et parallélisme / Scientific Computing and Parallelism

Jocelyne ERHEL, *Inria, France*
El Mostafa DAOUDI, *Univ. Mohamed I, Oujda, Morocco*
Laurent DEBREU, *Inria, France*
Laura GRIGORI, *Inria, France*
Abdou GUERMOUCHE, *Univ. Bordeaux, France*
Pierre MANNEBACK, *University of Mons, Belgium*
Maher MOAKHER, *Ecole Nationale d'Ingénieurs de Tunis, Tunisia*
Yanik NGOKO, *Univ. Paris 13, France*
Boniface NKONGA, *University of Nice-Sophia-Antipolis*
Patrice QUINTON, *ENS Rennes, France*
Denis TRYSTRAM, *Grenoble Institute of Technologu, France*

Intelligence artificielle et environnements informatique pour l'apprentissage humain / Artificial Intelligence and Computer-based Collaborative Environmental

Meziane AIDER, *USTHB, Algeria*
Djamil AISSANI, *Univ. Bejala, Algeria*
Monique BARON, *Sorbonne UPMC, Lip6, France*
Mahieddine DJOUDI, *Poitiers University, France*
Abdellatif ENNAJI, *Univ. Rouen, France*
Mokhtar SELLAMI, *Annaba University, Algeria*
Hassina SERIDI, *Annaba University,, Algeria*
Christophe SIBERTIN-BLANC, *Irit, France*
Salvatore TABBONE, *Univ. de Lorraine, France*
Claude TANGHA, *FTIC, Yaoundé, Cameroon*

Applications avancées en Génie logiciel / Software engineering and advances applications

Pascal ANDRE, *LINA, University of Nantes, France*
Eric BADOUEL, *Inria Rennes, France*
Kamel BARKAOUI, *Cedric, CNAM, France*
François CHAROY, *LORIA Univ. Henri Poincaré, France*
Yliès FALCONE, *Inria Grenoble, France*
Christian FOTSING, *Institut Universitaire de la Côte, Cameroon*
Georges-Edouard KOUAMOU, *ENSP Yaoundé I University, Cameroon*
Damien ROBERT, *Inria Bordeaux, France*
William SHU, *University of Buea, Cameroon*
François VERNADAT, *LAAS-CNRS, Toulouse, France*

Extraction et Organisation des Connaissances / Knowledge Organization and Mining

Hugo ALATRISTA-SALAS, *Pontificia Universidad Catolica del Peru*
Jérôme AZE, *LIRMM, University of Montpellier 2, France*
Nicolas BECHET, *Université de Bretagne Sud*
Hacene BELHADEF, *University of Constantine, Algeria*
Béatrice BOUCHOU MARKHOFF, *Univ. François Rabelais de Tours, France*
Ibrahim BOUNHAS, *LISI, Carthage University, Tunisia*
Sandra BRINGAY, *LIRMM, University Paul Valéry, France*
Patrice BUCHE, *Supagro INRA, France*
Gaoussou CAMARA, *Univ. Alioune Diop, Bambey, Senegal*
Célia DA COSTA TEREIRA, *Univ. Nice Sophia Antipolis, France*
Cheikh Taliboula DIOP, *UGB, Saint-Louis, Senegal*
Bilel ELAYEB, *ENSI, Tunisia*
Dino IENCO, *IRSTEA, France*
Clement JONQUET, LIRMM, *University of Montpellier 2, France*
Eric KERGOSIEN, *University of Lille, France*
Philippe LEMOISSON, TETIS, *Cirad, France*
Moussa LO, *Univ. Gaston Berger, Senegal*
Cédric LOPEZ, *Viseo Research Center, France*
Isabelle MOUGENOT, Espace-Dev, University of *Montpellier 2, France*
Mathieu ROCHE, TETIS, *Cirad, France*
Fatiha SAIS, *LRI, Paris Sud University, France*
Hassan SANEIFAR, *Raja University, Iran*
Joël SOR, *Cirad, France*
Maguelonne TEISSEIRE, *TETIS, Irstea, France*

Systèmes distribués et réseaux / Distributed systems and networks

Soraya AIT CHELLOUCHE, *University of Rennes, France*
Olivier BARAIS, *University of Rennes, France*
Melhem EL HELOU, *Univ. Saint-Joseph de Beyrouth, Lebanon*
Davide FREY, *Inria Rennes, France*
Abdoulaye GAMATIE, *LIRMM, France*
Bamba GUEYE, *UCAD, Dakar, Senegal*
Jean-Claude HOCHON, *Airbus SAS, France*
Michel HURFIN, *Inria Rennes, France*
Marc IBRAHIM, *Saint-Joseph University, Lebanon*
Samer LAHOUD, *University of Rennes, France*
Maryline LAURENT, *Telecom Sud-Paris, France*
Moussa LO, *Univ. Gaston Berger, Senegal*
Pascal LORENZ, *Univ. Haute Alsace, France*
Stéphane MAAG, *Telecom Sud-Paris, France*
Ludovic ME, *Supélec Rennes, France*
Congduc PHAM, *University of Pau, France*

Pierre ROLIN, *Institut Mines-Télécom, France*
Abed Ellatif SAMHAT, *Lebanese University, Lebanon*
Ousmane THIARE, *UGB, Saint-Louis, Senegal*
César VIHO, *Univ. De Rennes 1, France*

Mathématique pour la biologie / Mathematical biology

Nahla ABDELLATIF, *ENSI-ENIT, Tunisia*
Julien ARINO, *University of Manitoba, Canada*
Abdon ATANGANA, *University of Free State, South Africa*
Mejdi AZAIEZ, *University of Bordeaux, France*
Fethi Bin Muhammad BELGACEM, *PAAET, Kuwait*
Monia BELLALOUNA, *ENSI, Tunisia*
Hend BEN AMEUR, *LAMSIN-ENIT and IPEST, Tunisia*
Faker BEN BELGACEM, *UTC Compiègne, France*
Slimane BEN MILED, *University Tunis el Manar, Tunisia*
Adel BLOUZA, *University of Rouen, France*
Fabien CAMPILLO, *Inria, France*
Nicolas CHAMPAGNAT, *Inria, France*
Nadia CHOULAIEB, *ENIT, Tunisia*
Jean CLAIRAMBAULT, *Inria Paris- LJLL, UMPC, France*
Yves DUMONT, *Cirad, France*
Radhouene FEKIH SALEM, *University of Monastir, Tunisia*
Jean-Frédéric GERBEAU, *Inria, France*
Nabil GMATI, *LAMSIN-ENIT, Tunisia*
Lamia GUELLOUZ, *ENIT, Tunisia*
Abderrahmane HABBAL, *University of Nice-Inria, France*
Ridha HAMBLI, *Polytech Orleans, France*
Nejla HARIGA-TATLI, *INAT-LAMSIN, Tunisia*
Yousri HENCHIRI, *University of Montpellier 2, France*
Abderrahman IGGIDR, *Inria, France*
Adil KHALIL, *University of Marrakech, Morocco*
Michel LANGLAIS, *University of Bordeaux, France*
Claude LOBRY, *University of Nice, France*
John MADDOCKS, *EPFL, Switzerland*
Sylvie MELEARD, *Ecole Plotechnique, France*
Ali MOUSSAOUI, *Tlemcen University, Algeria*
Tri NGUYEN-HU, *IRD, France*
Gauthier SALLET, *Université de Lorraine, Nancy, France*
Tewfik SARI, *IRSTEA, Montpellier, France*
Suzanne TOUZEAU, *INRA-Inria, France*
Hatem ZAAG, *University Paris 13, France*
Nejib ZEMZEMI, *Inria, France*

LISTE DES RELECTEURS – *LIST OF REFEREES*

Ezzine ABDELHAK
Nahla ABDELLATIF
Meziane AIDER
Soraya AIT CHELLOUCHE
Abdelaziz AIT MOUSSA
Olivier ALATA
Hugo ALATRISTA-SALAS
Pascal ANDRE
Julien ARINO
Abdon ATANGANA
Mejdi AZAIEZ
Jérôme AZE
Eric BADOUEL
Olivier BARAIS
Kamel BARKAOUI
Monique BARON
Nicolas BECHET
Fethi Bin Muhammad BELGACEM
Hacene BELHADEF
Monia BELLALOUNA
Hend BEN AMEUR
Faker BEN BELGACEM
Slimane BEN MILED
Fethi BIN MUHAMMAD BELGACEM
Adel BLOUZA
Béatrice BOUCHOU MARKHOFF
Tarik BOUJIHA
Ibrahim BOUNHAS
Sandra BRINGAY
Patrice BUCHE
Fabien CAMPILLO
Gaoussou CAMARA
Jérôme CANALS
Bernard CAZELLES
Nicolas CHAMPAGNAT
François CHAROY
Nadia CHOULAIEB
Jean CLAIRAMBAULT
Célia DA COSTA PEREIRA
El Mostafa Daoudi
Mohamed DAOUDI
Laurent DEBREU
Cheikh Talibouya DIOP
Aissani DJAMIL
Mahieddine DJOUDI
Yves DUMONT
Bilel ELAYEB
Melhem EL HELOU
Abdellatif ENNAJI
Jocelyne ERHEL
Radhouene FEKIH SALEM
Davide FREY
Yles FALCONE
Christian FOTSING
Abdoulaye GAMATIE
Jean-Frédéric GERBEAU
Nabil GMATI
Arnaud GRIGNARD
Laura GRIGORI
Lamia GUELLOUZ
Bamba GUEYE
Abdou GUERMOUCHE
Abderrahmane HABBAL
Kais HADDAR
Denis HAMAD
Ridha HAMBLI
Ahmed HAMMOUCH
Nejla HARIGA-TATLI
Seridi HASSINA
Yousri HENCHIRI
Jean-Claude HOCHON
Michel HURFIN
Marc IBRAHIM
Dino IENCO
Abderrahman IGGIDR
Clement JONQUET
Eric KERGOSIEN
Adil KHALIL
Georges-Edouard KOUAMOU
Samer LAHOUD
Michel LANGLAIS
Maryline LAURENT
Philippe LEMOISSON
Christophe LETT
Moussa LO
Claude LOBRY
Cédric LOPEZ
Pascal LORENZ
Stéphane MAAG
John MADDOCKS
Pierre MANNEBACK
Lhoussaine MASMOUDI
Rachid MCHICH
Ludovic ME
Sylvie MELEARD
Nizar MESSAI
Maher MOAKHER
Isabelle MOUGENOT
Ahmed MOUSSA
Ali MOUSSAOUI
Yanik NGOKO
Tri NGUYEN-HUU
Boniface NKONGA
Rachid OULAD HAJ THAMI
Congduc PHAM
Patrice QUINTON
Damien ROBERT
Benjamin ROCHE
Mathieu ROCHE
Pierre ROLIN
Gauthier SALLET
Abed Ellatif SAMHAT
HASSAN SANEIFAR
Tewfik SARI
Abderrahmane SBIHI
Christophe SIBERTIN-BLANC
Fathia SAIS
Idrissa SARR
Mokhtar SELLAMI
William SHU
Yahya SLIMANI
Joël SOR
Salvatore TABBONE
Claude TANGHA
Maguelonne TEISSEIRE
Ousmane THIARE
Raja TOUAHNi
Suzanne TOUZEAU
Denis TRYSTRAM
François VERNARDAT
César VIHO
Hatem ZAAG
Lynda ZAOUI
Nejib ZEMZEMI
Djemel ZIOU

TABLE DES MATIERES / TABLE OF CONTENTS

Coupled bio-physical models for the transport of banana shrimps of the Sofala Bank, Mozambique

Bernardino S. Malauene[a,b,*], Atanasio Brito[a], Coleen L. Moloney[b], Michael J. Roberts[c], Francis Marsac[d], Pierrick Penven[e], Christophe Lett[f]

[a] Instituto Nacional de Investigacao Pesqueira, Av. Mao Tse Tung 309, Maputo, Mozambique dinomalawene@yahoo.com.br or MLNBER003@myuct.ac.za

[b] Biological Sciences Dept. and Marine Research Institute, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa.

[c] Nelson Mandela Metropolitan University, Port Elizabeth, South Africa

[d] Institut de Recherche pour le Développement, UMI 209, Centre de Recherche Halieutique Méditerranéenne et Tropicale, Avenue Jean Monnet - BP 171 - 34203 Sète Cedex, France.

[e] Institut de Recherche pour le Développement, Centre IRB de Bretagne, B.P. 70 – 29280, Plouzane, France.

[f] Institut de Recherche pour le Développement, UMR 248, Centre de Recherche Halieutique Méditerranéenne et Tropicale, Avenue Jean Monnet - BP 171 - 34203 Sète Cedex, France.

ABSTRACT. The Sofala Bank supports an important penaeid shrimps fishery where *Penaeus indicus* and *Metapenaeus monoceros* (banana shrimp) are the two main target species. The purpose of the present paper is to investigate the roles of biophysical processes on transport of larvae of banana shrimps on the Sofala Bank. A high-resolution two-way nested Regional Ocean Modeling System (ROMS) of the Sofala Bank is developed. The ROMS solution agrees well with available observations and literature. An individual-based model (IBM) using Ichthyop coupled to the ROMS outputs is developed for the banana shrimps larvae on the bank. Simulated larval transport are influenced by the offshore mesoscale eddy activity.

KEYWORDS: IBM, mesoscale eddies, larva, *Penaeus indicus, Metapenaeus monoceros*.

1. Introduction

The Sofala Bank is located within 16° S (near Angoche) and 21° S (Bazaruto archipelago) on the western side of the Mozambique Channel between Madagascar and the African mainland. The continental shelf is generally wide and shallow, with an average depth 20-30 m. The bank is a key habitat for the shallower water penaeid shrimps in the Southwest Indian Ocean (Ivanov and Hassan, 1976). It supports an important multi-sector and multi-species shrimp fishery. The two most important species are the closely related *Penaeus indicus* and *Metapenaeus monoceros* (so-called "banana shrimps) that contribute > 80 % of the total catch (de Sousa et al., 2008). The catch has been declining from >7000 tons in 2004 – 2006 to a low level of ~ 2000 tons in 2012 (de Sousa et al., 2013). This decrease is thought to be related to a combination of detrimental environmental factors and overfishing (de Sousa et al., 2013). However, no conclusive evidence of either overfishing or environmental factors has been found on the Sofala Bank.

Shrimp catch depends to a large extent on recruitment of juveniles into the fishery. This is driven by environmental factors that influence larval transport and dispersal (Ehrhardt and Legault, 1999). It is known that banana shrimps on the Sofala Bank spawn all year round (de Sousa et al., 2008; Malauene 2015) and their eggs develop to first postlarvae within 15 days (i.e. passive pelagic larval duration – PLD). During such PLD currents can transport shrimp larvae either shoreward or offshore (Penn, 1975).

The Mozambique Channel circulation is dominated by mesoscale

eddies and rings. These eddies, and particularly dipole eddies, can generate high velocity offshore-directed boundary currents (Roberts et al., 2014). Many studies have shown that eddy-induced currents can transport coastal biotic and abiotic material offshore (Tew-Kai and Marsac, 2009; Malauene et al., 2014). It is hypothesized that shrimp larvae similarly can be transported to offshore regions where they are unable to survive. The aim of this paper is to investigate the interactive roles of biophysical processes on transport of the banana shrimp larvae in tropical, shallow waters of the Sofala Bank. In the presence of limited observational data the present study is mostly based in numerical models.

2. Model and data

2.1. Hydrodynamic ROMS model for ocean circulation

The Regional Ocean Modeling System (ROMS) is a three-dimensional, split-explicit, free-surface, topography-following vertical and horizontal sigma-coordinate ocean model (Shchepetkin and McWilliams, 2005). The ROMS_AGRIF used in this study uses a fourth-order advection scheme, which reduces dispersive property errors and enhances model resolution of smaller scale processes.

A model domain encompasses the entire Sofala Bank and the offshore adjacent waters between roughly 14 – 24° S. The model uses a structured regular square grid in the horizontal plane with 6.36 km (1\16°) resolution for the large, i.e. parent grid. For a better representation of small-scale coastal features a second fine grid was created using two-way nesting (Debreu et al., 2012); the child grid at 2.12 km (1\48°) resolution.

The model topography was derived from the General Bathymetric Chart of the Oceans (GEBCO) One Minute Grid data set (available at

http://www.gebco.net/data_and_products/gridded_bathymetry_data/) and interpolated to the parent and child grid. Both parent and child grid has 50 vertical sigma-layers.

For surface forcing monthly climatologies are used. Sea surface wind stress from the Quick Scatterometer (QuikSCAT) satellite at a grid resolution of 1\2°. Surface fresh-water and heat fluxes from Comprehensive Ocean-Atmosphere Data Set (COADS) also at 1\2° resolution. Sea surface temperature (SST) from Pathfinder satellite observations at 9 km resolution.

The lateral boundaries of the model domain are open everywhere except at the coast. For the lateral open boundary conditions it was used outputs from the South-West Indian ocean Model (SWIM, Halo et al. 2014) applying the one-way nesting technique (Mason et al., 2010). Tides (ten constituents M2, S2, N2, K2, K1, O1, P1, Q1, Mf and Mm) at 1/4° resolution from the Global Inverse Tide Model data set (TPXO6.2) were also integrated into the model boundaries.

Four rivers Licungo, Zambezi, Pungue and Buzi that drain into the Sofala Bank were considered in one model experiment. Rivers were included as point sources of tracers (temperature and salinity) and momentum (realistic river flow) made available in monthly climatology by the Mozambican National Directorate of Water.

2.2. Individual-based model for larval transport

Individual-based model (IBM) is used here to simulate transport of banana shrimp larvae on the Sofala Bank. The IBM simulations were developed using Ichthyop version 3.1 (Lett et al., 2008, available at http://www.ichthyop.org) coupled to the nested ROMS model of the Sofala Bank. Ichthyop is a Lagrangian transport tool that tracks the trajectories of virtual eggs and larvae providing information of their state: position (longitude, latitude and depth), age (days) and status (alive or dead) at each time step.

Nine release areas (including spawning and non-spawning locations) were defined for the IBM simulations, based in the actual spawning locations for banana shrimp on the Sofala Bank identified by Malauene (2015). Simulations consisted of randomly releasing 30000 virtual banana shrimp eggs within the release areas every three days for five years and tracking their trajectories for 15 days (PLD). During this period simulated larvae could either stay on the bank (considered as successfully retained) or transported out (considered as lost).

2.3. Altimetry data

To evaluate the model ability to reproduce the mesoscale eddy activity, weekly “Delayed Time – DT” mapped absolute dynamic topography (MADT) at grid resolution of 1/4° from 1993 to 1999 data were used. The data combine sea surface height (SSH) observations merged from multi-satellite (TOPEX/Poseidon, Jason-1, GFO, ERS-1, ERS-2 and ENVISAT) altimeter missions processed by SSALTO/Duacs and distributed by AVISO with support from Centre National d'Etudes Spatiales (CNES, http://www.aviso.oceanobs.com), hereafter refereed to as AVISO observations.

3. Results

3.1 Simulated eddies variability, circulation and structure

The model sea surface height (SSH) and the AVISO altimetry agree reasonable well (Fig. 1A and B), in particular, the high mean SSH from the northernmost limit down the channel, the offshore low SSH centered at ~22° S and ~40° E, and the west-east SSH gradient over the slope following the bathymetry between 200 and 2000 m depth. A similar strong slope mean SSH gradient was found in another model study of the Mozambique Channel (Quartly et al., 2013).

Mean eddy kinetic energy (EKE) computed from the model SSH and from AVISO observations are in qualitative agreement (Fig. 1C and D), especially, the two centers of maximum energy one between 21-22° S and 38° E and the other between 19-20° S and 39° E. Quantitatively, the model energy doubled that of AVISO observations, suggesting that the model overestimated EKE by some ~50 %.

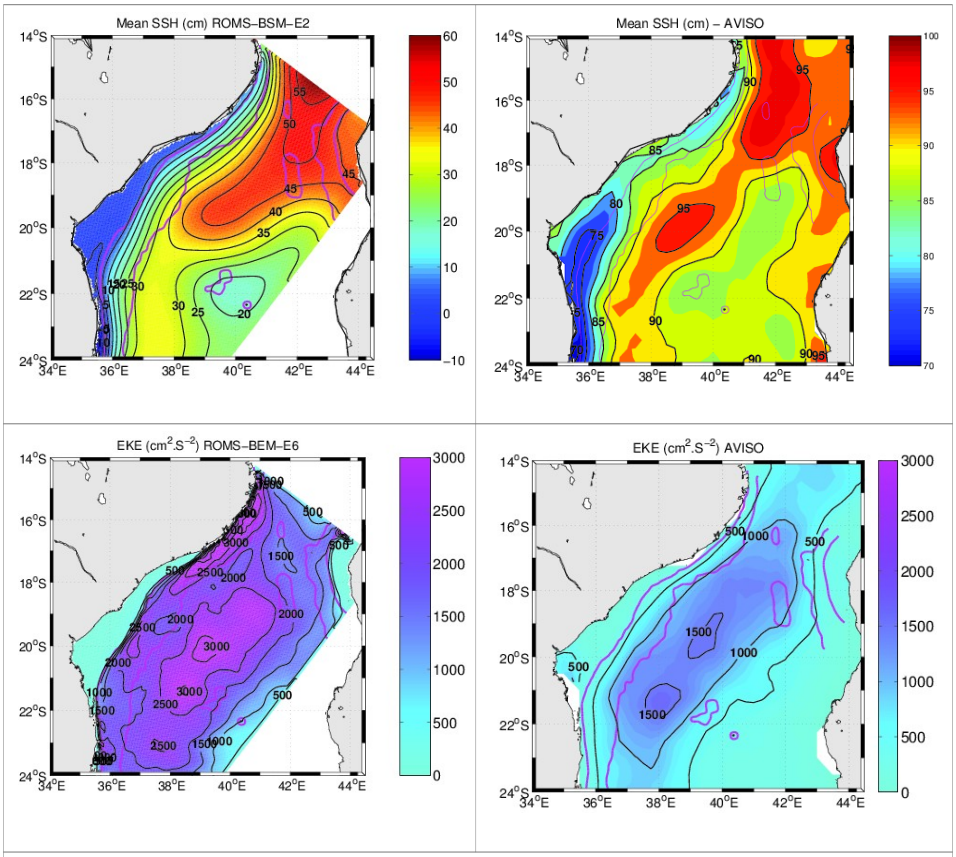


Figure 1 : Comparison between mean SSH (A) derived from ROMS years 4-10, (B) mean absolute dynamic topography derived from AVISO between 1993-1999. And mean EKE (cm^2s^{-1}) from (C) the model and (D) AVISO. Pink lines indicate the 200 and 2000 m bathymetry contours.

3.2 Simulated patterns of the banana shrimp larval transport

Simulated larvae density distribution indicated that larvae were found all over the model domain, with the high concentration on the shelf of the Sofala Bank, but some exited the bank (not shown). Snapshots of trajectories of simulated larvae show that most of the path of the larvae transported out of the bank display a circular shame (Fig. 2A, B and C), supporting the influence of the Mozambique channel eddies in advecting the banana shrimp larvae (Malauene 2015). In other cases, as depicted in Fig. 2D, nearly all larvae stayed on the bank for the full duration of the simulation. This coincided with period of weak or calm Mozambique Channel eddy activity (Malauene 2015). It is apparent that larvae originated from the southern release areas off Beira move little (Fig. 2A-D).

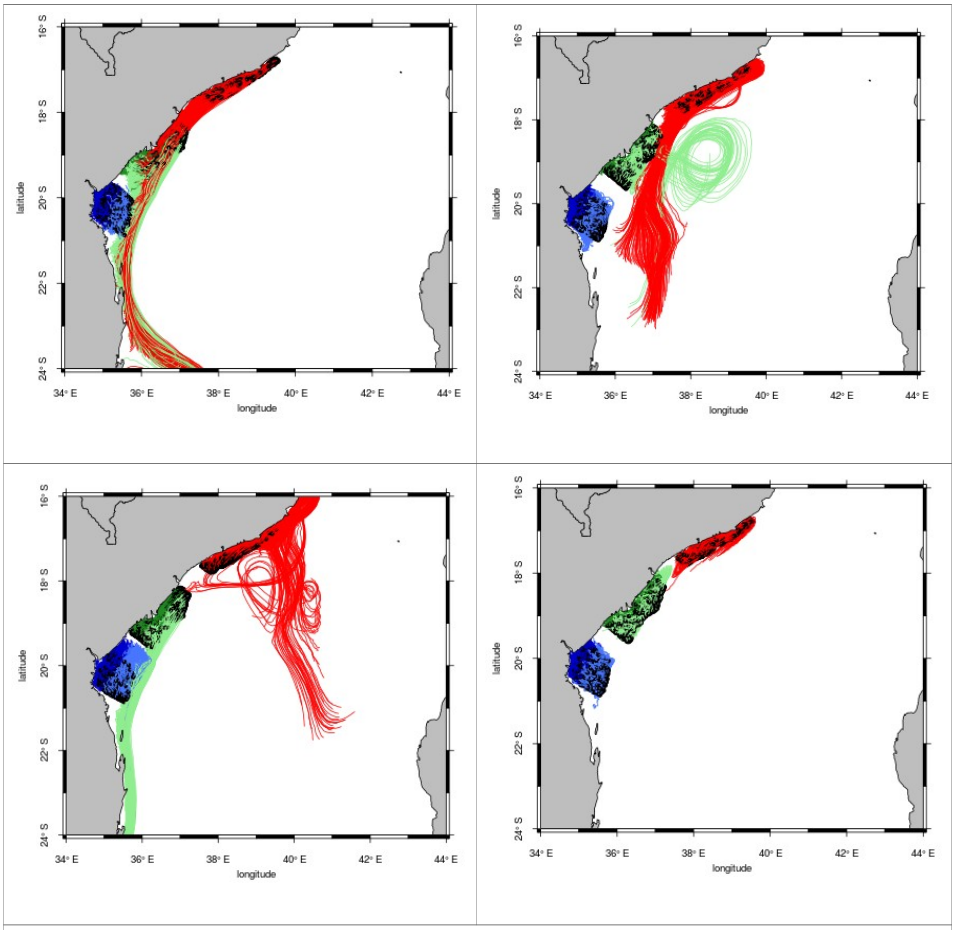


Figure 3 : Snapshots of trajectories of simulated larvae originated from the northern (red), central (green) and southern (blue) release areas. For green and blue releases areas, dark colors indicate inshore and light offshore. for simulations starting: (A) 2 February, (B) 21 March, (C) 3 June and (D) 5 July.

4. Discusion

The strong west-east gradient of mean SSH apparent over the Mozambican continental slope between the 200 and 2000 m isobaths in both the model and AVISO observations is an indication of the presence of a mean “Mozambique Current”. The model current, however, was stronger than that observed from AVISO. This probably because of the high-resolution (~6 km) of the model compared to the global, smoothing and coarser resolution (~25 km) of AVISO observations (Quartly et al., 2013).

The model overestimation of the mean EKE and thus the mean eddy variability in about 50% is gained from the SWIM climatology model used here for the lateral open boundary conditions. The elevated mean EKE from the model and AVISO, however, occurred at the same place, indicating that the model reproduce the Mozambique Channel eddy variability. According to Halo et al. (2014), the SWIM model overestimated the Mozambique Channel eddy variability relative to AVISO by about 40-50 % probably because SWIM reproduces the eddies with larger diameter and higher amplitude than AVISO.

The present study shows that the offshore highly energetic eddies of the Mozambique Channel strongly influence the Sofala Bank circulation and river plume direction. The direction and magnitude of the eddy impact depend on the eddy type, strength and proximity to the shelf. Offshore eddies have little impact on the dominant tidal region off Beira.

Offshore eddies influence the pattern of simulated larval transport on the Sofala Bank except off Beira Bay. Bay of Beira is semi-enclosed and thus protected from the impact of these eddies. Generally in the absence of mesoscale eddy activity larvae stay in the Sofala Bank. Eddies therefore are unlikely to produce a continuous declining in the catch.

5. Reference

de Sousa, L. P., Abdula, S., de Sousa, B. P., Penn, J., and Howell, D. (2013). *The shallow water shrimp at Sofala Bank Mozambique 2013*. Unpublished report Instituto Nacional de Investigação Pesqueira, Maputo.

de Sousa, L. P., Brito, A., Abdula, S., Penn, J., and Howell, D. (2008). *O Camarao do Banco e Sofala 2008*. Unpublished report Instituto Nacional de Investigação Pesqueira, Maputo. (In Portuguese).

Debreu, L., Marchesiello, P., Penven, P., and Cambon, G. (2012). Two-way nesting in split-explicit ocean models: Algorithms, implementation and validation. *Ocean Modelling*, 49 – 50:1 – 21.

Ehrhardt, N. M. and Legault, C. M. (1999). Pink Shrimp, *Farfantepenaeus duorarum*, Recruitment Variability as an Indicator of Florida Bay Dynamics. *Estuaries*, 22(2):pp. 471 – 483.

Halo, I., Backeberg, B., Penven, P., Ansorge, I., Reason, C., and Ullgren, J. E. (2014). Eddy properties in the Mozambique Channel: A comparison between observations and two numerical ocean circulation models. *Deep Sea Research Part II: Topical Studies in Oceanography*, 100(0):38 – 53.

Ivanov, B. G. and Hassan, A. M. (1976). Penaeid Shrimps (Decapoda, Penaeidae) Collected of East Africa by the Fishing Vessel "Van Gogh", 1. Solenocera ramadani sp. nov., and Commercial Species of the Genera Penaeus and Metapenaeus. *Crustaceana*, 30(3):241 – 251.

Lett, C., Verley, P., Mullon, C., Parada, C., Brochier, T., Penven, P., and Blanke, B. (2008). A Langrangian tool for modelling ichthyoplankton dynamics. *Enviromental Modelling & Software*, 23:1210 – 1214.

Malauene, B. S., Shillington, F. A., Roberts, M. J., and Moloney, C. L. (2014). Cool, elevated chlorophyll-*a* waters off northern Mozambique. *Deep Sea Research Part II: Topical Studies in Oceanography*, 100(0):68 – 78.

Malauene, B. S. (2015). *Environmental influences on the banana shrimps of the Sofala Bank, Mozambique Channel*. PhD thesis, Department of Biological Sciences - University of Cape Town, South Africa.

Mason, E., Molemaker, J., Shchepetkin, A. F., Colas, F., McWilliams, J. C., and Sangra, P. (2010). Procedures for offine grid nesting in regional ocean models. *Ocean Modelling*, 35(1-2):1 – 15.

Penn, J. W. (1975). The influence of tidal cycles on the distributional pathway of *Penaeus latisulcatus* Kishinouye in Shark Bay, Western Australia. *Australian Journal of Marine and Freshwater Research*, 26:93 – 102.

Quartly, G., de Cuevas, B., and Coward, A. (2013). Mozambique channel eddies in GCMS: A question of resolution and slippage. *Ocean Modelling*, 63(0):56 – 67.

Roberts, M. J., Ternon, J.-F., and Morris, T. (2014). Interaction of dipole eddies with the western continental slope of the Mozambique Channel. *Deep Sea Research Part II: Topical Studies in Oceanography*, 100(0):54 – 67.

Shchepetkin, A. F. and McWilliams, J. C. (2005). The regional ocean modeling system (ROMS): A split-explicit, free-surface, topography-following-coordinate oceanic model. *Ocean Modell.*, 9:347 – 404.

Tew-Kai, E. and Marsac, F. (2009). Patterns of variability of sea surface chlorophyll in Mozambique Channel: A quantitative approach. *Journal of Marine Systems*, (77):77-88.

Rubrique

Social Network Analysis

Novel method to find directed community structures based on triads cardinality

Gamgne Domgue Félicité* — Tsopze Norbert* — René Ndoundam*

* Computer Science Department - University of Yaounde I
BP 812 Yaounde - Cameroon
felice.gamgne@gmail.com, tsopze@uy1.uninet.cm, ndoundam@gmail.com

RÉSUMÉ. La détection des communautés est davantage un challenge dans les l'analyse des réseaux orientés. Plusieurs algorithmes de détection de communautés ont été developpés et considèrent la relation entre les nœuds comme symmétrique, car ils ignorent l'orientation des liens, ce qui biaise les résultats en produisant des communnautés aléatoires. Ce document propose un algorithme plus eff cace, TRICA, basé sur l'extraction des kernels qui sont des ensembles de nœuds inf uents dans le réseau. Cette approche découvre des communautés plus signif catives avec une complexité temporelle meilleure que celles produites par certains algorithmes de détection de communautés de l'état de l'art.

ABSTRACT. Community structure extraction is once more a major issue in Social network analysis. A plethora of relevant community detection methods have been implemented for directed graphs. Most of them consider the relationship between nodes as symmetric by ignoring links directionality during their clustering step, this leading to random results. This paper propose TRICA, an eff cient clustering method based on kernels which are inf uencial nodes, that takes into account the cardinality of triads containing those inf uencial nodes. To validate our approach, we conduct experiments on some networks which show that TRICA has better performance over some of the other state-of-the-art methods and uncovers expected communities.

MOTS-CLÉS : Réseaux orientés, détection des communautés kernel, Triade

KEYWORDS : Directed graphs, Community kernel detection, Triad.

1. Introduction

Community detection in directed networks appears as one of dominant research works in network analysis. The top meaning of community is a set of nodes that are densely connected with each other while sparsely connected with other nodes in the network [1]. This definition is interesting for undirected graphs ; like this many community detection algorithms implemented for directed networks simply ignore the directionality during the clustering step while other technics transform the directed graph into an undirected weighted one, either unipartite or bipartite, and then algorithms for undirected graph clustering problem can be applied to them.

These simplistic technics are not satisfactory because the underlying semantic is not retained. For example, in a food web network, according to them, the community structure will be corporated of predator species with their prays. This reflexion is not quite right. To make up for that idea, a generic definition of community detection consists of clustering nodes with homogeneous semantic characteristics(nodes centred around a set of objects owning the same interest). Our approach is based on extending the idea that within "good"communities, there are influencial nodes [6], ***kernels***, that centralize information, so that it will easily be attainable. Influancial nodes are crossed by a maximal number of triads in a community. A triad is a set of 3 nodes whose at least 2 are the *in-neighbor nodes* (target vertices) of the 3^{rd} vertex, or according to the triadic closure. Consequently, triads are the basis of many community structures [3]. Here we focus on the link orientation in triads. The specific contributions of our paper are :

– we mainly define a new concept named *kernel degree* to measure the strength of the pair of nodes and the similarity of vertices and give a new sense definition to kernel community based on the triadic closure.

– we develop a novel algorithm based on kernel degree to discover kernels and then communities from real social networks.

– We conduct to better quality improvement over the commmunity kernel detection algorithms.

The rest of paper is organized as follows. Section 2 is an introduction to related works. In Section 3, we formally define several concepts used into the proposed clustering method. In Section 4, we develop the algorithm. Section 5 is experiment study and Section 6 concludes this study.

2. Related works

Most approaches focused on symmetric models which lose the semantics of link directions, a key factor that distinguishes directed networks from undirected networks. For detecting communities in directed networks [2], some studies propose a simple scheme that converts a directed graph into undirected one, this enabling to utilize the richness and complexity of existing methods to find communities in undirected graphs, thus, to mesure cluster strength, they use an objective function, *the modularity*. Yet, this mesure has a limit resolution [1]. More recently, various probabilistic models have been proposed for community detection [7]. Among them, stochastic block models are probably the most successful ones in terms of capturing meaningful communities, producing good performance, and offering probabilistic interpretations. However, its complexity is enough because in pratice, if the number of iterations goes beyond 20, the method discontinue

and results become insignificant. To make up for this complexity, some authors define "kernels"like described below.

A *kernel* is considered as a set of influencial nodes inside a group. It seems to be information centralizing nodes. Some methods explored the problem of detecting community kernels, in order to either reduce the number of iterations, and consequently the time-complexity of algorithms defined for complex social networks or uncover the hidden community structure in large social networks. [4] identifies those influential members, *kernel* and detects the structure of community kernels and proposed efficient algorithms for finding community kernels. Through these algorithms, there is a random choice of the initial vertex, and the size of communities is fixed, leading to an arbitrary result estimation. To keep going, [3] proved that triangles(short cycles) play an important role in the formation of complex networks, especially those with an underlying community structure [5] and converts directed graph into an undirected and weighted one. This transformation misses the semantic of links. We propose a method which extracts triads based on Social properties to characterize the structure of real-world large-scale networks.

3. Method formalization

We propose in this section the kernel community model and introduce several related concepts and necessary notations.

3.1. Kernel community model

In directed networks, the link direction gives a considerable semantic to the graph and to the information flow. On twitter network for example, the notion of authority is pointed up as illustrated in Fig 1.(a), because of the relationship between a set of authoritative or hub blogs (nodes u and v) and a set of non-popular one called followers (nodes x) as presented in Fig 1.(b) and Fig 1.(c).

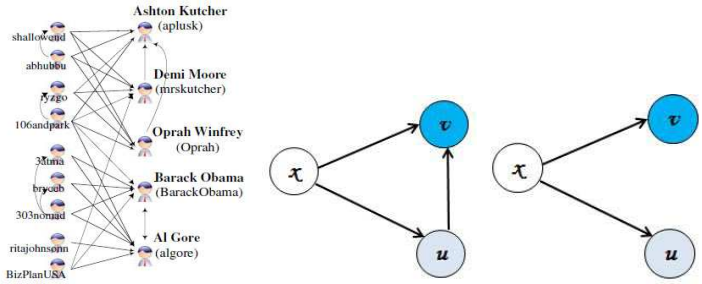
We integrated this concept of authority as one concept named **kernel degree**. Fig 1.(a) is a visualization of an extract from a twitter network. Kernel communities consist of nodes owning the same "in-neighbourhood "which corresponds to nodes that have more connections to the kernel (and not from the kernel) than a vertex outside the kernel. We consider only ingoing edges to the kernel vertices to express the strength these nodes get in some kind of network treated in this paper ; in a twitter network for example, hub blogs are viewed by many others followers and not the opposite ; in a citation network for example, authoritative authors like pioneers in a research area are more quoted by the others junior researchers. On the beginning, the kernel consists of two vertices sharing the same properties, leading to the notion of "triad "which consists of the idea that two vertices of the kernel share the same friend, like defined in the following sub-section.

3.2. Basic terminology and concepts

Given a directed graph $G(V, E)$ with $n = |V|$ vertexes and $m = |E|$ edges. Let Γ_u be the neighborhood vertices set of vertex u. We now give some following useful definitions :

Definition 1 (*Triad weight*). Let the identifier of vertex x in G be j. The triad weight of any edge (u, v) in graph G can be represented as Δ. We can use TW_{uv} to represent the number of triads (triad cardinality) crossing u and v according to the scheme presented in the Fig 1.(b) and Fig 1.(c).

$TW_{uv} = \frac{|\Delta_{uv}|}{|\Delta_j|}$.



(a) Illustration of Twitter Network (b) Closed triad (c) Opened triad

Figure 1. *Basic structures of our kernel community model.*

Definition 2 (*Neighborhood overlap*). Given two vertices u and v, let Γ_u be the set of vertices that are the neighborhood of vertex u, let Γv be the set of vertices that are the neighborhood of vertex v. Let NO_{uv} be the neighborhood overlap of u and v. $NO_{uv} = \frac{|\Gamma_v \cap \Gamma_u|}{|\Gamma_v \cup \Gamma_u| - 2}$ if there is an edge between u and v and 0 otherwise.

Definition 3 (*The kernel degree*). The Kernel degree of a pair of vertex u and v is : $K_{uv} = TW_{uv} * NO_{uv}$. K_{uv} can measure the strength of the pair (u, v) and the similarity of nodes.

Definition 4 (*New sense Kernel Community*). A new definition of the kernel community in the sense of this paper is a set of vertices with the same neighborhood such as these neighbors expand inward to the kernel, according the kernel degree K_{uv} gradually until its minimum.

Definition 5 (*Triadic Closure*). If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future.

The algorithm is structured into two steps : detecting kernel communities and then migrating the others vertexes to the kernel to whom they are more connected to.

4. Our Method for extracting communities

The new algorithm is structured in two steps : identifying kernels, then migrating the other vertices to the kernel as described in the following subsections. The algorithm for extracting Kernel communities, TRICA (Triads Cardinality Algorithm) we propose here makes use of a new concept *Kernel degree*, that measures the strength of a kernel gradually until it decreases. This concept is based on the triadic closure for emphasis the semantic proximity that links community members conducting to efficient propagation of information over the network. We focus on triads cardinality that is the number of neighboors two nodes own.

Data set	Vertices	Edges	Types
Extract from Twitter Network	14	31	Directed
American Football Network	115	613	Undirected
Celegansneural	297	2359	Directed

Tableau 1. *Data sets description*

4.1. TRICA algorithm

We assume that the network we want to analyze can be represented as a connected, directed, nonvalued graph G of $n = |N|$ nodes and $m = |E|$ edges. This step for identifying kernels is described in four sub-steps as follow :

1) Detect the *in-central* vertex v, which is the vertex with the maximal in-degree in the graph.

2) Determine the neighborhood overlap of each edge *(u,v)* through a variant of *Jaccard Index*[1] represented by NO_{uv} as defined in **Definition 2**

3) Store neighboorhood vertices u of v like $NO_{uv} > \varepsilon$

4) Compute K_{uv} through the *triad weight* TW_{uv} as described in **Definition 1**. This action is repeated to measure the strength of a kernel gradually until K_{uv} decreases.

These 4 substeps are repeated n/k times, k being the *in-degree* of vertex v.
The space complexity of TRICA is $O(n+m)$, and it runs in time more quickly than some of the state-of-the-art algorithms like shown in experiments.

The TRICA implementation for kernel communities is presented in Algorithm 1.

4.2. Deduction of global communities

After extracting kernels, it remains the other nodes which don't belong to the kernels ; they are called *non-kernels vertices*. The process of generating *global communities* (communities containing both kernels and non-kernels vertices) consists of migrating the other members (belonging to a set called "auxiliary communities") to the kernel whith whom they have a maximum number of connections, as described in Algorithm 2.

Algorithm 1 TRICA implementation for kernels extraction

Data: Directed graph $G = (N, E)$
Result: K Kernels
Initialisation : $K = \emptyset$;
repeat
  $k = d^{in}(v)/d^{in}(v) = max\{d^{in}(t), \forall t \in V\}$;
  Calculate NO_{uv} for each $(u, v) \in E$;
  $\Gamma_v[] \leftarrow \{t \in V/\exists t \in V, NO_{tv} > 0, 8\}$;$\Gamma_v[].sort$;$i \leftarrow 1$;
  $S \leftarrow \emptyset$;
  $j \leftarrow i$; $u \leftarrow \Gamma_v[j]$;$K^*_{uv} \leftarrow 0$;
  repeat
    Compute K_{uv} ;
    if $(K_{uv} > K^*_{uv})$ **then**
      $S \leftarrow S \cup u$;
    end if
    $u \leftarrow \Gamma_v[i++]$;
  until $(K_{uv} < K^*_{uv})$;
  $K \leftarrow K \cup S$;
until $(|V|/k)$
Return K ;

5. Experiments

To study the effectiveness and accuracy of TRICA, we compare it with following comparative methods :

– NEWMAN : Method for finding community structure in directed networks using the betweenness based on modularity [6].

Algorithms	Extract from Twitter		American Football		Celegansneural	
	$\%\Delta$	Comm Numb	$\%\Delta$	Comm Numb	$\%\Delta$	Comm Numb
Newmann	98%	2	39%	10	28%	194
Louvain	98%	2	63%	9	35%	5
Weba	98%	2	-	8	-	-
Triad Cardinality	**98%**	2	**70%**	12	**64%**	21

Tableau 2. *Community detection performance on the triad cardinality rate where the best rate are in bold.*

– LOUVAIN : Community detection algorithm based on modularity ; (we use Gephi tool for visualizing LOUVAIN results).

– WEBA [4] :Algorithm for community kernel detection in large social networks.

Algorithm 2 Algorithm implementation for non-kernels vertices migration

Data: Communities Kernels $K = \{K_1, K_2, ..., K_t\}$
Result: Global Communities $G_K = \{G_{K_1}, G_{K_2}, ..., G_{K_t}\}$
Let N be set of auxiliary communities ; $N = \{N_{K_1}, N_{K_2}, ..., NG_{K_t}\}$;
$\forall i \in \{1, ..., t\}, G_{K_i} = \emptyset$;
repeat
  $\forall i \in \{1, ..., t\}, G_{K_i} = K_i \cup N_{K_i}$;
  For $i \leftarrow 1$ **to** t **do**
    $S \leftarrow \{v \notin \cup G_{K_i} / \forall j \in \{1, ..., t\}\}$, $|E(v, G_{K_i})| \geq |E(v, G_{K_j})| > 0$;
    $N_{K_i} \leftarrow N_{K_i} \cup S$;
    $G_{K_i} \leftarrow K_i \cup N_{K_i}$;
  End For
until (No more vertices can be added)
Return G_K ;

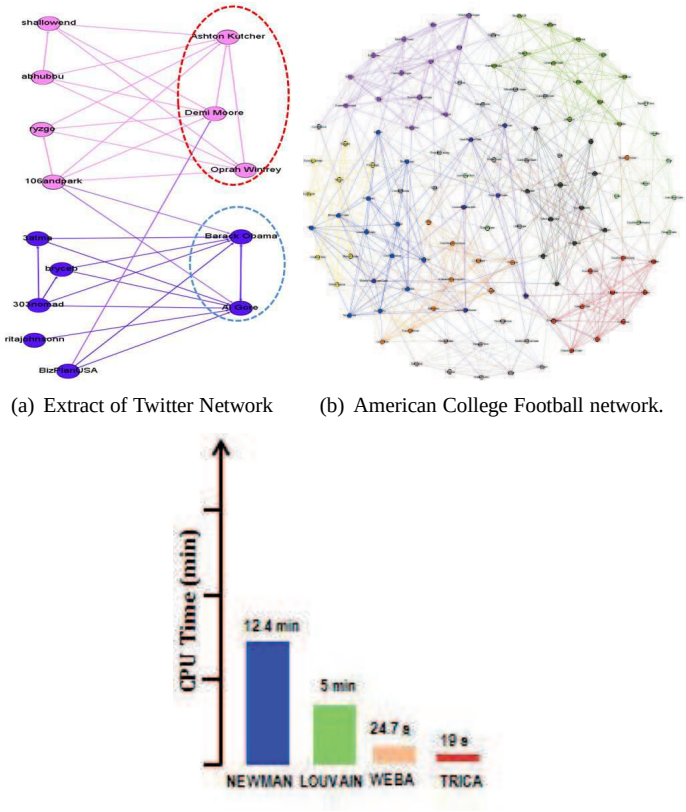
Our method is evaluated on directed and undirected networks. We use two levels of evaluation : The first is based on the time complexity, and the second on the **triad cardinality rate in communities**, that is the percentage of communities in the partition with highest triad cardinality rate. We use the function TCR defined as following, to evaluate our method :

$TCR = \frac{\Sigma_i |\Delta_i|}{|\Delta|}$

where i is one community and $|\Delta|$ the number of triads.

When we apply TRICA on the data sets described in the table 1, results in Fig 2 are following : The Fig 2.(a) illustrates the 2 expected communities of the Extract from Twitter Network, for all of the methods compared, with a triad cardinality rate in communities of 98% with kernels and followers [6]. But TRICA CPU time is better than other methods CPU time, as shown if 2.(c) The table 2 summarizes the comparison with some state-of-the-art methods. It shows that Triad Cardinality algorithm provides the highest triad cardinality rate in communities. As far as the Football network is concerned, Triads cardinality algorithm can divide the network into 12 communities exactly as shown in Fig 2.(b). In this result, 8 communities are completely consistent, this revealed by the triad cardinality rate of 70%. Meanwhile Newmann algorithm can divide it into 10 communities and LOUVAIN into 9. This number of communities does not reflect the real structure of the American College Football network. On the other hand, the result for applying

TRICA to Celegansneural network shown in Table2 presents that TRICA detects 21 communities, while LOUVAIN detects 5 and NEWMAN 194. But the triad cardinality rate is the best, 64%, certifying that our method uncovers a better structure of social networks.



(a) Extract of Twitter Network (b) American College Football network.

(c) Efficiency comparison of TRICA and others algorithms on Twitter Network.

Figure 2. *Results of applying TRICA to data sets.*

6. Conclusion

In this paper, we focus on the problem of kernel community detection in directed graphs, kernels being the key tool for understanding the role of networks and its structure. We mainly interested on extracting kernels which are influential nodes on the network. Our kernel community model define triads according to some social properties to characterize the structure of real-world large-scale network, and we develop a novel method based on the proposed new concept, the *kernel degree* which defines the strength of kernel community. Experiments proved that TRICA detects efficiently expected communities and achieves 20 % performance improvement over some other state-of-the-art algorithms, but it only works for unweighted graphs. Our next work is to optimize Triad cardinality-

based property, and adjust it to suit for detecting kernel communities from large-scale directed and weighted networks.

7. Bibliographie

[1] S. FORTUNATO, « Community detection in graphs », *Physics Reports* 486(3) 75-174, 2010.

[2] F. D. MALLIAROS and M.VAZIRGIANNIS, « Clustering and community detection in directed networks : A survey. », *arXiv* 1308.0971, 2013.

[3] C.KLYMKO , D.F GLEICH and T.G KOLDA, « Using Triangles to Improve Community Detection in Directed Networks », *Conference Stanford University*.

[4] LIAORUO WANG , TIANCHENG LOU , JIE TANG and JOHN E. HOPCROFT, « Detecting Community Kernels in Large Social Networks ».

[5] A. PRAT-PÉREZ , D. DOMINGUEZ-SAL , J. M. BRUNAT and J. L. LARRIBA-PEY, « Shaping communities out of triangles. », *In CIKM 12* n° 1677-1681, 2012.

[6] FÉLICITÉ GAMGNE and NORBERT TSOPZE, « Communautés et rôles dans les réseaux sociaux », *in : CARI '14 : Proceedings of the 12th African Conference on Research in Computer science and Applied Mathematics* n° 157 - 164, 2014.

[7] TIANBAO YANG , YUN CHI , SHENGHUO ZHU and YIHONG GONG and RONG JIN,« Directed network community detection : A popularity and productivity link model. », *In SIAM Data Mining'10* n° 2010.

A comparative study of three membrane fouling models

Towards a generic model for optimization purposes

N. KALBOUSSI[a,*] J.HARMAND[b] N.BEN AMAR[a,**] F.ELLOUZE[a,***]

[a] Département de chimie, Institut National des Sciences Appliquées et de Technologie (INSAT)
Charguia, Tunis 1080, TUNISIE
Laboratoire de Modélisation Mathématique et Numérique dans les Sciences de l'Ingénieur (LAMSIN)

[b] LBE, INRA, 11100, Narbonne, France
jerome.harmand@supagro.inra.fr

* nesrinekalboussi@gmail.com

**nihel.benamar@insat.rnu.tn

***ellouze_fatma@yahoo.fr

ABSTRACT. Most of the published models of membrane fouling are too complex and contain too many parameters to be estimated from experimental data. This works aims to justify the choice from the literature of a simple model of membrane fouling for control and optimization design purposes. To do so, we identify a simple and generic model from the literature and we show, using preliminary results, that this model can reproduce the same results than those much more complicated and specific published models with less parameters to estimate.

RÉSUMÉ. La plupart des modèles de colmatage de la membrane sont compliqués avec beaucoup de paramètres à estimer à partir des données expérimentales. L'objectif de ce travail est de justifier le choix, à partir de la littérature, d'un model simple de colmatage de la membrane pour des fins de contrôle et d'optimisation. Pour ce faire, on identifie un modèle simple et générique et on montre que ce modèle peut reproduire les mêmes résultats que d'autres modèles publiés plus compliqués et spécifiques, avec moins de paramètres à estimer.

KEYWORDS: membrane bioreactor(MBR), fouling, modeling, mathematical models, optimization.

MOTS-CLÉS : bioréacteur à membrane (BRM), colmatage, modélisation, modèles mathématiques, optimisation.

1. Introduction

The membrane bioreactors (MBR) are an increasingly used technology in wastewater treatment. Such a process combines a biological reactor with a filtration membrane that separates microorganisms and suspended matters from the purified water. The advantages are: a high quality effluent, a high solid retention time (SRT), a high possible biomass concentration and a small footprint. Despite its benefits and its widespread use, the MBR technology is constrained by membrane fouling. Fouling is due to the attachment of particles on membrane surface which leads to sever flux decline and an increase of the operating costs. Therefore, several authors have proposed different mathematical models to simulate the MBR process in order to be used in the prediction and control of membrane fouling. However, those models either include a lot of parameters to be estimated from experimental data and they are thus too complex to be really operational, or they make too many assumptions that limit their interest. In this paper, we propose to evaluate a simple generic mathematical model proposed by Benyahia et al. [1] by comparing this model to two other models published in the literature: the model of Pimentel et al. [2] and the model of Di Bella et al. [3]. In particular, we are interested in investigating the generic character of the model proposed by Benyahia et al. [1] for two purposes. The first is to illustrate its usefulness for control and optimization design purposes by justifying the high prediction capabilities of this model despite its simplicity. The second is to prove that if MBRs are very complex systems, yet they can be modeled by simple and generic mathematical models.

To do so, the models [2] and [3] are used as virtual processes to generate data that are then utilized to identify the model parameters of Benyahia et al. model [1] by using an optimization strategy. Model simulations and parameter estimation were conducted using Matlab.

2. The model proposed by Pimentel et al.

Pimentel et al. [2] have proposed an integrated model coupling a biological model and a filtration model. The coupled model is formed of eight ordinary differential equations (ODEs) with six parameters to be estimated from experimental data. The biological model is designed using a simple chemostat reactor, involving one substrate and one biomass. The short-term evolution of the cake deposit on the membrane surface was modeled by equation (1) and the long-term evolution due to irreversible clogging was described by equation (2). In this model, the total resistance is calculated as the cake resistance while the intrinsic resistance of the membrane was neglected (equations (3) and (4)). The trans-membrane pressure can be determined according to equation (5).

For the relaxation phase, the model is represented by equations (6) to (7). The nomenclature used in the model is presented in Appendix 1.

- Coupled model for the filtration phase:

$$\dot{m} = Q_{perm}\, X - J_{air}\, \mu_{air}\, m \quad (1)$$

$$\dot{\beta} = -\gamma\, \beta \quad (2)$$

$$R_{tot} = R_{cake} \quad (3)$$

$$R_{cake} = \rho\, \frac{m + m_0}{A} \quad (4)$$

$$TMP = \frac{Q_{perm}}{A}\, \eta\, R_{tot} \quad (5)$$

- Coupled model for the relaxation phase :

$$\dot{m} = -J_{air}\, \mu_{air}\, m \quad (6)$$

$$\dot{\beta} = -\gamma\, \beta \quad (7)$$

3. The model proposed by Di Bella et al.

The membrane bioreactor mathematical model of Di Bella et al. [3] consists of two sub-models. The biological activity is described in the first sub-model through twenty-six ODEs. This sub-model is a modified version of the well-know ASM1 [4] to consider the influence of the Soluble Microbial Products (SMPs), known as playing a key role in membrane fouling [5]. The cake layer formation was modeled by equation (10). The latter is regulated by two opposite phenomena: the suction which leads to attachment and the friction drag caused by the turbulent air flow. The attachment is proportional to the total suspended concentration as expressed by equation (8) while the friction drag is proportional to the local shear intensity as in equation (9). During backwashing phase, the detachment action of the cake layer is evaluated by equation (11) where η_c is a calibrated parameter. The nomenclature used in Di Bella et al.'s model is given in Appendix 2.

- The model for the filtration phase:

$$MLSS = i_{SS,X_I}\, X_I + i_{SS,X_S}\, X_S + i_{SS,BH}\, X_{BH} + i_{SS,BA}\, X_{BA} \quad (8)$$

$$G = \sqrt{\frac{\rho_S\, g\, Q_a}{\mu_S}} \quad (9)$$

$$\dot{M_{sf}} = \frac{24\, MLSS\, Q_{perm}^2}{24\, Q_{perm} + C_d\, d_p\, G} - \frac{\beta\, (1-\alpha) G\, M_{sf}^2}{\gamma\, V_f\, t_f + M_{sf}} \quad (10)$$

- The model for the backwashing phase:

$$\dot{M_{sf}} = -\eta_c \, M_{sf} \quad (11)$$

Di Bella et al. model includes forty-four parameters to be estimated from experimental data and it does not give equations to calculate the resistance of membrane fouling and thus the transmembrane pressure (*TMP*).

4. The model proposed by Benyahia et al.

Benyahia et al. [1] have proposed a simple model of membrane fouling and have connected it to a b iological process to demonstrate its utility in a large number of situations. In this model, two main fouling phenomena were considered: the attachment of solids onto the membrane surface (cake formation) and the retention of compounds inside the pores (pores clogging), in particular the SMP.

The coupled model of Benyahia et al. is formed of fourteen ODEs: ten ODEs to describe the biological activity and four ODEs to represent the filtration process, with twenty-six parameters. In their work, the authors [1] assume that total filtering membrane surface is not constant, contrary to many models of the literature. Instead, it is modeled by a decreasing function of both the mass of matter attached on the surface of the membrane *m(t)* and the mass of deposited matter into pores *Sp(t)* (notably *SMP*). The dynamic of *m(t)* is proportional to the particulate matter (X_T) and the total soluble (S_T and *SMP*), as in equation (12). The evolution of *Sp(t)* is proportional to *SMP* (cf equation 13). The filtration model of Benyahia et al. [1] is represented by the following dynamical equations and the nomenclature used in this model is given in Appendix 3:

- The model for the filtration phase:

$$\dot{m} = \delta \, Q_{out} \, (C_S \, S_T + C_x \, X_T + C_{SMP} \, SMP) - f_m \, m \quad (12)$$

$$\dot{S_p} = \delta' \, Q_{out} \, (\beta \, SMP + \frac{\beta}{15} \, (S_1 + S_2)) \quad (13)$$

$$R_{tot} = R_0 + \alpha \, \frac{m}{A} + \alpha' \, \frac{V_p \, S_p}{\epsilon \, A} \quad (14)$$

$$TMP = \frac{Q_{perm}}{A} \, \eta \, R_{tot} \quad (15)$$

- The model for the relaxation/backwashing phase:

$$\dot{m} = -\omega \, m \quad (16)$$

$$\dot{S_p} = -\omega' \, S_p \quad (17)$$

5. Identification of Benyahia et al. 's model parameters using Pimentel data

At this stage, Pimentel et al. model [2] is considered as a virtual process to generate data in order to identify the parameter of the model [1]. All the hypothesis considered in Pimentel et al. model [2] were applied to Benyahia et al. model [1]. Therefore, the parameter to be optimized of the model [1] are: δC_x, f_m , α and ω.

The optimization of these parameters was done by the least squares method programmed with Matlab R2013a. The functional cost that was minimized is the sum of the error between the mass of attached matter calculated according to the model [1] and that determined according to the model [2] and the difference between the trans-membrane pressure calculated with the model [1] and that determined with the model [2]. It should be noticed here that in the model [1] the contribution of the SMP in the fouling was neglected in order to fit the hypothesis considered in [2].

The optimal values of the different parameters of the model [1] are presented in the table 1. The results of the simulation of the optimization problem are shown in the Figs.1 and 2. The comparison of the simulated data of the two models confirms the possibility of the model proposed by Benyahia et al. [1] to capture the mean value and the dynamics of the attached mass and the trans-membrane pressure.

Table 1 *Optimal results for parameter estimation of the model [1] from the data of the model [3]*

Parameters	Unit	Value	Lower bound	Upper bound
δC_X	dimensionless	1	0.9	1
f_m	day^{-1}	184.2	160	190
ω	day^{-1}	184.2	160	190
α	$m.g^{-1}$	2.371e+07	2.2e+07	2.4e+07

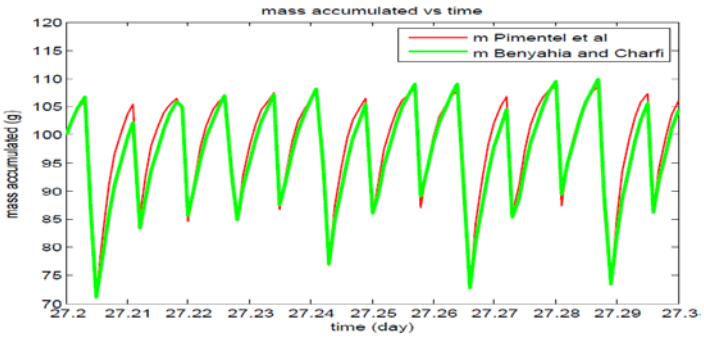


Figure 1. *the accumulated mass on the membrane surface versus time*

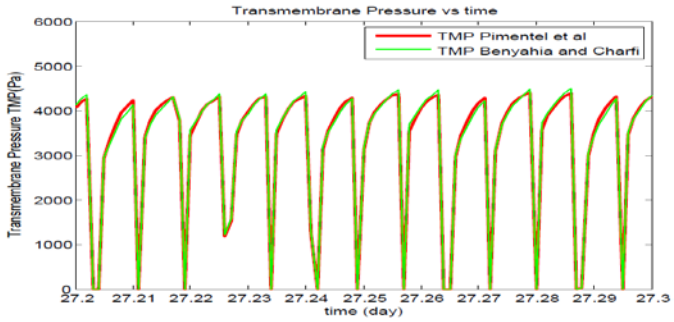


Figure 2. *the TMP trends versus time*

6. Identification of Benyahia et al. 's model parameters using Di Bella data

In this part, the model of Di Bella et al.[3] was considered to evaluate the genericity of Benyahia et al. model [1]. To do that, the same approach as before was considered. The objective function that was minimized by the optimization problem is the difference between the mass of attached matter calculated according to the model [1] and that determined according to the model [3]. The optimal solution and the search ranges of the unknown parameters of Benyahia et al. model are presented in the table 2.

Table 2. *Optimal results for parameter estimation of the model [1] from the data of the model [3]*

Parameters	Unit	Value	Lower bound	Upper bound
δC_x	dimensionless	0.3	0	1
f_m	day^{-1}	3.25e+3	1e+3	3.5e+3
ω	$m.kg^{-1}$	3300	3000	3500

Fig.3 shows the simulation results of the mass attached. These results demonstrate that the model proposed by Benyahia et al. [1] can reproduce well the dynamic of the mass attachment on the membrane. However, this model estimate a mean value of the attached mass slightly different from that evaluated with the model [3]. We explain this difference by the fact that DiBella et al. suppose in their model that the friction drag (the second term of the equation (10)) is a function of the square of the sludge cake which reduce the rate of the sludge deposition in the time. Contrary to DiBella et al., Benyahia et al. model consider that the friction drag is proportional only to the mass attached. For that, the attachment rate evaluated with the model [1] increase much more than that calculated with the model proposed by DiBella et al.

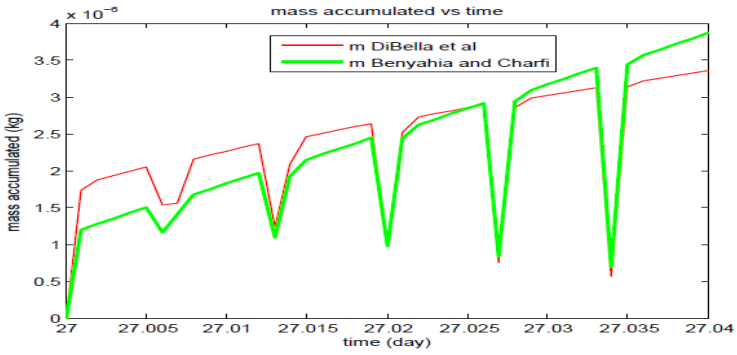


Figure 3. *the mass accumulated on the membrane surface versus time*

7. Conclusion

The aim of this paper is to investigate the capability of the model of Benyahia et al.[1] to capture the dynamics of more complex models. For this purpose, simulations of

two models of the literature, the model of Pimentel el al.[2] and Di Bella et al.[3], were performed. The generated data are used to identify Benyahia et al. model [1] parameters by minimizing the difference between the models predictions. Simulations of the different models were performed by solving a set of differential equation by using the Matlab function ODE. The optimization problem was resolved with the fmincon function in MATLAB. Certainly the model of Pimentel et al. [2] is a simple model but with many assumptions which limit its application. Likewise, comparing to the model of Benyahia et al.[1], the model of DiBella et al. [3] is not taking into account all the fouling mechanisms and it is composed of large number of ODE with many parameters to estimate.

The optimization results show that Benyahia et al. model [1] can capture important properties of the model proposed by Pimentel et al. [2] as the mean value of the trans-membrane pressure and the attached mass on the membrane and their dynamics. The model of Benyahia et al. [1] was able to reproduce the evolution of the attached mass of the model proposed by DiBella et al. [3] but with a little deviation in the values. This deviation can be explained by the difference in the mathematical formulation of the two models [1] and [3]. So, we suggest to add to the drag force of the model of Benyahia et al. a squared term in order to increase the applicability of this model.

Finally, we conclude that the model of Benyahia et al. is generic enough to be used for optimization and control purposes.

8. Bibliography

1. Benyahia, B., et al. *A simple model of anaerobic membrane bioreactor for control design: coupling the "AM2b" model with a simple membrane fouling dynamics*. in *13. World Congress on Anaerobic Digestion: Recovering (bio) Resources for the World. AD13*. 2013. Juan M. Lema, Fernando Fdez-Polanco, Marta Carballa, Jorge Rodriguez, Sonia Suarez 2013.
2. Pimentel, G.A., et al., *Validation of a Simple Fouling Model for a Submerged Membrane Bioreactor.* IFAC-PapersOnLine, 2015. **48**(1): p. 737-742.
3. Di Bella, G., G. Mannina, and G. Viviani, *An integrated model for physical-biological wastewater organic removal in a submerged membrane bioreactor: Model development and parameter estimation.* Journal of Membrane Science, 2008. **322**(1): p. 1-12.
4. Henze, M., *Activated sludge models ASM1, ASM2, ASM2d and ASM3*. Vol. 9. 2000: IWA publishing.
5. Le-Clech, P., V. Chen, and T.A.G. Fane, *Fouling in membrane bioreactors used in wastewater treatment.* Journal of Membrane Science, 2006. **284**(1–2): p. 17-53.

Annexe 1: Pimentel et al. model nomenclature

Symbol	Meaning and Unit
A	The membrane area [m^2]
J_{air}	Air crossflow [$m^3/m^2.d$]
m	Mass cake state [g]
m_0	Initial value of solids m(t) attached onto the membrane area
Q_{in}	Inflow [m^3/d]
Q_w	Waste flux [m^3/d]
Q_{perm}	Permeate flux [m^3/d]
R_{tot}	The total fouling resistance [m^{-1}]
R_{cake}	The cake resistance [m^{-1}]
S	Substrate concentration [g/m^3]
S_{in}	Input substrate concentration [g/m^3]
TMP	The trans-membrane pressure [Pa]
V	Tank volume [m^3]
X	Solid matter concentration [g/m^3]
Y	Yield coefficient of the substrate consumption [-]
μ	Monod's law [1/d]
γ	Constant [day^{-1}]
β	Resistance of detachable cake by air crossflow [m^{-1}]
ρ	The specific cake resistance [m/g]
η	The apparent bulk viscosity [Pa.s]

Annexe 2: Di Bella et al. model nomenclature

Symbol	Meaning and Unit
C_d	Lifting force coefficient [dimensionless]
d_p	Particle size [m]
G	Local shear intensity [day^{-1}]
g	Gravity acceleration [m s^2]
i_{SS,X_I}	Mass particular inert/mass COD in biomass [kg X_I $kgcod^{-1}$]
i_{SS,X_S}	Mass biodegradable organic matter/mass COD in biomass [kg X_S $kgcod^{-1}$]
$i_{SS,BH}$	Mass active heterotrophic biomass/mass COD in biomass [kg X_{BH} $kgcod^{-1}$]
$i_{SS,BA}$	Mass active autotrophic biomass/mass COD in biomass [kg X_{BA} $kgcod^{-1}$]
MLSS	The mixed liquor suspended solids
M_{sf}	Dynamic sludge film cake on the membrane [kg m^2]
Q_a	Air flow [m^3 day^{-1}]
Q_{perm}	Effluent flow rate [m^3 day^{-1}]
V_f	Volume of permeate produced [m^3]
X_I	Particulate inert organic matter [kg COD m^{-3}]
X_S	Particulate biodegradable organic matter [kg COD m^{-3}]
X_{BH}	Active heterotrophic biomass [kg COD m^{-3}]
X_{BA}	Active autotrophic biomass [kg COD m^{-3}]
α	Stickiness of biomass [dimensionless]
β	Erosion rate coefficient of dynamic sludge film [dimensionless]
γ	Compression coefficient for dynamic sludge layer [kg m^{-3} day^{-1}]
ρ_S	Density of activated sludge [kg m^{-3}]
η_c	Efficiency of backwashing [dimensionless]
μ_S	Viscosity of activated sludge [Pa s]

Annexe 3: Benyahia et al. model nomenclature

Symbol	Meaning and Unit
A	Membrane surface [m^2]
C_S	Fraction of ST = S1 + S2 attached onto the membrane at a given time [day^{-1}]
C_x	Fraction of $X_T = X_1 + X_2$ attached onto the membrane at a given time [day^{-1}]
C_{SMP}	Fraction of SMP attached onto the membrane at a given time [day^{-1}]
f_m	Coefficient [day^{-1}]
Q_{out}	The output flow of the bioreactor [m^3. Day^{-1}]
R_{tot}	The total membrane resistance
R_0	Intrinsic membrane resistance
m	Value of solids attached onto the membrane area [kg]
SMP	Soluble microbial products [$kg.m^{-3}$]
S_T	Total substrate [$kg.m^{-3}$]
S_p	Value of the suspended solids blocked into the pores [$kg.m^{-3}$]
V_p	The total volume of the pores [m^3]
X_T	Total biomass [$kg.m^{-3}$]
α	Specific resistance of the sludge [$m.kg^{-1}$]
α'	Specific resistance of the sludge [$m.kg^{-1}$]
β	SMP fraction leaving the bioreactor [-]
δ	Parameter to normalize units [day]
δ'	Parameter to normalize units [-]
ω	Efficiency of backwashing/relaxation [dimensionless]
ω'	Efficiency of backwashing/relaxation [dimensionless]
η	The permeate viscosity [Pa.s]
∈ A	The porous surface [m^2]

Rubrique

Well's location in porous media using topological asymptotic expansion

Wafa Mansouri[1,2], Thouraya Nouri Baranger[2], Hend Ben Ameur[1] and Nejla Tlatli[1]

[1] National Engineering School of Tunis
University of Tunis El Manar
wafa.mansouri@insa-lyon.fr
hbenameur@yahoo.ca
tlatli@topnet.tn
[2] University of Lyon, CNRS, University lyon1
LAMCOS UMR5259, INSA-Lyon
Thouraya.Baranger@univ-lyon1.fr

ABSTRACT. We study the inverse problem of identification of well's location in a porous media via boundary measurements. Our main tool is the topological gradient method applied to a convenient design function.

RÉSUMÉ. Nous étudions le problème inverse d'identification des positions des puits dans un milieu poreux par des mesures sur la frontière. Notre outil principal est la méthode du gradient topologique appliqué à une fonction objectif.

KEYWORDS : Inverse Problem, Topological Gradient, Wells Location.

MOTS-CLÉS : Problème Inverse, Gradient Topologique, Identifications des Puits.

1 Introduction

In hydrogeology, it is very difficult to construct an accurate simulation model for a groundwater system. However, in many real situations, uncertainties can be related to parameters caracterising the aquifer itself, or to external constraints, such as withdrawal rates in wells, drilling and recharge. The knowledge of the aquifer withdrawal rates can represent a largely unknown factor in real problems of groundwater resources modelling. The inverse problem under consideration is to determine the location of wells using boundary measurements. We consider the cases where we have Neumann or Dirichlet condition on boundary of the wells and we use the topological sensitivity method.

The topological sensitivity analysis has been recognized as a promising method to solve topology optimization problems. It consists to derive an asymptotic expansion of a shape functional with respect to the size of a small hole created inside the domain. This method was introduced by Schumacher [12] in the context of compliance minimization. Then, Sokolowski and Zochowski [13] generalized it to more general shape functionals.

To present the basic idea of this method, let us consider a domain Ω in $\mathbb{R}^2$ and a cost functional $j(\Omega) = J(u_\Omega)$ to be minimized, where u_Ω is the solution to a given PDE (model) defined in Ω. For a small parameter $\varepsilon \geq 0$, let $\Omega \backslash \overline{B(x_0, \varepsilon)}$ be the perturbed domain obtained by the creation of a circular hole of radius ε around the point $x_0 \in \Omega$. The topological sensitivity analysis provides an asymptotic expansion of j when ε goes to zero in the form:

$$j(\Omega \backslash \overline{B(x_0, \varepsilon)}) - j(\Omega) = f(\varepsilon) g(x_0) + o(f(\varepsilon)). \tag{1}$$

In this expansion, $f(\varepsilon)$ is a positive function going to zero with ε. The function g is commonly called topological gradient, or topological derivative. It is usually simple to compute and is obtained using the solution of direct and adjoint problems defined on the initial domain. To minimize the criterion j, one has to create holes at some points x where $g(x)$ is negative.
The topological derivative has been obtained for various problems, arbitrary shaped holes and a large class of shape functionals [5, 10].

This work is outlined as follows: Section 2 is devoted to the model setting; section 3 is devoted to the formulation of the inverse problem and the introduction of the topological asymptotic analysis in the case of a well with a Dirichlet or Neumann condition on its boundary. In section 4 we illustrate the efficiency of the proposed method by several numerical experiments then we conclude.

2 The model setting

Let Ω be a domain of $\mathbb{R}^2$ and $\Gamma = \partial\Omega$. We assume that the wells are well separated and have a circular form $\mathcal{O}_{x_k,\varepsilon} = x_k + \varepsilon \mathcal{O}^k$, $1 \leq k \leq m$, where ε is the common diameter and $\mathcal{O}^k \subset \mathbb{R}^2$ are bounded and smooth domains containing the origin. The points $x_k \in \Omega$, $1 \leq k \leq m$, determine the location of the wells (Figure 1). These are the unknowns of our inverse problem.

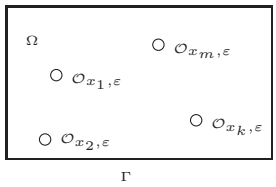


Figure 1: Domain containing m wells having the same radius.

For simplicity, we assume that Ω is a homogeneous geological zone. Following [4], the forward two-dimensional problem of groundwater flow in isotropic and homogeneous medium, with wells $\mathcal{O}_\varepsilon = \bigcup_{k=1}^{m} \mathcal{O}_{x_k,\varepsilon}$ in the domain Ω, can be formulated as follows:

$$\begin{cases} -div(T\nabla u_\varepsilon) & = 0 & \text{in } \Omega\backslash\overline{\mathcal{O}_\varepsilon} \\ u_\varepsilon & = H & \text{on } \Gamma \\ T\nabla u_\varepsilon.n & = 0 \ (\text{or } u_\varepsilon = 0) & \text{on } \Sigma_\varepsilon \end{cases} \tag{2}$$

Where $\Sigma_\varepsilon = \partial\mathcal{O}_\varepsilon$ is the wells boundary; T is the transmissivity and u_ε is the pieozometric head . Let $\Omega_\varepsilon = \Omega\backslash\overline{\mathcal{O}_\varepsilon}$.

3 The inverse problem and the misfit function

We consider the inverse problem of determining well's location from overspecified boundary data on Γ. These data correspond to both Neumann and Dirichlet conditions. We split these data in such a way to build two well posed problems:

- The first problem uses a Neumann condition on Γ,

$$\begin{cases} -div(T\nabla u_\varepsilon^N) & = 0 & \text{in } \Omega_\varepsilon \\ T\nabla u_\varepsilon^N.n & = \Phi & \text{on } \Gamma \\ T\nabla u_\varepsilon^N.n & = 0 \ (\text{or } u_\varepsilon^N = 0) & \text{on } \Sigma_\varepsilon \end{cases} \tag{3}$$

- The second problem uses a Dirichlet condition on Γ,

$$\begin{cases} -div(T\nabla u_\varepsilon^D) & = 0 & \text{in } \Omega_\varepsilon \\ u_\varepsilon^D & = H & \text{on } \Gamma \\ T\nabla u_\varepsilon^D.n & = 0 \ (\text{or } u_\varepsilon^D = 0) & \text{on } \Sigma_\varepsilon \end{cases} \tag{4}$$

We define a misfit function:

$$J(u_\varepsilon^D, u_\varepsilon^N) = \|u_\varepsilon^N - u_\varepsilon^D\|^2_{L^2(\Omega_\varepsilon)} \tag{5}$$

One can remark that if Σ_ε coincides with the actual well boundary Σ_ε^* then the misfit between the solutions vanishes $u_\varepsilon^N = u_\varepsilon^D$.

Our identification problem can be formulated as a topological optimization problem as follows: given a flow ϕ and the measured H, find the optimal location of wells $\mathcal{O}_\varepsilon$ inside the domain Ω minimizing the shape function j

$$(\mathcal{P}_{min}) \quad \min_{\mathcal{O}_\varepsilon \subset \Omega} j(\varepsilon)$$

where

$$j(\varepsilon) = J(u_\varepsilon^D, u_\varepsilon^N). \tag{6}$$

The solution of this inverse problem depends on the boundary condition on the well's boundary taken in (4) and (3).

3.1 Case 1: the topological gradient considering Neumann condition on Σ_ε

The aim of this section is to derive a topological asymptotic expansion for equation (7) with Neumann condition on Σ_ε:

$$\begin{cases} -div(T\nabla u_\varepsilon) & = 0 & \text{in } \Omega_\varepsilon \\ T\nabla u_\varepsilon.n & = \Phi & \text{on } \Gamma \\ u_\varepsilon & = H & \text{on } \Gamma \\ T\nabla u_\varepsilon.n & = 0 & \text{on } \Sigma_\varepsilon \end{cases} \tag{7}$$

A topological sensitivity analysis using Neumann boundary condition has already been obtained for the elasticity equations in [5], for Laplace equation in [2] and for Maxwell equations in [9].

Inspired by the master thesis work [8], the topological gradient method provides an asymptotic expansion of a function j defined in (6) of the form:

$$\begin{aligned} j(\varepsilon) - j(0) &= -2\pi T\varepsilon^2[\nabla u_0^D(z)\nabla v_0^D(z) + \nabla u_0^N(z).\nabla v_0^N(z) \\ &+ \frac{1}{2}|u_0^N(z) - u_0^D(z)|^2] + o(\varepsilon^2), \end{aligned}$$

In this case the topological gradient is defined by:

$$g(z) = -2\pi T[\nabla u_0^D(z)\nabla v_0^D(z) + \nabla u_0^N(z).\nabla v_0^N(z) + \frac{1}{2}|u_0^N(z) - u_0^D(z)|^2]$$

where u_0^N and u_0^D are respectively the solution of the problems (3) and (4) with $\varepsilon = 0$ (in the domain without wells). Then v_0^N and v_0^D are respectively the solution of the adjoint problems associated to the problems (3) and (4) in the domain without wells.

3.2 Case 2: the topological gradient considering Dirichlet condition on Σ_ε

Instead to the case 1, our goal here is to present a topological asymptotic expansion for equation (8) with Dirichlet condition on Σ_ε:

$$\begin{cases} -div(T\nabla u_\varepsilon) & = 0 & \text{in } \Omega_\varepsilon \\ T\nabla u_\varepsilon.n & = \Phi & \text{on } \Gamma \\ u_\varepsilon & = H & \text{on } \Gamma \\ u_\varepsilon & = 0 & \text{on } \Sigma_\varepsilon \end{cases} \tag{8}$$

Topological sensitivity analysis with Dirichlet boundary condition on the boundary of the hole $\mathcal{O}_\varepsilon$ was considered in [6] for Stokes equations, in [7] for quasi-Stokes equations and in [1] for Navier-Stokes equations.

In this section, we derive a topological asymptotic expansion for function j. It consists in studying the variation of j with respect to the presence of a small wells $\mathcal{O}_\varepsilon$ with a

Dirichlet boundary condition on Σ_ε. As mentioned above, we will derive an asymptotic expansion for j on the form:

$$j(\varepsilon) - j(0) \quad = \quad \frac{-2T\pi}{log(\varepsilon)}[u_0^N(z)v_0^N(z) + u_0^D(z)v_0^D(z)] + o(\frac{-1}{log(\varepsilon)}),$$

In this case the topological gradient is defined by:

$$g(z) = -2\pi T[u_0^N(z)v_0^N(z) + u_0^D(z)v_0^D(z)].$$

4 Numerical result

4.1 One-shot reconstruction algorithm

The identification procedure is a one shot algorithm based on the following steps:

– Step 1: solve the direct and adjoint problems,

– Step 2: compute the topological gradient g,

– Step 3: determine the negative local minima of g.

To test the efficiency of the proposed reconstruction process, different cases are studied. Wells are likely to be located at spots where the topological gradient g is most negative. The discretization of the direct problems (4) and (3) for $\varepsilon = 0$ with Dirichlet or Neumann conditions is based on triangular mesh and the finite element method. The numerical simulations are done using a $2D$ version of the software Comsol and Matlab [3].

The numerical tests are performed on a $20\,km \times 10\,km$ rectangular domain, with a homogeneous transmissivity $T = 0.001m^2s^{-1}$.
We define the relative errors as:

$$\tau_z = 100|\frac{\|OP_{ex}\| - \|OP_{id}\|}{\|OP_{ex}\|}|, \tag{9}$$

for position's identification: Where OP the position and O is the origin of the coordinate system. We denote by subscripts *ex* and *id* the exact and identified solutions.

4.2 Effects of mesh size

The aim of these first numerical experiment is to study the influence of mesh size on the results of the algorithm defined on the previous section. We consider a unique well centred at $z_{exact} = (0.3, 0.3)$ and having a radius ε.

Mesh	Mesh size h	Finite elements number	P_{id}	τ_z [%]
1	0.00625	1025	$(0.306, 0.305)$	1.8
2	0.02	736	$(0.31, 0.312)$	3.66
3	0.1	59	$(0.315, 0.315)$	5.16

Table 1: Effects of mesh size for the case of single well located at ($x = 0.3$, $y = 0.3$).

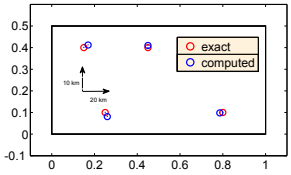
In Table 1, we give a summury of results obtained with different mesh size h. The finer is the mesh the smaller is the error.

4.3 A case of four wells

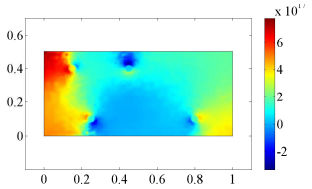
We identify four wells with Neumann (Case 1) and Dirichlet (Case 2) conditions on the boundary of the wells. Computed positions and the relative errors are shown in Table 2. In both cases, errors indicate that we have a good identification. We observe from the Figure 2 that the region where the most negative gradient is located in the vicinity of exact wells' position.

Identified \ Exact	$(0.25, 0.10)$	$(0.45, 0.40)$	$(0.80, 0.10)$	$(0.15, 0.40)$
Case 1	$(0.26, 0.81)$	$(0.45, 0.41)$	$(0.78, 0.09)$	$(0.17, 0.41)$
τ_z [%]	4.29	2	2.82	4.66
Case 2	$(0.24, 0.08)$	$(0.45, 0.41)$	$(0.80, 0.11)$	$(0.16, 0.41)$
τ_z [%]	4.23	3.23	3.68	3.95

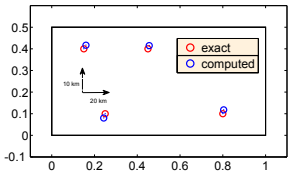
Table 2: Exact and computed wells' locations and corresponding relative errors.



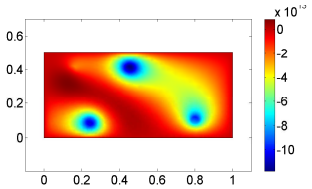
(a) Neumann Condition on well's boundary.



(b) Neumann Condition on well's boundary.



(c) Dirichlet Condition on well's boundary.



(d) Dirichlet Condition on well's boundary.

Figure 2: (a) and (c) represent the exact and estimated positions, (b) and (d) are the topological gradient distribution.

4.4 Sensitivity to the relative position

In the third case, we test the sensitivity to the relative position of two wells. We consider two wells separated by a variable distance d, and we compute the relative error for each distance (see Table 3). One can observe that if the wells are "well separated" (far from each other), the wells' locations are well identified, but when the distance between the wells decreases, the identification process is less accurate.

	d	0.63	0.53	0.4	0.25	0.18	0.15
Case 1	τ_z [%]	2.4	3.5	5.22	9.48	11.37	14.4
Case 2	τ_z [%]	2.1	3.22	4.84	9.16	10.59	14.2

Table 3: Influence of the relative distance between the wells.

4.5 Effect of noisy data

In the hope to try later this algorithm on real experimental data, we are interested in this paragraph to the robustness of the algorithm with respect to noisy data. We consider the case of single well located at point $z_{exact} = (0.4, 0.2)$. Then, we disrupt the over-specified data on Γ by adding relative white noise with different noise level. The results are presented in the Table 4. Notice that for noise level less than 6% the method remains efficient and results are in good agreement with the exact ones, whereas, since the noise level is higher fictitious flaws show up (see Table 4).

	Noise level (%)	4	6	10
Case 1	Location	$(0.429, 0.197)$	$(0.438, 0.191)$	$(0.459, 0.178)$
Error	τ_z [%]	5.83	7.81	12.59
Case 2	Location	$(0.381, 0.192)$	$(0.368, 0.181)$	$(0.354, 0.171)$
Error	τ_z [%]	4.12	7.44	10.87

Table 4: Effects of various noise levels for the case of single well located at ($x = 0.4$, $y = 0.2$).

We consider the case of three well separated located at point $z_i,\ i = 1:3$ having the same radius ε, the coordinates of the wells are $z_1 = (0.15, 0.4)$, $z_2 = (0.8, 0.15)$ and $z_3 = (0.45, 0.4)$. Then, we disrupt the overspecified data on Γ by adding relative white noise with different noise level.

	$Noise\ level$ (%)	4	6	10
Cas I	P_{id}^1	$(0.131, 0.42)$	$(0.123, 0.432)$	$(0.093, 0.476)$
	τ_z [%]	2.98	5.14	13.52
	P_{id}^2	$(0.819, 0.159)$	$(0.829, 0.169)$	$(0.839, 0.175)$
	τ_z [%]	2.5	4	5.29
	P_{id}^3	$(0.455, 0.444)$	$(0.462, 0.44)$	$(0.475, 0.475)$
	τ_z [%]	5.59	5.96	11.57
Cas II	P_{id}^1	$(0.136, 0.412)$	$(0.128, 0.425)$	$(0.102, 0.456)$
	τ_z [%]	1.56	3.89	9.37
	P_{id}^2	$(0.823, 0.165)$	$(0.85, 0.18)$	$(0.89, 0.21)$
	τ_z [%]	3.12	6.74	12.34
	P_{id}^3	$(0.459, 0.401)$	$(0.465, 0.42)$	$(0.472, 0.439)$
	τ_z [%]	1.32	4.07	7.06

Table 5: Effects of various noise levels for three wells.

One can note that for less than 6% of noise the wells are very well located wheras for a noise great than 10% it will be difficult to locate their positions (see Table 5).

5 Conclusion

In this work a new procedure for location of wells from overspecified boundary data based on the minimization of a misfit function type for identifying the positions of the wells. We develop an identification process related to the choice of type of boundary condition on well boundary.

The developed algorithm is fast since it a one shot algorithm. The method seems relevant in all the tested cases: multiplicity of wells with different position and noisy data. However, it is important to note that the present inverse problem is very sensitive to the quantity and quality of *"available"* data. The more over-specified data are available, the better are the recovered boundary data.

References

[1] S. AMSTUTZ, "The topological asymptotic for Navier Stokes equations", *ESAIM, Cont. Optim. Cal. Var*, vol. 3, p. 401-425, 2005.

[2] S. AMSTUTZ, I. HORCHANI, M. MASMOUDI, "Crack detection by the topological gradient methods", *Control and Cybernetics*, vol. 34(1), p. 81-101, 2005.

[3] "COMSOL Multiphysics Modeling Guide", *COPYRIGHT by COMSOL AB*, , 1998-2008.

[4] M. DE MARSILY, "Cours d'HydrogÃ©ologie", *UniversitÃ© de Paris 5*, 2004.

[5] S. GARREAU, PH. GUILLAUME, M. MASMOUDI, "The topological asymptotic for pde systems: the elasticity case", *SIAM J. Cont. Optim*, vol. 39, p. 1756-1778, 2001.

[6] PH. GUILLAUME, K. SID IDRIS, "Topological sensitivity and shape optimization for the Stokes equations", *SIAM J. Cont. Optim*, vol. 43, p. 1-31, 2004.

[7] M. HASSINE, M. MASMOUDI, "The topological sensitivity analysis for the Quasi-Stokes problem", *ESAIM, COCV J*, vol. 10, p. 478-504, 2004.

[8] I. KALLEL, "Analyse de sensibilité topologique pour l'opérateur de Laplace anisotrope.", *École Supérieur des Sciences et de Technologie, memoire de master* , 2012.

[9] M. MASMOUDI, J. POMMIER, B. SAMET, "The topological asymptotic expansion for the Maxwell equations and some applications", *J. Inverse Problems*, vol. 21(2), p. 547-564, 2005.

[10] M. MASMOUDI, "The topological asymptotic, in: Computational Methods for Control Applications (R. Glowinski, H. Karawada and J. Periaux, eds)", *GAKUTO Internat. Ser. Math. Sci. Appl*, vol. 16, p. 53-72, 2001.

[11] B. SAMET, S. AMSTUTZ, M. MASMOUDI, "Topological sensitivity analysis", *SIAM J. Control*, vol. 2(5), p. 1523-1544, 2003.

[12] A. SCHUMACHER, "Topologieoptimisierung von Bauteilstrukturen unter Verwendung von Lopchpositionierungkrieterien, Thesis", *Universitat-Gesamthochschule-Siegen*, 1995.

[13] J. SOKOLOWSKIA. ZOCHOWSKI, "On the topological derivative in shape optimization", *SIAM J. Control*, vol. 37, p. 1241-1272, 1999.

CARI 2016

Data assimilation for coupled models

Toward variational data assimilation for coupled models : first experiments on a diffusion problem

Rémi Pellerej[1], Arthur Vidard[2], Florian Lemarié[3]

Inria, Univ. Grenoble-Alpes, CNRS, LJK, F-38000 Grenoble, France
[1]remi.pellerej@imag.fr
[2]arthur.vidard@imag.fr
[3]florian.lemarie@inria.fr

ABSTRACT. Nowadays, coupled models are increasingly used in a wide variety of fields including weather forecasting. We consider the problem of adapting existing variational data assimilation methods to this type of application while imposing physical constraints at the interface between the models to be coupled. We propose three data assimilation algorithms to address this problem. The proposed algorithms are distinguished by their choice of cost function and control vector as well as their need to reach convergence of the iterative coupling method (the Schwarz domain decomposition method is used here). The performance of the methods in terms of computational cost and accuracy are compared using a linear 1D diffusion problem.

RÉSUMÉ. De nos jours, les modèles couplés sont de plus en plus utilisés dans de nombreux domaines, dont les prévisions météorologiques. Nous essayons ici d'adapter les méthodes courantes d'assimilation de données variationelles à ce type d'applications tout en imposant des contraintes physiques entre les deux modèles couplés. Nous proposons trois méthodes d'assimilation de données pour ce problème. Les différents algorithmes se distinguent par le choix de leur fonction coût, de leur vecteur de contrôle et du nombre d'itérations de couplage nécessaires (nous utilisons les méthodes de Schwarz pour coupler nos modèles). Ces méthodes sont comparées dans le cadre d'un problème linéaire de diffusion 1D en analysant leur coût de calcul et la qualité de leur analyse.

KEYWORDS : Coupled data assimilation, Schwarz methods, Optimal control

MOTS-CLÉS : Assimilation de données couplée, Méthodes de Schwarz, Contrôle optimal

1. Introduction

In the context of operational meteorology and oceanography, forecast skills heavily rely on proper combination of model prediction and available observations via data assimilation techniques. Historically, numerical weather prediction is made separately for the ocean and the atmosphere in an uncoupled way. However, in recent years, fully coupled ocean-atmosphere models are increasingly used in operational centers to improve the reliability of seasonal forecasts and tropical cyclones predictions. For coupled problems, the use of separated data assimilation schemes in each medium is not satisfactory since the result of such assimilation process is generally inconsistent across the interface, thus leading to unacceptable artefacts [4]. Hence, there is a strong need for adapting existing data assimilation techniques to the coupled framework, as initiated in [5]. In this paper, three general data assimilation algorithms, based on variational data assimilation techniques [3], are presented and applied to a simple coupled problem. The dynamical equations of this problem are coupled using an iterative Schwarz domain decomposition method [1]. The aim is to properly take into account the coupling in the assimilation process in order to obtain a coupled solution close to the observations while satisfying the physical conditions across the air-sea interface. The paper is organized as follows. The model problem and coupling strategy are given in Sec. 2. In Sec. 3 we briefly recall some theoretical aspects of variational data assimilation techniques, and we introduce and discuss three algorithms to solve coupled constrained minimization problems. The performance of the proposed schemes are illustrated by numerical experiments in Sec. 4.

2. Model problem and coupling strategy

We consider a problem defined on $\Omega = \mathbb{R}$. We decompose Ω in two nonoverlapping subdomains Ω_1 and Ω_2 with an interface $\Gamma = \{z = 0\}$. A model is defined on each space-time domain $\Omega_d \times [0, T]$ $(d = 1, 2)$ thanks to a differential operator $\mathcal{L}_d$ which acts on the variable u_d. The problem is to couple the two models at their interface Γ. To do so, we introduce the operators $\mathcal{F}_d$ and $\mathcal{G}_d$ which define the interface conditions. Those operators must be chosen to satisfy the required consistency on Γ. We propose to use a global-in-time Schwarz algorithm (a.k.a. Schwarz waveform relaxation, see [1] for a review) to solve the corresponding coupling problem. This method consists in solving iteratively each model on their respective space-time subdomain using the interface conditions on Γ computed during the previous iteration. For a given initial condition $u_0 \in H^1(\Omega_1 \cup \Omega_2)$ and *first-guess* $u_1^0(0, t)$, the corresponding coupling algorithm reads

$$\left\{\begin{array}{rl} \mathcal{L}_2 u_2^k = f_2 & \text{on } \Omega_2 \times T_W \\ u_2^k(z,0) = u_0(z) & z \in \Omega_2 \\ \mathcal{G}_2 u_2^k = \mathcal{G}_1 u_1^{k-1} & \text{on } \Gamma \times T_W \end{array}\right. \quad \left\{\begin{array}{rl} \mathcal{L}_1 u_1^k = f_1 & \text{on } \Omega_1 \times T_W \\ u_1^k(z,0) = u_0(z) & z \in \Omega_1 \\ \mathcal{F}_1 u_1^k = \mathcal{F}_2 u_2^k & \text{on } \Gamma \times T_W \end{array}\right. \tag{1}$$

where k is the iteration number, $T_W = [0, T]$, and $f_d \in L^2(0, T; L^2(\Omega_d))$ is a given right-hand side. At convergence, this algorithm provides a mathematically strongly coupled solution which satisfies $\mathcal{F}_1 u_1 = \mathcal{F}_2 u_2$ and $\mathcal{G}_2 u_2 = \mathcal{G}_1 u_1$ on $\Gamma \times T_W$. The convergence speed of the method greatly depends on the choice for $\mathcal{F}_d$ and $\mathcal{G}_d$ operators, and the choice of the *first-guess*. Note that in this paper, for the sake of simplicity, we restrict ourselves to linear differential operators for $\mathcal{L}_d$, $\mathcal{G}_d$, and $\mathcal{F}_d$, and to the multiplicative form of the Schwarz method where each model is run sequentially.

3. Data assimilation

Let us now suppose that some discrete estimates $\mathbf{y}$ of the solution to problem (1) are available over an irregular set of points in the interval $\Omega \times T_W$. In this context we are interested in using a data assimilation (DA) procedure to account for this additional source of information. For the present study we use the variational methods of DA, based on optimal control theory. Our aim is to evaluate a set of parameter $\mathbf{x}_0$, including for instance the initial condition u_0 of problem (1), through the minimization of a cost function $J(\mathbf{x}_0)$ ($\mathbf{x}_0$ is the control vector) which quantifies in some sense the misfit between the observations $\mathbf{y}$ and the model prediction. This minimization requires the gradient of $J(\mathbf{x}_0)$, which can be computed using adjoint methods [3].

3.1. Uncoupled variational data assimilation

We first briefly describe the variational DA approach in the uncoupled case to introduce the necessary notations. The control vector is restricted to subdomain Ω_d and is noted $\mathbf{x}_{0,d} = u_0|_{z \in \Omega_d}$. The optimal control problem amounts to find $\mathbf{x}^a_{0,d}$, the *analysed state*, which best fit observations $\mathbf{y}$ and a previous estimate of the initial state $\mathbf{x}^b_d$ called the *background*. Noting H the observation operator that goes from model space to the observations space and $\mathbf{x}_d = u_d$ the state vector, the cost function to minimize reads

$$J(\mathbf{x}_{0,d}) = \overbrace{\left\langle \mathbf{x}_{0,d} - \mathbf{x}^b_d, \mathbf{B}^{-1}(\mathbf{x}_{0,d} - \mathbf{x}^b_d) \right\rangle_{\Omega_d}}^{J^b(\mathbf{x}_{0,d})} + \overbrace{\int_0^T \left\langle \mathbf{y} - H\left(\mathbf{x}_d\right), \mathbf{R}^{-1}(\mathbf{y} - H\left(\mathbf{x}_d\right)) \right\rangle_{\Omega_d} \mathrm{dt}}^{J^o(\mathbf{x}_{0,d})} \tag{2}$$

where $\mathbf{R}$ is the covariance matrix associated to observation errors, $\mathbf{B}$ is the background error covariance matrix, and $\langle \cdot \rangle_\Sigma$ is the usual Euclidian inner product on a spatial domain Σ. Obviously, if the DA process is done separately on each subdomain (with prescribed boundary conditions on the interface Γ), the initial condition $u_0 = (\mathbf{x}^a_{0,1}, \mathbf{x}^a_{0,2})^T$ obtained on Ω does not satisfy the interface conditions, hence $u_0 \notin H^1(\Omega)$ and well-posedness of the coupled problem is no longer guaranteed. In practice this type of imbalance in the initial condition can severely damage the forecast skills of coupled models [4].

3.2. Toward a coupled variational data assimilation

Our objective is now to properly take into account the coupling in the assimilation process. To do this, we introduce in this section three types of variational DA algorithms whose aim is to provide a solution close to the observations while satisfying the interface conditions on Γ; or at least a weak form of it. The key properties of those algorithms are summarized in Tab. 1.

Full Iterative Method (FIM)

A first possibility is to consider a monolithic view of the problem by ignoring the presence of an interface in the assimilation process. In this case the state vector is $\mathbf{x_0} = u_0(z),\ z \in \Omega$ and for each model integration we iterate the models on Ω_1 and Ω_2 till convergence of the Schwarz algorithm. If we note k_{cvg} the number of iterations to satisfy the stopping criterion, the cost function for the FIM is

$$J(\mathbf{x}_0) = J^b(\mathbf{x}_0) + \int_0^T \left\langle \mathbf{y} - H\left(\mathbf{x}^{\mathrm{cvg}}\right), \mathbf{R}^{-1}(\mathbf{y} - H\left(\mathbf{x}^{\mathrm{cvg}}\right)) \right\rangle_\Omega \mathrm{dt} \tag{3}$$

where $\mathbf{x}^{\mathrm{cvg}} = (u_1^{k_{\mathrm{cvg}}}, u_2^{k_{\mathrm{cvg}}})^T$. Since the first-guess u_1^0 in (1) is updated after each minimization iteration with the converged solution obtained during the previous model integration, the Schwarz algorithm will converge more rapidly over the minimization iteration. It can readily be seen that cost function (3) is identical to the cost function we would use for an uncoupled problem defined on Ω. The solution provided by this approach is strongly coupled. Note that the FIM requires the adjoint of the strongly coupled model (1) which can be tedious to derive. The main drawback of this method is that it possibly requires a very large number of Schwarz iterations since we systematically iterate till convergence.

Truncated Iterative Method (TIM)

In order to improve the computational cost of the FIM algorithm, we propose to truncate the Schwarz iterations in the direct and adjoint model after $k_{\max}$ iterations, with $k_{\max} < k_{\mathrm{cvg}}$. Because we do not iterate till convergence, the coupled solution strictly satisfies only one of the two interface conditions, for example we would have $\mathcal{F}_1 u_1 = \mathcal{F}_2 u_2$ and $\mathcal{G}_2 u_2 \neq \mathcal{G}_1 u_1$ if iteration $k_{\max}$ is done first on Ω_2 and then on Ω_1. As proposed by [2] in the context of river hydraulics, a convenient way to propagate the information from one subdomain to the other during the minimization iterations is to use an extended cost function which includes the misfit in the interface conditions. The idea behind this approach is to enforce a weak coupling within the minimization iterations. The control vector $\mathbf{x}_0 = (u_0(z), u_1^0(0,t))^T$ now includes the first-guess on the interface and the cost function reads

$$J(\mathbf{x}_0) = J^b(\mathbf{x}_0) + \int_0^T \left\langle \mathbf{y} - H\left(\mathbf{x}^{\mathrm{trunc}}\right), \mathbf{R}^{-1}(\mathbf{y} - H\left(\mathbf{x}^{\mathrm{trunc}}\right)) \right\rangle_\Omega \mathrm{dt} + J^s \tag{4}$$

where $J^s = \alpha_{\mathcal{F}} \|\mathcal{F}_1 u_1(0,t) - \mathcal{F}_2 u_2(0,t)\|^2_{[0,T]} + \alpha_{\mathcal{G}} \|\mathcal{G}_1 u_1(0,t) - \mathcal{G}_2 u_2(0,t)\|^2_{[0,T]}$ with $\|a\|^2_\Sigma = \langle a, a \rangle_\Sigma$ and $\mathbf{x}^{\mathrm{trunc}} = (u_1^{k_{\max}}, u_2^{k_{\max}})^T$. As mentioned above, if the model is integrated

first on Ω_2 and then on Ω_1 we have $\mathcal{F}_1 u_1 = \mathcal{F}_2 u_2$ and only $\alpha_{\mathcal{G}}$ is a relevant parameter in the penalization of the interface conditions in (4). Note that, unlike FIM, the first-guess is part of the control vector here, but this method still requires the adjoint of the coupling. Since the first-guess u_1^0 is updated at the end of each minimization iteration, we can expect that we will converge toward a good approximation of the strongly coupled solution.

Coupled Assimilation Method with Uncoupled models (CAMU)

The last possibility we propose to investigate is to suppress the coupling iterations and rely only on the minimization iterations to weakly couple the two models. This approach only requires the adjoint of each individual model but not the adjoint of the coupling as for the previous algorithms. The control vector is $\mathbf{x}_0 = (\mathbf{x}_{0,1}, \mathbf{x}_{0,2})^T$ with $\mathbf{x}_{0,d} = (u_0|_{z\in\Omega_d}, u_d^0(0,t))$. The corresponding cost function is

$$J(\mathbf{x}_0) = \left\{ \sum_{d=1}^{2} (J^b(\mathbf{x}_{0,d}) + J^o(\mathbf{x}_{0,d})) \right\} + J^s.$$

It is straightforward to see that this algorithm provides only a weakly coupled solution. We proceed only to one iteration of the models (which can be run in parallel) with boundary conditions on Γ provided by the term $u_d^0(0,t)$ taken from the control vector. Note that both parameters $\alpha_{\mathcal{F}}$ and $\alpha_{\mathcal{G}}$ have an impact on the solution of the minimisation. In the next section the three DA algorithms presented so far are compared in terms of computational cost and accuracy.

Algo	Control vector	# of coupling iterations	extended cost function	Adjoint of the coupling	Coupling
FIM	$(u_0(z))$	k_{cvg}	no	yes	strong
TIM	$(u_0(z), u_1^0)^T$	$k_{\max}$	yes	yes	$\sim$strong
CAMU	$(u_0(z), u_1^0, u_2^0)^T$	0	yes	no	weak

Table 1. *Overview of the properties of the coupled variational DA methods described in Sec. 3.2. Notations are consistent with those introduced in the text.*

4. Application to a 1D diffusion problem

In this section, previous algorithms are applied on a 1D diffusion problem. We, thus, consider $\mathcal{L}_d = \partial_t + \nu_d \partial_z^2$ in (1) with $\nu_1 \neq \nu_2$ the diffusion coefficients in each subdomain. The computational domain is $\Omega =]-L_1, L_2[$ with $L_1, L_2 \in \mathbb{R}^{+*}$. We choose the interface operators on Γ to obtain a Dirichlet-Neumann algorithm, i.e. $\mathcal{F}_d = \nu_d \partial_z$ and $\mathcal{G}_d = \mathrm{Id}$.

We consider the analytical solution $u_d^\star$, and the corresponding right hand side $f_d = \mathcal{L}_d u_d^\star$, of the coupled problem on each subdomain as :

$$u_d^\star(z,t) = \frac{U_0}{4} e^{-\frac{|z|}{\alpha_d}} \left\{ 3 + \cos^2\left(\frac{3\pi t}{\tau}\right) \right\} \qquad \text{on } \Omega_d \times T_W. \tag{5}$$

where $U_0 = 20\ ^\circ\mathrm{C}$ and $\tau = 22\ \mathrm{h}$. Note $\alpha_1\nu_2 = \alpha_2\nu_1$ is required to ensure the proper regularity of the coupled solution across the interface Γ. To satisfy this constraint we choose $\alpha_1 = 4\ \mathrm{km}$, $\alpha_2 = 0.4\ \mathrm{km}$, $\nu_1 = 1\ \mathrm{m}^2/\mathrm{s}$, $\nu_2 = 0.1\ \mathrm{m}^2/\mathrm{s}$. The model problem (1) is discretized using a backward Euler scheme in time and a second-order scheme in space. The resolution in each subdomain is $\Delta z = 20\ \mathrm{m}$ with $L_1 = L_2 = 1\ \mathrm{km}$ and the time-step is $\Delta t = 180\ \mathrm{s}$. The total simulation time is $T = 12\ \mathrm{h}$ and we start the Schwarz iterations with a random first-guess.

For the assimilation experiments, we consider that the true-state $\mathbf{x}^t$ is the solution of the Schwarz algorithm (1) while the background $\mathbf{x}^b$ corresponds to the solution obtained with a biased initial condition. In both cases, the Schwarz algorithm converges in $k_{\mathrm{cvg}} = 50$ iterations with a tolerance $\epsilon = 10^{-6}$. Some observations $\mathbf{y}$ of the true-state are generated such that $\mathbf{y} = H(\mathbf{x}^t)$, with H the observation operator. The observation and background errors covariance matrices are considered diagonal such that $\mathbf{R} = 10\ \mathrm{Id}$ and $\mathbf{B} = 100\ \mathrm{Id}$. For the extended cost function we consider $\alpha_{\mathcal{F}} = \frac{\alpha_1}{\nu_1}\alpha_{\mathcal{G}}$ with different values of $\alpha_{\mathcal{G}}$. All the minimisation are done until convergence of a conjugate gradient algorithm with a stopping criterion $\| \nabla J(\mathbf{x}_0) \|_\infty < 10^{-5}$.

Single column observation experiment

For our experiments, we consider that observations are available in $\Omega \setminus \{\Gamma\}$ only at the end of the time-window (i.e. at $t = T$). In this case, the results obtained for different assimilation schemes are reported in table 2 where the performance of each scheme is presented in terms of the number of minimisation and models runs. Note that the computational cost of a given method is almost entirely dominated by the model integration. To evaluate the strength of the coupling we define an *interface imbalance indicator* which corresponds to the value of J_s at the end of the DA process, with $\alpha_{\mathcal{G}} = 0.01$ and $\alpha_{\mathcal{F}} = 40$. Values of J^s close to zero indicate that the analysed state is strongly coupled. In table 2, a root mean square error (RMSE) defined as $\sqrt{\mathbb{E}\left((\mathbf{x}^a - \mathbf{x}^t)^2\right)}$ on $\Omega \times T_W$ is also used to evaluate how much the analysed state is close to the true-state.

From table 2, we can first note that the FIM algorithm requires few minimisation iterations to obtain a low RMSE value and a strongly coupled analysed state ($J^s \sim 10^{-12}$). A drawback of this approach is a high computational cost (1169 models runs). Since in the TIM approach the coupling iterations are truncated and the first-guess u_1^0 is part of the control vector, we expect a reduced computational cost compared to FIM. It is however the case only if the J^s term is included in the cost function (i.e. $\alpha_{\mathcal{G}} \neq 0$ or $\alpha_{\mathcal{F}} \neq 0$), otherwise the TIM requires a very large number of models runs to reach an analysed state which is of a lesser quality than with FIM. On the one hand decreasing the

Algo	$\alpha_{\mathcal{G}}$	$\alpha_{\mathcal{F}}$	$k_{\max}$	# of minimisation iterations	# of models runs	Interface imbalance indicator	RMSE in °C
FIM	-	-	k_{cvg}	58	1169	$3.69\ 10^{-12}$	0.220
TIM	0	-	k_{cvg}	48	2016	$5.63\ 10^{-12}$	0.220
TIM	0	-	5	320	1600	$2.95\ 10^{-2}$	0.216
TIM	0	-	2	1521	3042	3.77	0.272
TIM	0.01	-	2	391	782	$9.25\ 10^{-7}$	0.217
TIM	0.01	-	1	350	350	$8.60\ 10^{-7}$	0.215
CAMU	0.01	40	0	1308	1308	$1.40\ 10^{-4}$	0.229
CAMU	0.001	4	0	268	268	$9.38\ 10^{-3}$	0.240
CAMU	0.0001	0.4	0	758	758	$3.30\ 10^{-1}$	0.327
Uncoupled	0	0	0	101	101	29.0	1.717

Table 2. *Results obtained for the three coupled variational DA methods described in Sec. 3.2 with observations available in* $\Omega \setminus \{\Gamma\}$ *at the end of the time-window.*

value of $k_{\max}$ increases the number of minimization iterations. Indeed, going to Schwarz convergence ($k_{\max} = k_{\text{cvg}}$) procures the best model solution, it then needs few minimisation iterations. However, for the next iteration, the background interface is given by the control vector rather than the previous converged estimate; therefore it requires again numerous Schwarz iterations. On the other hand, by reducing the $k_{\max}$ value, the number of Schwarz iterations is reduced and the update of the first-guess more significant, but the quality of the coupling is affected and this leads to a slower minimisation convergence. Here, a good compromise is to choose $k_{\max} = 5$. When taking J^s into account in TIM (i.e. for $\alpha_{\mathcal{G}} \neq 0$), it leads to a better analysed state with significantly less models runs. Smaller values of $k_{\max}$ provide a faster convergence of the algorithm. With $k_{\max} = 1$, which corresponds to a one-way coupling, it requires only 350 models runs to provide a good approximate of the strongly coupled solution ($J^s = 8.6\ 10^{-7}$, RMSE $= 0.215\,°\text{C}$). In this case, the interface condition $\mathcal{F}_1 u_1 = \mathcal{F}_2 u_2$ is imposed in a strong way in the coupling iterations while the other condition $\mathcal{G}_1 u_1 = \mathcal{G}_2 u_2$ is established in a weak way through J^s during the minimisation. For $k_{\max} > 1$ the interface condition $\mathcal{G}_1 u_1 = \mathcal{G}_2 u_2$ is also imposed in a strong way in the coupling iterations, and seems to conflict with the weak constraint from J^s. By considering uncoupled models in the CAMU algorithm, a proper choice for $\alpha_{\mathcal{G}}$ and $\alpha_{\mathcal{F}}$ to balance J^s and J^o in the cost function can lead to an efficient method (268 models runs). Too big values imply a more constrained cost function, which leads to more minimisation iterations. At the opposite, too small values do not constrain enough the interface and therefore produce poor model solutions. The analysed state shows a larger interface imbalance indicator compared to FIM and TIM, which confirms that CAMU provides a weakly coupled solution, but is significantly better than the uncoupled DA in that respect.

5. Conclusion and perspectives

We addressed in this paper the problem of variational data assimilation for coupled models. The aim of was to introduce coupled DA algorithms. In this context, a difficulty is to determine how to combine the two iterative processes at play, namely the Schwarz iterations in the coupling and the minimisation iterations in the DA problem. The proposed algorithms are distinguished by their choice of cost function and control vector as well as their need to reach convergence of the Schwarz coupling method. We showed that adding a physical constraint on the interface conditions in the cost function can have a beneficial effect on the performance of the method and allow to save coupling iterations. Moreover, an approach which only requires the adjoint of each individual model but not the adjoint of the coupling showed promising results. Since the objective is to apply such methods to ocean-atmosphere coupled models, increasingly complex models including physical parameterisations for subgrid scales will be considered in future work.

6. Acknowledgements

The work described in this article was supported by the ERA-CLIM2 project, funded by the European Union's Seventh Framework Programme under grant n°607029.

References

[1] M.J. Gander. Schwarz methods over the course of time. *Electron. Trans. Numer. Anal.*, 31:228–255, 2008. http://etna.math.kent.edu/vol.31.2008/pp228-255.dir.

[2] I.Y. Gejadze and J. Monnier. On a 2d 'zoom' for the 1d shallow water model: Coupling and data assimilation. *Comput. Methods Appl. Mech. and Engrg.*, 196(45–48):4628 – 4643, 2007.

[3] F.X. Le Dimet and O. Talagrand. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A*, 38A(2):97–110, 1986.

[4] D.P. Mulholland, P. Laloyaux, K. Haines, and M.A. Balmaseda. Origin and impact of initialization shocks in coupled atmosphere-ocean forecasts. *Mon. Wea. Rev.*, 143:4631–4644, 2015.

[5] P. Smith, A. Fowler, and A. Lawless. Exploring strategies for coupled 4d-var data assimilation using an idealised atmosphere-ocean model. *Tellus A*, 67(0), 2015.

Calcul numérique de solutions de l'équation de Schrödinger non linéaire faiblement amortie avec défaut

Laurent Di Menza [a] — Olivier Goubet [b], — Emna Hamraoui [b,c,*], — Ezzeddine Zahrouni [c,d]

[a] Laboratoire de Mathématiques de Reims (LMR) - EA 4535, U.F.R. Sciences Exactes et Naturelles, 51687 REIMS cedex 2, France
laurent.di-menza@univ-reims.fr

[b] Laboratoire Amiénois de Mathématique Fondamentale et Appliquée, CNRS UMR 7352, Faculté des Sciences, Université de Picardie Jules Verne, 80039 Amiens CEDEX 1, France
Olivier.Goubet@u-picardie.fr

[c] Unité de Recherche : Multifractals et Ondelettes, Faculté des Sciences de Monastir, Université de Monastir, 5019 Monastir,Tunisie
emna.hamraoui@gmail.com

[d] Faculté des Sciences Économiques et de Gestion de Nabeul, Université de Carthage, 8000 Nabeul, Tunisie
zahrouniez@gmail.com
[*] Corresponding author

RÉSUMÉ. Dans ce travail, on étudie numériquement l'influence d'un défaut ponctuel sur le comportement des solutions de l'équation de Schrödinger non linéaire faiblement amortie. Notre méthode numérique repose sur l'utilisation des couches PML (a Perfectly Matched Layer) pour les conditions aux limites, d'un schéma Cranck-Nicolson en temps et la méthode des différences finies en espace. On observe tout d'abord que le défaut décompose l'onde incidente en deux parties, une réfléchie et une transmise, dont les normes L^2 sont des fonctions décroissantes du temps. D'autre part, on trouve que le défaut peut jouer le rôle d'une barrière.

ABSTRACT. In this work, we study numerically how a single defect influences the behaviour of solutions of the weakly damped non linear Schrödinger equation. Our numerical method is based on a Crank-Nicolson scheme in the time, finite difference method in space including a Perfectly Matched Layer (PML) treatment for the boundary conditions. First, we observe that the defect splits the incident wave in two parts, one reflected and one transmitted. For each, the L^2-norm are decreasing functions with respect to time. More over, we find that the defect can be considered as a barrier.

MOTS-CLÉS : NLS faiblement amortie, masse de Dirac, couche parfaitement absorbante.

KEYWORDS : weakly damped NLS equation, Dirac potential, perfectly matched layer.

1 Introduction

Les équations de Schrödinger non linéaires (NLS) ont toujours fait l'objet d'études intensives dans la littérature sur des thèmes variés tels que l'explosion en temps fini ou la propagation d'états stationnaires. En optique non linéaire, elles sont utilisées pour décrire la propagation d'un faisceau lumineux intense (laser) dans une fibre optique [12]. Ici, plusieurs phénomènes peuvent participer à l'atténuation de la lumière dans la fibre et à la perte de l'énergie lumineuse [1, 11].

Ce travail porte sur deux causes de perte d'énergie. La première est la dispersion du signal. Dans ce cas, le phénomène est modélisé par l'équation NLS faiblement amortie

$$i\frac{\partial u}{\partial t} + i\gamma u + \frac{\partial^2 u}{\partial x^2} + |u|^2 u = 0, \tag{1}$$

où l'inconnue $u = u(t, x)$ est définie sur $\mathbb{R} \times \mathbb{R}^+$ à valeurs dans $\mathbb{C}$, avec γ une constante positive qui représente le paramètre d'amortissement. Le problème de Cauchy associé à (1) a été étudié par plusieurs auteurs dont J. Ginibre [6], T. Cazenave [2] et T. Kato [9].

La deuxième cause de perte d'énergie résulte de l'existence d'impuretés dans le milieu d'étude et qui sont la conséquence du mode de fabrication. Ici, notre problème est modélisé par l'équation NLS avec défaut ponctuel à l'origine

$$i\frac{\partial u}{\partial t} + Zu\delta_0 + \frac{\partial^2 u}{\partial x^2} + |u|^2 u = 0, \tag{2}$$

où Z est l'amplitude de défaut et δ_0 est la masse Dirac en zéro. L'existence et l'unicité de la solution du problème (2) à été étudiée par R. H Goodmana et all [7], ils ont prouvé que le problème est bien posé en $H^1(\mathbb{R})$.

Dans ce travail, on étudie numériquement l'influence du défaut sur le comportement des solutions de l'équation de NLS faiblement amortie. Notre problème est donné par l'équation de Schrödinger suivante

$$i\frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial x^2} + i\gamma u + Zu\delta_0 + |u|^2 u = 0. \tag{3}$$

L'objectif de notre étude est l'analyse de l'influence du défaut sur le comportement des solutions de l'équation NLS faiblement amortie (pour γ fixé), et en particulier sur la décroissance en temps de la norme L^2 des solutions numériques calculées.

2 Formulation forte du problème modèle

2.0.0.1 Mise en place des PML

Soit Ω le domaine de calcul, $\Omega =]x_g, x_d[\subset \mathbb{R}$ tel que $0 \in]x_g, x_d[$. Notre problème s'écrit

$$\begin{cases} i\dfrac{\partial u}{\partial t} + \dfrac{\partial^2 u}{\partial x^2} + i\gamma u + Zu\delta_0 + |u|^2 u = 0, & x \in \Omega, t > 0, \\ u(0, x) = u_0(x), & x \in \Omega, \\ u(t, x_d) = u(t, x_g) = 0, & t > 0. \end{cases} \tag{4}$$

L'utilisation d'une condition aux limites classique aux points x_d et x_g (par exemple la condition de type Dirichlet homogène) provoque la réflexion de la solution à l'intérieur

du domaine du calcul, et perturbe ainsi la solution numérique recherchée. D'où l'utilisation des couches PML (a Perfectly Matched Layer), qui sont des bandes ajoutées autour du domaine du calcul et destinées à absorber les ondes proches du bord sans les réfléchir [13].
Soit L la largeur de la bande PML. Le nouveau domaine de calcul est défini par $]x_g - L, x_d + L[$, que l'on note aussi $]xg_{pml}, xd_{pml}[$.

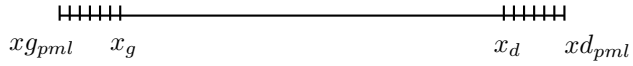


Figure 1 – Domaine et PML considérés en 1D.

La nouvelle formulation PML est basée sur la transformation suivante de x en coordonnées complexes (on se restreint au cas $+x$)

$$x' = x + R\int_{x_d}^{x} \sigma(s)ds,$$

où $R \in \mathbb{C}$ et σ est une fonction positive, continue et nulle à l'extérieur de $[x_r + \infty)$ appelée coefficient d'absorption. Soit $v = u(t, x')$ une solution du problème (4), v vérifie l'équation

$$i\frac{\partial v}{\partial t} + \frac{\partial^2 v}{\partial x'^2} + i\gamma v + Zv\delta_0 + |v|^2 v = 0, \tag{5}$$

l'équation (5) se réécrit en fonction de la dérivée spatiale en x

$$i\frac{\partial v}{\partial t} - \frac{1}{1+R\sigma}\frac{\partial}{\partial x}(\frac{1}{1+R\sigma}\frac{\partial v}{\partial x}) + i\gamma v + Zv\delta_0 + |v|^2 v = 0,\ x \in]xg_{pml}, xd_{pml}[, \tag{6}$$

avec des conditions de types Dirichlet homogène sur les deux points limites $\{xg_{pml}, xd_{pml}\}$.
Pour l'expression de σ, différents choix existent dans la littérature, dans notre étude on prend le choix des fonctions quadratiques [13]

$$\sigma = \begin{cases} \sigma_0(x - x_g)^2, & xg_{pml} < x < x_g \\ 0, & x_g < x < x_d \\ \sigma_0(x - x_d)^2, & x_d < x < xd_{pml} \end{cases} \tag{7}$$

où σ_0 est une constante positive.

2.0.0.2 Modélisation de la masse de Dirac

Pour la modélisation de la masse de Dirac, on utilise l'approche donnée dans les travaux de Le Coz et al. [10], J. Holmer et C. Liu [8] et aussi dans [5]. Elle traduit la présence du défaut par une condition de transmission en zéro

$$\begin{aligned} &i\frac{\partial u}{\partial t} + \frac{1}{(1+R\sigma)^2}\frac{\partial^2 u}{\partial x^2} - \frac{R\sigma'}{(1+R\sigma)^3}\frac{\partial u}{\partial x} + i\gamma u + |u|^2 u = 0, \quad x \in]xg_{pml}, xd_{pml}[,\ x \neq 0,\ t > 0, \\ &\frac{\partial u}{\partial x}(t, 0^+) - \frac{\partial u}{\partial x}(t, 0^-) = -Zu(t, 0), \qquad\qquad t > 0. \end{aligned} \tag{8}$$

3 Discrétisation du problème

Dans cette section, on utilise la méthode des différences finies pour l'approximation en espace et le schéma de Crank-Nicolson pour la discrétisation en temps. L'approximation numérique du terme défaut est donnée dans [10].

3.1 Semi-discrétisation en espace

Pour l'approximation en espace on utilise la méthode des différences finies. Soit Δx le pas de discrétisation, $\forall x_j \in]xg_{pml}, xd_{p_ml}[$, $x_j = xg_{pml} + j\ \Delta x$, $j = 1 : N$.

$$\frac{\partial u}{\partial x}(t, 0^+) = \frac{4u(t, \Delta x) - u(t, 2\Delta x) - 3u(t, 0)}{2\Delta x}. \tag{9}$$

$$\frac{\partial u}{\partial x}(t, 0^-) = \frac{u(t, -2\Delta x) - 4u(t, 0 - \Delta x) + 3u(t, 0)}{2\Delta x}. \tag{10}$$

Par conséquent, on résout en 0 un schéma d'ordre deux

$$4u(t, \Delta x) - u(t, 2\Delta x) + (2\ Z\ \Delta x - 6)u(t, 0) - u(t, 0 - 2\Delta x) + 4u(t, -\Delta x) = 0.$$

Pour $x_j \neq 0$, on utilise un schéma centré pour la approximation de la dérivé première

$$\frac{\partial u}{\partial x}(t, x_j) \approx \frac{u(t, x_{j+1}) - u(t, x_{j-1})}{2\Delta x},$$

et un schéma d'ordre deux pour l'approximation de la dérivée seconde

$$\frac{\partial^2 u}{\partial x^2}(t, x_j) \approx \frac{u(t, x_{j+1}) - 2u(t, x_j) + u(t, x_{j-1})}{(\Delta x)^2}.$$

3.2 Semi-discrétisation en temps

Pour l'approximation en temps on utilise un schéma de Crank-Nicolson. Soit Δt le pas discrétisation en temps, $t^n = n\ \Delta t$. On note $u_j^n \approx u(t^n, x_j)$. Pour tout $j \neq j_\delta$ où $x_{j_\delta} = 0$, on résout un système non linéaire

$$\begin{aligned}
&i\ \frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{1}{2}\frac{1}{(1 + R\sigma_j)^2}\frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{(\Delta x)^2} - \frac{1}{2}\ \frac{R\sigma_j'}{(1 + R\sigma_j)^3}\frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{2\Delta x} \\
&+i\frac{\gamma}{2}u_j^{n+1} = -\frac{1}{2}\frac{1}{(1 + R\sigma_j)^2}\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\Delta x)^2} + \frac{1}{2}\ \frac{R\sigma_j'}{(1 + R\sigma_j)^3}\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} \\
&-i\frac{\gamma}{2}u_j^n - \frac{1}{2}(|u_j^{n+1}|^2 u_j^{n+1} + |u_j^n|^2 u_j^n).
\end{aligned}$$

En revanche, en $j = j_\delta$ (au point défaut) on résout l'équation linéaire suivante

$$\begin{aligned}
&\frac{\Delta t}{2}\ (4u_{j_\delta+1}^{n+1} - u_{j_\delta+2}^{n+1} + (2\ Z\ \Delta x - 6)\ u_{j_\delta}^{n+1} - u_{j_\delta-2}^{n+1} + 4u_{j_\delta-1}^{n+1}) \\
&\qquad = -\frac{\Delta t}{2}\ (4u_{j_\delta+1}^n - u_{j_\delta+2}^n + (2\ Z\ \Delta x - 6)\ u_{j_\delta}^n - u_{j_\delta-2}^n + 4u_{j_\delta-1}^n).
\end{aligned}$$

Comme le schéma contient des termes non linéaires, l'implémentation nécessite la résolution d'un problème de point fixe à chaque pas du temps.

4 Résultats numériques

Dans notre étude, on s'intéresse aux solutions issues d'une donnée initiale voyageuse de type gaussienne

$$u_0 = q\exp(ikx)\exp(-(x-x_0)^2).$$

Pour les simulations numériques on pose que $t \in [0,T]$, avec $T=1$. On prend $x_g = -25$, $x_d = 25$, $L = 2$, $\Delta x = 10^{-2}$, $\Delta t = 10^{-4}$, $R = exp(\frac{i\pi}{4})$, et $\sigma_0 = 1$. Dans tous les cas test, on prend la même donnée initiale

$$u_0 = \exp(i10x)\exp(-(x-x_0)^2),$$

avec $x_0 = -5$.

4.1 Cas test 1 : sans défaut $(Z=0)$

On rappelle que l'équation (1) vérifie l'estimations à priori suivante, établi par la masse

$$\frac{\partial}{\partial t}\|u\|_{L^2}^2 = -2\gamma\|u\|_{L^2}^2, \tag{11}$$

L'égalité (11) montre que la norme L^2 de la solution de l'équation NLS faiblement amortie est une fonction décroissante du temps. Dans la figure 2, on visualise pour différentes valeurs d'amortissement la variation de la masse au cours du temps.

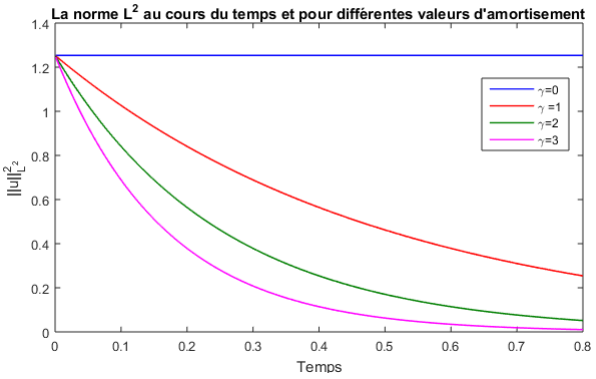


Figure 2 – La norme L^2 au cours du temps pour différentes valeurs d'amortissement.

En absence de dissipation ($\gamma = 0$), on observe bien la conservation de la masse qui est un invariant pour l'équation NLS. Pour $\gamma > 0$, on remarque que plus la valeur de l'amortissement est grande, plus la décroissance de la masse est rapide. Maintenant, afin de mieux comprendre l'influence du défaut, on visualise le comportement de la solution de l'équation NLS faiblement amortie pour $\gamma = 1$.

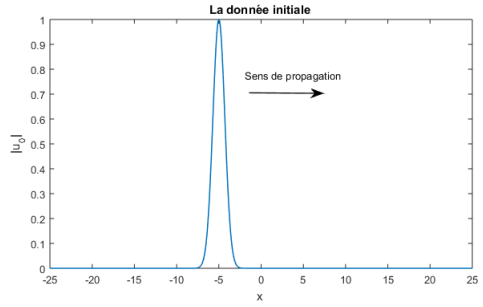


Figure 3 – La donnée initiale.

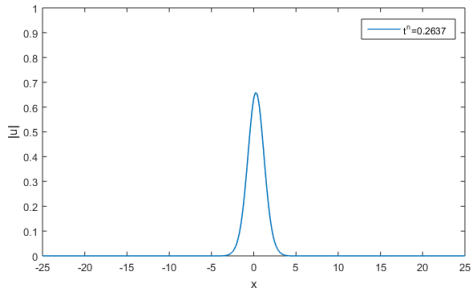


Figure 4 – La solution numérique calculée pour $\gamma = 1$ à l'instant $t^n = 0.2637$.

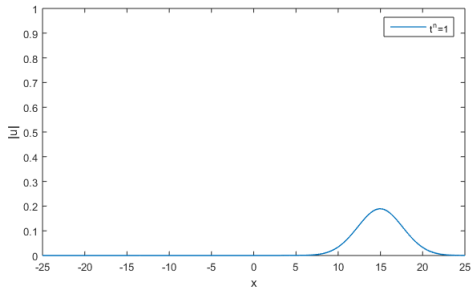


Figure 5 – La solution solution numérique calculée pour $\gamma = 1$ à l'instant final.

On observe que la solution numérique calculée se déplace vers les x positifs, ce qui est en bonne concordance avec le choix de k strictement positif (paramètre de la donnée initiale, $k = 10$). On remarque aussi l'influence du paramètre d'amortissement sur la décroissance de la norme L^2 : Soient $M_i = ||u_0||^2_{L^2}$ la masse de la donnée initiale, $M_n = ||u^n||^2_{L^2}$ la masse de la solution à l'instant n, on a

$$M_i = 1.2533,\ M_T = 0.1696, \text{ et } M_T << M_i.$$

4.2 Cas test 2 : avec défaut ($Z = 10$)

Dans ce cas test, on analyse l'influence de la présence du défaut sur le comportement dynamique de la solution. On prend un défaut localisé à l'origine, d'amplitude $Z = 10$.

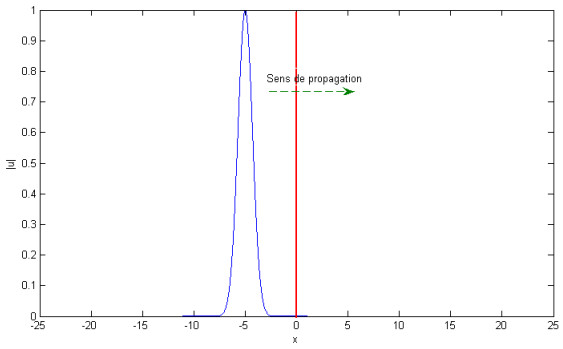


Figure 6 – La donnée initiale.

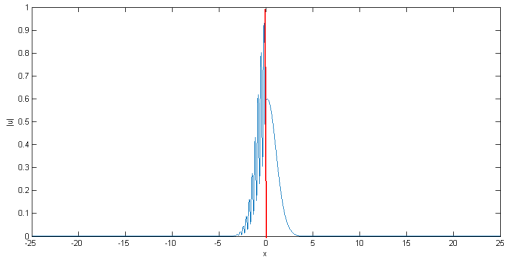


Figure 7 – La solution numérique calculée pour $\gamma = 1$, $Z = 10$ à l'instant $t^n = 0.2553$.

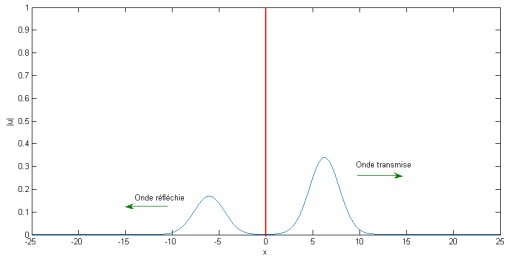


Figure 8 – La solution numérique calculée pour $\gamma = 1$, $Z = 10$ à l'instant $t^n = 0.5592$.

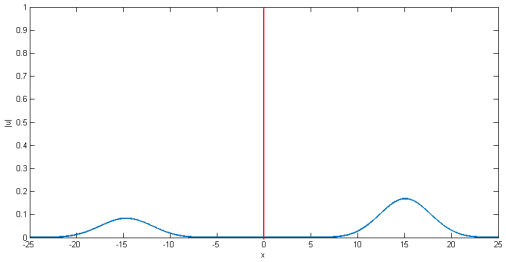


Figure 9 – La solution numérique calculée pour $\gamma = 1$, $Z = 10$ à l'instant final.

On observe que la solution se décompose en deux parties, une onde transmise et une onde réfléchie à la suite du passage par le défaut. On note t_δ le temps lors de l'interaction avec le défaut, $\forall t > T_\delta$, on a $u = u_{\mathbf{t}} + u_{\mathbf{r}}$, où $u^{\mathbf{t}}$ est l'onde transmise qui représente dans notre cas test la restriction de u sur $]0, x_d]$, pendant que $u^{\mathbf{r}}$ est l'onde réfléchie qui s'accorde avec la restriction de u sur $[x_g, 0[$.
Soient $M_{\mathbf{r}}$ la masse associée à l'onde réfléchie et $M_{\mathbf{t}}$ la masse associée à l'onde transmise. A l'instant final, on a

$$M_{\mathbf{t}} = 0.1354,\ M_{\mathbf{r}} = 0.0342, \text{ ainsi } M_{\mathbf{r}} < M_{\mathbf{t}} < M_i.$$

où M_i est la masse de la donnée initiale.
Les deux figures suivantes représentent la variation de la masse de l'onde réfléchie et l'onde transmise au cours du temps.

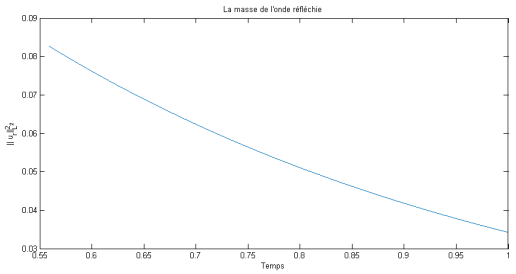


Figure 10 – La masse de l'onde réfléchie en fonction du temps pour $\gamma = 1$, et $Z = 10$.

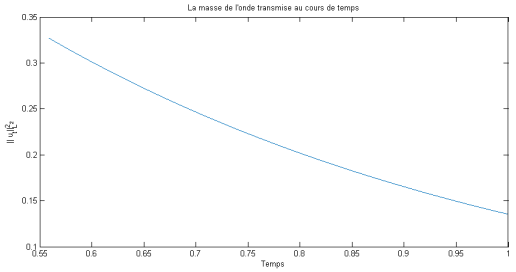


Figure 11 – La masse de l'onde transmise en fonction du temps pour $\gamma = 1$, et $Z = 10$.

On remarque que le défaut décompose l'onde incidente en deux parties (onde transmise et onde réfléchie), dont la norme L^2 de chacune décroit en fonction du temps.

4.3 Cas test 3 : influence du défaut sur la norme L^2

Dans un premier temps, on visualise la variation de la masse de la solution globale (sur tout le domaine de calcul) pour deux différentes amplitudes du défaut, $Z = 0$ et $Z = 10$, avec $\gamma = 1$.

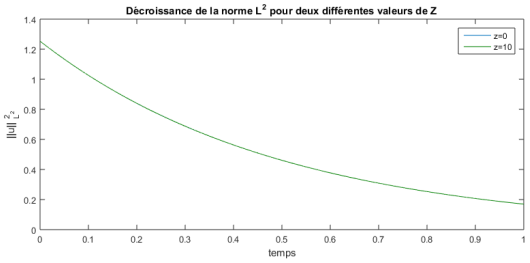


Figure 12 – La masse en fonction du temps pour $\gamma = 1$, $Z = 0$ et $Z = 10$.

On observe que les valeurs de la norme L^2 des solutions calculées se superposent au cours du temps. Ainsi, le présence de défaut n'influe pas sur la manière de la décroissance de la norme L^2. Dans la figure suivante, on visualise la norme L^2 de l'onde transmise en fonctions du temps et pour différentes valeurs de Z.

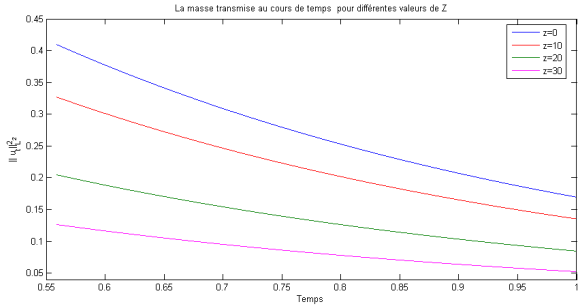


Figure 13 – La masse de l'onde transmise en fonction du temps pour $\gamma = 1$, et différentes valeurs de Z.

On remarque que plus l'amplitude du défaut est importante, plus la masse de la partie transmise est faible.

4.4 Cas test 4 : Z assez grand

Ici, on étudie le comportement de la solution suite au passage par un défaut d'amplitude assez grande, $Z = 10000$ avec $\gamma = 1$.

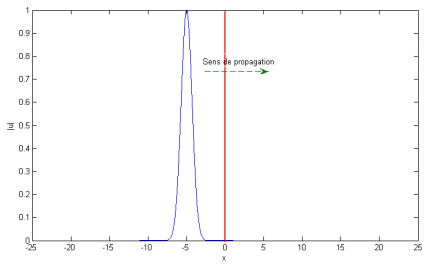


Figure 14 – La donnée initiale.

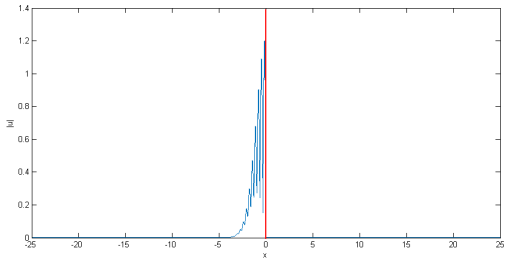


Figure 15 – La solution numérique calculée pour $\gamma = 1$, $Z = 10000$ à l'instant $t^n = 0.2683$.

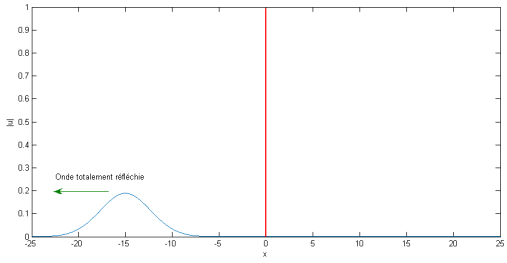


Figure 16 – La solution numérique calculée pour $\gamma = 1$, $Z = 10000$ à l'instant final.

On observe que le défaut a joué le rôle d'une barrière : la solution est totalement réfléchie et la partie transmise est nulle. On remarque aussi que la masse de la solution réfléchie à l'instant final est égale à la masse de la solution de l'équation NLS faiblement amortie sans défaut, $M_{\mathbf{r}} = M_T = 0.1696$.

Références

[1] B. BALLAND, « Optique géométrique - imagerie et instruments ». *Presses polytec*, 2007.

[2] T. CAZENAVE « Semilinear Schrodinger Equations ». Courant Lecture Notes in Mathematics, vol. 10, New York University Courant Institute of Mathematical Sciences, New York, 2003.

[3] M. DELFOUR, M. FORTIN, G. PAYRE. « Finite-difference solutions of a nonlinear Schrödinger equation ». *J. Comput. Phys*, vol. 44(2) (1981), 277-288

[4] E. EZZOUG, O. GOUBET, E. ZAHROUNI « Semi-discrete weakly damped nonlinear 2-D Schrödinger equation ». *Differential and Integral Equations*, vol. 23, (2010), 237-252.

[5] F. GENOUD, B. A. MALOMED, R. M. WEISHÄUPL. « Stable NLS Solitons in a cubic-quintic medium with a delta-function potential ». *Nonlinear Analysis : Theory, Methods & Applications*, vol. 133, (2016), 28-50

[6] J. GINIBRE « Introduction aux équations de Schrödinger non linéaires ». Cours de DEA 1994-1995, Orsay, Paris 11 éditions.

[7] R. H GOODMAN, P. J HOLMES, M. I WEINSTEIN « Strong NLS soliton defect interactions ». *Physica D : Nonlinear Phenomena*, 2004.

[8] J. HOLMER, C. LIU. « Blow-up for the $1D$ nonlinear Schrödinger equation with point nonlinearty I : Basic theory ». arXiv :1510.03491, october 2015.

[9] T.KATO « On nonlinear Schrödinger equation ». *Ann. Inst. H. Poincaré Phys. Théor.*, vol. 46(1987), 113-129.

[10] S. LE COZ, R. FUKUIZUMI, G. FIBICH, B. KSHERIM, Y. SIVAN. « Instability of bound states of a nonlinear Schrödinger equation with a Dirac potential », Physica D, vol. 237 (2008) 1103-1128.

[11] H. C. NGUYEN, B. T. KUHLMEY, E. C. MÄGI , M. J. STEEL, P. DOMACHUK, C. L. SMITH, B. J. EGGLETON, « Tapered photonic crystal fibres : properties, characterisation and applications ». *Applied Physics B*, vol. 81, Issue 2-3, 2005.

[12] C. SULEM, P. L SULEM, « The Nonlinear Schrödinger Equation. Self-Focusing and Wave Collapse ». *Applied Mathematical Sciences*, vol. 139, 1999.

[13] C. ZHENG. « A perfectly matched layer approach to the nonlinear Schrödinger equations ». *Journal of Computational Physics*, 2007.

Rubrique

Towards a recommender system for healthy nutrition

An automatic planning-based approach

Ngoko Yanik

Laboratoire d'Informatique de Paris Nord
University of Paris 13
Paris
France
yanik.ngoko@lipn.univ-paris13.fr

RÉSUMÉ. Ce papier introduit le problème de la planification automatique des repas. Étant donné un ensemble de repas caractérisés par une composition calorifique, un prix, des éléments de qualité nutritionnelle et la période de consommation (petit-déjeuner, dîner etc.), le but est de construire l'affectation optimale des repas dans les périodes afin d'optimiser la qualité de l'alimentation et/ou le plan. Le problème de la planification automatique des repas est particulièrement intéressant en Afrique. Avec la récente émergence des systèmes de télécommunications ainsi que la construction des bases de données sur l'alimentation africaine, il peut servir à construire des systèmes de recommandation, permettant aux consommateurs d'optimiser leur budget de nutrition tout en maintenant une alimentation équilibrée. Ce papier contribue sur ce défi en formalisation le problème, l'analysant et en proposant un algorithme par séparation-évaluation pour sa résolution. Enfin, nous effectuons une validation expérimentale à partir des données sur l'alimentation en Tanzanie. Les résultats montrent que nous pouvons produire des plans nettement meilleurs que ceux issus d'approches plus naïves.

ABSTRACT. In this paper, we focus on the automatic meal planning problem (AMP). Assuming a set of meals characterized by a calorific composition, a price, a nutrient composition and mealtimes, the objective is to decide on the meal to assign to each mealtime such as to obtain an optimal plan in term of nutritional quality and prices. AMP is particularly interesting in Africa. Indeed, thanks to the emergence of telecommunication networks and works done on the statistical modeling of nutrition in Africa, it can serve for designing recommender systems, accessible on cellular phones that will help consumers to make a better planning of their budget while keeping a balanced nutrition. Our study contributes to this objective by formalizing the problem, analyzing it and providing a branch and bound algorithm for its resolution. Finally, we did an experimental evaluation based on open data available for nutrition in Tanzania. The results show that the plan we produce can largely overpass naive solutions.

MOTS-CLÉS : Système de recommandation, alimentation équilibrée, problème d'optimisation de contraintes

KEYWORDS : Recommender system, balanced nutrition, constraint optimization problem

1. Introduction

The global objective of this study is to help to improve the world nutrition by providing adequate numeric tools for a balanced and responsible nutrition. For this purpose, we introduce the automatic meal planning problem (AMP). Assuming a set of meals that we mainly characterize by their ingredients, nutrient composition, price and a window of mealtimes, the objective is to automatically build a meal plan that states *what to eat in each mealtime* while optimizing a budget and a nutritional balance.

The AMP is particularly interesting in African countries. Indeed, it can serve for building recommender systems that will help people to keep a balanced nutrition. The idea of building such recommender systems is not new. It is part of a general trend that consists of investigating nutrition challenges with computer algorithms. Some popular questions in this tendency are the finding of equivalence between ingredients [1], the analysis of flavor between recipes [2] or the discovery of the structure similarity in recipes [3]. In this paper, our focus is on the automatic composition of meal plans.

Closer to our objective, we can refer to the CHEF system [4], a recommender system based on user preference [5], the smart kitchen system [6] or the daily meal plan recommender system [8]. Our work shares several common features with these works. These are : the formalization of meals through cook recipes, the management of users preferences, the distinction of meals in types or the objective to converge towards a healthy nutrition. However our work differs from these studies on three main points. The first point is that these works tackle the problem mainly in an Artificial Intelligence (case-based reasoning, expert systems etc.), information system or data analysis perspective (data clustering, statistical analysis etc.) while we are interesting in modeling, formalizing and solving the combinatorial problem of meal plan composition. The second novelty that our work introduces is to consider qualitative evaluation of meals based on the a nutritional score system that we will refer to as the Hercberg score [9]. The Hercberg score is a classification that ranges foods in distinct classes depending on the quantity of nutrients they include and the type of ingredients they have. This classification is becoming a popular standard and was adopted for food labeling in France. Finally, we provided experiments that demonstrate how our system can be used for healthy nutrition based on open data available for nutrition in Tanzania.

The remainder of this paper is organized as follows. In Section 2 , we discuss the related work. Section 3 introduces the theoretical formulation of the automatic meal planning problem. In Section 4, we propose a heuristic for solving the problem and evaluate it. We conclude in Section 5.

2. Related work

The design of automatic meal planner was investigated rather early in Artificial Intelligence. One of the first proposed system was CHEF [4], a case-based reasoning system that was able to recommend dishes based on their types and the taste expected by the consumer. As the CHEF system, in the problem we propose, dishes are ranged in types and are characterized by their ingredients. Unlike CHEF, we do not explicitly account on the tasting but consider a general concept of preference. Finally, we consider the quality of meals based on the nutrient composition. In [5], the authors propose a recommender system for recipes based on users preferences. From several observations made on recipes chosen by the users, the system is able to detect what are the user favorite ingredients.

Based on these ingredients, a classification of recipes is proposed and then used for recommending recipes to users. In our work, we also handle user preferences. However, it is only a criteria for deciding on the best recipes. In [6] the authors introduce the smart kitchen, an intelligent kitchen that returns qualitative data about cooking processes. The smart kitchen includes sensors and camera that serve for detecting any cooking action and estimating its *nutritional and calorific value*. The system also provides recommendations for adjusting the real-time composition of a meal towards a nutritional balance. While the smart kitchen is a hardware and software innovation, we focus in this paper on the software aspect of meals planning. In [7] the authors propose a planning system for healthy nutrition. The system is based on propositional logic and can be used on mobile devices. As our work, the objective of this study is to propose a digital assistant to fill the lack of experts in poor countries. However, we differ from this study on our modeling of the meals planning problem. In [8] the authors propose a recommender system for building daily nutrition plan. They demonstrate that their solution can provide balanced nutrition plans that respect users' preferences. A common feature between our formulation and this recommender system is the idea of considering the quality of meal plans. But, while this paper proposes a custom classification of foods, we consider the Hercberg score. In addition, we do not only focus on daily plan (as it is the case in this study) : our mealtime window can include weeks and months. Finally, it is important to observe that our study is possible because of existing theoretical formulations for characterizing the quality of nutrition based on discrete quantity. These are for instance the Hercberg score [9] and nutrient composition tables [11]. In the next, we present our model.

3. Problem description and analysis

3.1. General view

We consider a family that has a finite set of mealtimes $D = \{t_1, t_2, \ldots, t_k\}$. Typically, we might have $k = 3$ with t_1 being the breakfast, t_2 the lunch and t_3 the diner. We also assume that the family is interested in a meal plan in a horizon of $\Delta = \{1, \ldots, T\}$ days. $T = 30, 1$ are meaningful values when considering cultural notion as the concept of "ration" in Africa [1]. At each mealtime, the family can opt for a meal issued from a finite set $M = \{m_1, m_2, \ldots, m_n\}$. The general goal in AMP is to build an assignment σ that for each meal m_i, mealtime $t_u \in D$ and days $d_j \in \Delta$ is such that $\sigma(m_i, t_u, d_j) = 1$ if on day d_j and at the mealtime t_u, the meal m_i was chosen. The built assignment must satisfy objectives and criteria specified by the family. This general formulation is subject to constraints and objectives that we will define in the next.

3.2. Formal definition

For the sake of simplicity, we reduce the family to a single person. This choice will impact the formulation of constraints related to the nutritional balance. We also assume the following (additional) input data :

– K_i the calories provided by meal m_i ;

1. In several African countries, husbands give a budget for cooking to their wives everyday or at the beginning of the month.

– $\alpha_i^1, \alpha_i^2, \alpha_i^3$ the percentage of carbohydrates, fat and proteins in meal m_i ;

– p_i, the price of the m_i ;

– A Boolean function $\gamma(i, c)$ such that $\gamma(i, c) = 1$ is meal m_i belongs to culture c. We also assume that we have a set C of cultures.

– H, W, G, A the height (cm), weight (kg), gender, age of the person we consider ; $G = 1$ for female, 0 for male.

– $R = \{R_1, \ldots, R_5\}$ the classes of recommendations the user could follow. R_1 corresponds to a consumer that makes little or no exercise ; R_2 is a consumer that makes $1-3$ days of exercise per week ; R_3 a consumer with $3 - 5$ days of exercise ; R_4 a consumer with $6 - 7$ days of exercise ; and R_5 a consumer with very intensive exercises. We also assume the Boolean variables r_is that are such that $r_i = 1$ if the user chose the class R_i.

– $E(t_u)$ the set of acceptable meal in the mealtime t_u ;

– q_i , the Hercberg score of meal m_i ; the lower is q_i, the better is the quality of m_i.

We consider the percentage of proteins, fat and carbohydrates because as mentioned in [8], they are crucial for a balanced diet. We range each meal in a culture. This choice is among other things motivated by an observation made in prior studies [6] : consumers choose their dishes according to cultural preferences. With the set $E(t_u)$ of acceptable meals, our objective is to distinguish between types of meals that are appropriate depending on the mealtime. Finally, we consider the height and weight of the consumer because this will serve to estimate his requirement in term of calories. Assuming these data, in the next, we will now define the constraints.

3.2.1. Constraints

We consider the following constraints :

C_1 **: One meal per mealtime**

$$\forall d_j, t_u, \sum_{m_i \in M} \sigma(d_j, t_u, m_i) = 1$$

C_2 **: The meal must be accepted**

$$\forall d_j, t_u, \sum_{m_i \in M | m_i \notin E(t_u)} \sigma(d_j, t_u, m_i) = 0$$

C_3 **: Maximum budget limit per day**

$$\forall d_j, \sum_{t_u \in D} \sum_{m_i \in M} \sigma(d_j, t_u, m_i).p_i \leq B$$

(the daily budget for eating is B)

C_4 **: Diversity in meal choice**

$$(1)\forall m_i, \sum_{t_u \in D} \sum_{d_j \in \Delta} \sigma(d_j, t_u, m_i) \leq F_i; (2)\forall m_i, d_j \sum_{t_u \in D} \sigma(d_j, t_u, m_i) \leq 1$$

(1) means that a meal is chosen at most F_i times. (2) means that per day, a meal cannot be chosen twice.

C_5 **: Cultural preferences**

$$\forall c \in C, \sum_{d_j \in \Delta} \sum_{t_u \in D} \sum_{m_i \in M} \sigma(d_j, t_u, m_i).\gamma(i, c) \geq I_c;$$

(meal from a culture c will be chosen at least I_c times)

C_6 **: Calorific recommendation based on the Harris-Benedict equation**

$$\forall d_j, \sum_{t_u \in D} \sum_{m_i \in M} \sigma(d_j, t_u, m_i).K_i = [G.bmr_1 + (1-G).bmr_2].bmrFactor$$

Here, $bmr_1 = 447.593 + 9.247W + 3.098H - 4.330A$; $bmr_2 = 88.362 + 13.397W + 4.799H - 5.677A$ and $bmrFactor = 1.2r_1 + 1.375r_2 + 1.55r_3 + 1.725r_4 + 1.9r_5 + 200\epsilon_1$

This constraint expresses the calorific need according to the Harris-Benedict equation [10]. We added a margin error factor $\epsilon_1 \in [-1, 1]$ that ensures that the proposed plan will exceed or be lower of at most 200 calories from the standard recommendation.

C_7 **: Balanced diet requirement**

$$\forall d_j, \sum_{t_u \in D} \sum_{m_i \in M} \sigma(d_j, t_u, m_i).\alpha_i^1.K_i = (0.55 + 0.10\epsilon_2) \sum_{t_u \in D} \sum_{m_i \in M} \sigma(d_j, t_u, m_i)K_i$$

$$\forall d_j, \sum_{t_u \in D} \sum_{m_i \in M} \sigma(d_j, t_u, m_i).\alpha_i^2.K_i = (0.275 + 0.75\epsilon_3) \sum_{t_u \in D} \sum_{m_i \in M} \sigma(d_j, t_u, m_i)K_i$$

$$\forall d_j, \sum_{t_u \in D} \sum_{m_i \in M} \sigma(d_j, t_u, m_i).\alpha_i^1.K_i = (0.225 + 0.125\epsilon_4) \sum_{t_u \in D} \sum_{m_i \in M} \sigma(d_j, t_u, m_i)K_i$$

Here, $\epsilon_2, \epsilon_3, \epsilon_4 \in [-1, 1]$. The idea in balanced diet requirement is to ensure that in the calories gained each day, between 45 to 65% come from carbohydrates, 20 to 35% from fat and 10 to 35% from proteins. Let us notice that these values are recommended by experts in nutrition [8].

3.3. Objective functions in the automatic meal planning problem

In AMP, we want to minimize the price and the Hercberg score of the plan : the lower is this score, the better is the quality. We modelize the price and quality of a plan as follows : $Price = \sum_{t_u \in D} \sum_{m_i \in M} \sum_{d_j \in \Delta} \sigma(d_j, t_u, m_i).p_i$; $Quality = \sum_{t_u \in D} \sum_{m_i \in M} \sum_{d_j \in \Delta} \sigma(d_j, t_u, m_i).q_i$. The objective function in AMP is the normalized function

$$Cost = \lambda \frac{Price}{|Price| + |Quality|} + (1 - \lambda) \frac{Quality}{|Price| + |Quality|}$$

Here $\lambda \in [0, 1]$ is a parameter defined by the consumer to give more interest in either price or quality.

3.4. Analysis

It is straightforward to notice that AMP is a Constraint Optimization Problem. The interest in the observation is that we can therefore consider general Constraint optimization framework like Branch and Bound for its resolution. We also have the following result.

Theorem 3.1 *If we only consider the constraints C_1, C_2 and C_3 then AMP is NP-hard.*

The proof is given in the appendix. It is based on a reduction to the 3-partition problem. Finally, let us notice that several variants of AMP can be proposed. For instance, we can model the diversity in considering *neighbor meals*. A neighbor could refer to meals of the same day or those in consecutive days.

4. Heuristic and evaluation

From the mathematical formulation proposed in Section 3.2, we can derive an Integer Linear Program (ILP) for solving AMP. The only difficulty will could come from the nonlinear objective function. Despite the interest in ILP, let us notice that the runtime can quickly explode when we consider big problems. However, let us observe that we described AMP as a constraint optimization problem. For such problems, branch and bound algorithms (B&B) are efficient. We will describe such an algorithm in the next.

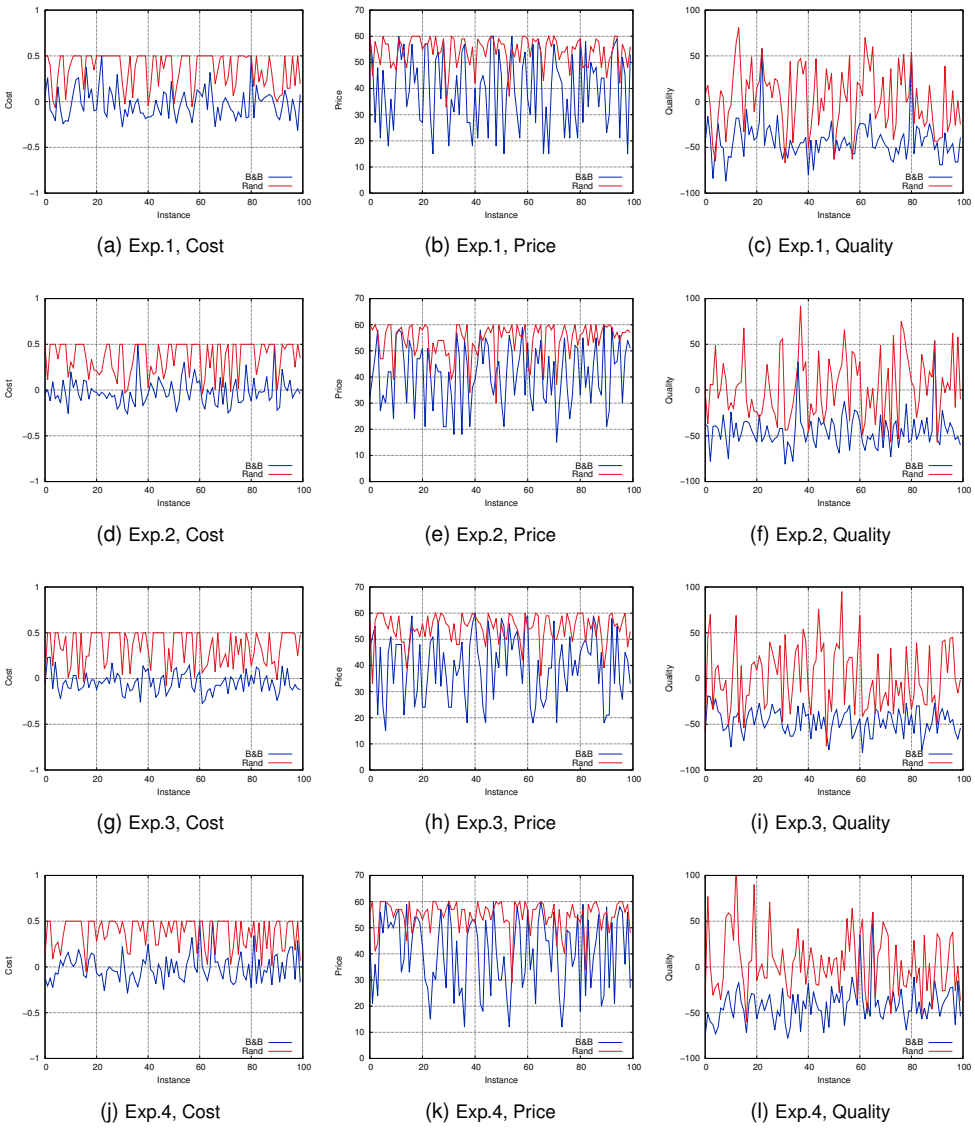
4.1. A branch and bound algorithm for AMP

In this algorithm, we consider that a solution to AMP is a one dimensional vector X such that each $X(e)$ states for a pair $e = (d_j, t_u)$ the meal m_i that was chosen. Consequently, $|X| = |D|.|\Delta|$ and the domain of possible values for $X(e)$ is $dom(X(e)) = E(t_u)$. In the B&B algorithm, we start by assigning a value to $X(0)$ and evaluate partially all the constraints from C_3 to C_7. For instance, the partial evaluation of C_3 consists of checking whether or not we already exceeded the maximal budget. If no violation is found, we continue in assigning a value to $X(1)$ and repeating the process. Let us now assume that at a moment, we have a sub-vector $X(1...i)$ and that we detect a violation with the assignment made to $X(i+1)$. Then, we backtrack by changing the value of $X(i+1)$. If no possible values could be assigned to $X(i+1)$ we backtrack to $X(i)$. Finally, in this algorithm we keep every time a lower bound : the partial value of $Cost$ for the assignment we made. If this bound exceeds the best found solution, we backtrack.

4.2. Experimental evaluation

We evaluated the B&B algorithm in using a database of Tanzanian food composition [11]. We chose from this database 106 recipes of Tanzanian meals for which we have the ingredients and nutrient composition. Based on these data, we computed the quality of each meal and their calorific values. In the experiments, we randomly generated the price of each meal in choosing a value between 1 and 50. We also assume that half of the recipes belong to one culture and the remaining to another one. We also fixed the following values $k = 3, T = 4, I_c = 3, F_i = 0.3.(3 \times 4), \lambda = 0.5$. Finally, we assumed different settings where the consumer has one of the *standard profile* defined in [12]. We chose 4 of these profiles : female sedentary, 31-50 (Exp. 1), female sedentary, 51+ (Exp.2), male sedentary, 51+ (Exp.3), male sedentary, 31-50 (Exp.4). For each experiment, we randomly generated 100 price distributions. We then compared the best solution obtained by the B&B algorithm after at most 5 min, with a randomized algorithm. This latter solution was obtained by running a randomized version of the B&B that was interrupted once a feasible solution was found. The randomization was applied here on the ordering we used for processing the $X(i)$s. The solution issued from the randomized algorithm could correspond to the consumer choice. Indeed, we do not believe that in practice, consumers will make a deep exploration of the huge space of potential solutions. Therefore, the first feasible solution (naive solution) could probably be the one they will adopt.

The results of our experiments are presented in Figure 1. As expected, the solutions of the B&B are better (in cost) than the naive ones. But more interestingly, they are not only better when considering the objective function : as showed by the curves on prices and quality, we are able to find plans that a both cheaper and of better quality. Let us recall indeed that in the Hercberg score, the lower is the score, the better is the quality.



(a) Exp.1, Cost (b) Exp.1, Price (c) Exp.1, Quality

(d) Exp.2, Cost (e) Exp.2, Price (f) Exp.2, Quality

(g) Exp.3, Cost (h) Exp.3, Price (i) Exp.3, Quality

(j) Exp.4, Cost (k) Exp.4, Price (l) Exp.4, Quality

Figure 1 – Cost, price and quality in different experiments

5. Conclusion

In this paper, we modeled the automatic design of balanced meals plans and proposed an algorithm for its construction. Our modeling is based on keys mathematical concepts in nutrition like the Harris-Benedict equation and the distribution of calories in healthy diet. We then validated our algorithm in considering a database of Tanzanian foods. The experimental results showed that with our modeling, we are able to find balanced nutrition plans that outperform naive solutions on both prices and quality. For continuing this work, we have three main perspectives. The first is to refine the modeling and evaluation in including other elements like tasting and enlarging the database of meals. The second is

to validate the approach in considering a pool of real consumers. Finally, we envision to reduce the runtime of the B&B algorithm in using parallelism and advanced constraint optimization techniques.

6. Bibliographie

[1] YUKA SHIDOCHI, TOMOKAZU TAKAHASHI, ICHIRO IDE, HIROSHI MURASE, « Finding replaceable materials in cooking recipe texts considering characteristic cooking actions », *CEA '09 Proceedings of the ACM multimedia 2009 workshop on Multimedia for cooking and eating activities*, pp. 9-14, 2009.

[2] YONG-YEOL AHN, SEBASTIAN E. AHNERT, JAMES P. BAGROW, ALBERT-LÁZLÓ BARABÁSI, « Flavor network and the principles of food pairing », *Nature, Scientific Reports*, vol. 1, n° 196, pp 1-7, 2011.

[3] LIPING WANG, QING LI, NA LI, GUOZHU DON, YU YANG, « Substructure Similarity Measurement in Chinese Recipes », *Proceedings of the International World Wide Web Conference*, pp. 979-988, 2008.

[4] KRISTIAN J. HAMMOND, « CHEF : A model of Case-based Planning », *Proceedings of AAAI*, pp. 267-271, 1986.

[5] MAYUMI UEDA, SYUNGO ASANUMA, YUSUKE MIYAWAKI, SHINSUKE NAKAJIMA, « Recipe Recommendation Method by Considering the User's Preference and Ingredient Quantity of Target Recipe », *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, pp. 519-523, 2014.

[6] JEN-HAO CHEN, PEGGY PEI-YU CHI, YUSUKE MIYAWAKI, HAO-HUA CHU, CHERYL CHIA-HUI CHEN, POLLY HUANG, « A Smart Kitchen for Nutrition-Aware Cooking », *Pervasive computing*, vol. 9, n° 4, pp. 58-65, 2010.

[7] FERNANDO ZACARIAS F., ROSALBA CUAPA, ERICK MADRID, DIONICIO ZACARIAS, « Healthy Nutrition Under ASP-PROLOG », *International Journal of Computer Networks & Communication*, vol. 5, n° 3, pp. 91-102, 2013.

[8] DAVID ELSWEILER, MORGAN HARVEY, « Towards Automatic Meal Plan Recommendations for Balanced Nutrition », *Proceedings of the 9th ACM Conference on Recommender Systems*, pp. 313-316, 2015.

[9] SERGE HERCBERG, « Propositions pour un nouvel élan de la politique nutritionnelle française de santé publique », *http ://www.sante.gouv.fr/IMG/pdf/rapport_Hercberg_15_11_2013.pdf*, Accessed 26 October 2015

[10] J. ARTHUR HARRIS, FRANCIS G. BENEDICT, « A Biometric Study of Human Basal Metabolism », *Published by The Carnegie Institute of Washington. Proc Natl Acad Sci U S A.* vol. 4, n° 12, pp 370-373, 1918.

[11] ZOHRA LUKMANJI, ELLEN HERTZMARK, NICOLAS MLINGI, VINCENT ASSEY, GODWIN NDOSSI, WAFAIE FAWZI, « Tanzania Food composition Tables », *http ://www.hsph.harvard.edu/nutritionsource/food-tables/* Accessed 26 October 2015.

[12] KATHLEEN M. ZELMAN, « Estimated Calorie Requirements », *http ://www.webmd.com/diet/estimated-calorie-requirement*, Accessed 26 October 2015

7. Appendix

7.1. Proof of theorem 3.1

Let us recall that in this proof, we consider a restricted version of AMP that only includes the constraints C_1, C_2, C_3. The NP-hardness proof is based on a reduction to the 3-partition problem. Given a set S of $3l$ positive integers $s_1, \dots s_{3l}$, the objective in 3-partition is to subdivide S into l triplets $S_1, \dots S_l$ such that the sum of number in each subset is equal and the sets $S_1, \dots, S_l$ cover S.

From this instance, we propose to build the following AMP instance : We set $T = l$ and $k = 3$. This means that the AMP instance has 3 mealtimes per day and covers l days. We assume $3l$ meals and associate each meal m_i with the price $p_i = s_i$. We fix $F_j = 1$ (all chosen meals are distinct) and

$$B = \frac{\sum_{u=1}^{3l} e_i}{l}$$

Finally, we set the quality of each meal to 0 (such values exist in the Hercberg score).

For solving any instance of the 3-partition problem, we formulate the associated AMP instance and solve it. If σ is the solution, then we associate each S_j with a day d_j as follows :

$$S_j = \{p_i | \sigma(d_j, t_u, m_i) = 1\}(a)$$

It is straightforward to notice that if there is a solution to the 3-partition then there is a solution to the associated AMP instance where the maximal budget spent by day is exactly B. Reciprocally, in any solution of the associated AMP instance, the total value spent by day is B and each meal corresponds to a distinct s_i. This implies that the S_j as defined in (a) will constitute a cover and the sum of each S_j will be equal to B. For concluding the proof, we must now ensure that the reduction is done in polynomial time.

Given an instance of 3-partition, the construction of the associated AMP can be done in $O(l)$. Once, the instance is solved, the construction of S_j can be done in $O(l^2)$. Indeed, it suffices to loop over each $\sigma(d_j, t_u, m_i)$. Consequently, we have a polynomial time reduction.

Algorithmes Hybrides Pour la Résolution du Problème du Voyageur de Commerce

Baudoin TSOFACK NGUIMEYA* — Mathurin SOH* — Laure Pauline FOTSO **

* Département de Mathématiques-Informatique
Université de Dschang
BP 67 Dschang, Cameroun
nguimeyabaudoin@yahoo.fr
mathurin.soh@univ-dschang.org

** Département d'Informatique
Université de Yaoundé I, Cameroun
laurepfotso@yahoo.com

RÉSUMÉ. Cet article traite de la résolution du Problème du Voyageur de Commerce (PVC). A partir de l'algorithme de Lin-Kernighan (LK) modif é par Helsgaun (LKH) qui est actuellement la meilleure heuristique d'amélioration, nous proposons deux nouvelles heuristiques hybrides de résolution du PVC. Elles sont basées d'une part sur l'hybridation d'un algorithme de colonie de fourmis(ACF) et de l'algorithme de LKH, sur la combinaison de l'algorithme génétique (AG) et de l'algorithme LKH d'autre part. Les résultats obtenus sur 10 problèmes choisis au hasard dans la librairie TSPLIB montrent que les algorithmes proposés sont très eff caces. Une solution optimale a été obtenue au moins 9 fois sur 10 pour tous les problèmes avec un optimum connu et à des temps meilleurs. Cela montre ainsi que les hybridations (LKH-AG) et (LKH-ACS) sont d'une grande importance dans la résolution du PVC.

ABSTRACT. This article deals with solving the Traveling Salesman Problem (TSP). From the Lin-Kernighan algorithm (LK) modif ed by Helsgaun (LKH), which is currently the best heuristic improvement, we propose two new hybrid heuristic for solving the TSP. They are based in part on the hybridization of a Ant Colony Algorithm (ACF) and the LKH algorithm, and on the combination of a genetic algorithm (GA) and the LKH algorithm in other part. The tests on 10 randomly selected problems in the TSPLIB show that the proposed algorithms are very effective. An optimal solution was obtained at least 9 out of 10 for all the problems with known optimal and better times, showing that (AG-LKH) and (ACS-LKH) hybridizations are of great importance in the resolution of the TSP.

MOTS-CLÉS : Colonie de fourmis, Heuristique, Hybridation, Lin-Kernighan, PVC.

KEYWORDS : Ants colony, Heuristic, Hybridization, Lin-Kernighan, TSP.

1. Introduction

Le Problème du Voyageur de Commerce (en anglais Traveling Salesman Problem, TSP in abbreviated form) consiste pour un voyageur de commerce de visiter un certain nombre de villes, débutant et finissant son parcours dans la même ville en visitant chacune des autres villes une et une seule fois. Le voyageur désire sélectionner la tournée qui minimise la distance totale parcourue[1]. Les algorithmes de résolution du TSP peuvent être répartis en deux classes : les algorithmes déterministes qui trouvent la solution optimale et les algorithmes d'approximation qui fournissent une solution presque optimale. Cependant, une autre classe dite de méthodes hybrides émerge, tirant profit des avantages des deux premières classes de méthodes.

Nous nous intéressons dans ce travail à la problématique des approches hybrides de résolution du PVC. Dans un premier temps, l'algorithme de Lin-Kernighan modifié par Helsgaun (LKH) est hybridé avec un algorithme de colonie de fourmis (ACS). Ensuite, il est hybridé avec un algorithme génétique (AG). Nous développons ainsi deux nouvelles méthodes hybrides LKH-ACS et LKH-AG en se basant sur les heuristiques LKH, ACS , et AG. Nous nous appuyons surtout sur deux techniques d'hybridation :l'hybridation de bas niveau en ce qui concerne le LKH-ACS et l'hybridation de haut niveau dans le cas LKH-AG.

Le reste de ce travail est organisé comme suit. La section 2 présente l'état de l'art. La section 3 présente les différentes approches et méthodes de résolution du TSP. Enfin, dans la section 4, nous présentons deux nouvelles techniques hybrides pour la résolution du TSP. Nous combinons ainsi les avantages du LKH avec ceux de ACS et AG. Les résultats expérimentaux sont mis en exergue dans la section 5. Dans la section 6, nous interprétons et discutons de ces résultats.

2. Le problème du Voyageur de Commerce (TSP)

Le TSP se définit comme suit : Etant données n points (villes) séparés par des distances, il faut trouver un chemin de longueur totale minimale qui passe exactement une fois par chaque point et revienne au point de départ (une tournée). En effet, selon l'ordre dans lequel on visite les villes, on ne parcourt pas la même distance totale. La notion de distance peut-être remplacée par d'autres notions comme le temps qu'il met ou l'argent qu'il dépense : dans tous les cas, on parle de coût. C'est un problème d'optimisation combinatoire qui consiste à trouver la meilleure solution parmi un ensemble de choix possibles. Il est généralement modélisé comme un graphe dans lequel chaque noeud représente une ville à visiter et les arêtes représentent les routes reliant les villes. En terme de complexité, le TSP est considéré comme étant NP-difficile [1, 2]. On ne connaît pas de méthode de résolution permettant d'obtenir des solutions exactes en un temps raisonnable pour de grandes instances (grand nombre de villes) du problème [2].

3. Méthodes de résolution du TSP

Plusieurs méthodes de résolution existent pour résoudre le TSP. Ces méthodes sont classées en deux groupes. D'un coté, nous avons les méthodes exactes qui donnent des

solutions optimales pour des problèmes de taille raisonnable. Seulement, le temps nécessaire pour trouver cette solution augmente exponentiellement avec la taille du problème. Nous pouvons citer à titre d'exemple, les méthodes de recherche arborescente (branch and bound), la programmation dynamique, la programmation linaire en nombres entiers, la méthode force brute... De l'autre coté, nous avons les méthodes approchées (ou heuristiques) qui permettent de trouver une solution dont le coût est proche du coût de la solution optimale au bénéfice d'un temps meilleur [1, 2]. Elles constituent une alternative très intéressante pour traiter les problèmes d'optimisation de grande taille si l'optimalité n'est pas primordiale. Ces méthodes sont fondées principalement sur diverses stratégies (heuristiques).

3.1. Les heuristiques d'amélioration

Leur principe est le suivant : une fois qu'une tournée est générée par une heuristique de construction, elles améliorent cette solution pour obtenir une tournée de qualité meilleure. A titre d'exemple, nous avons les procédures k-opt.

3.1.1. Procédures k-opt

L'idée de cette méthode est de partir d'une solution connue et d'explorer systématiquement une sphère de rayon k autour de cette solution à la recherche d'un cycle qui serait plus court encore. Cette transformation consiste à effacer k arêtes (k = 1,2,3,...) du tour, et à recomposer un autre tour en reconnectant cette chaîne d'une autre manière. L'algorithme k-opt, a une complexité en $O(n^k)$ [7]. A titre d'illustration, nous présentons l'heuristique 3-opt.

3.1.2. L'heuristique 3-opt

L'heuristique 3-opt commence avec une tournée admissible donnée et cherche ensuite dans le voisinage de la solution courante toute tournée améliorant la configuration courante. A chaque étape de l'itération, l'algorithme examine si l'échange de 3 arêtes produit une tournée plus courte. L'algorithme continue ainsi jusqu'à ce qu'aucune amélioration ne soit plus possible. Cette heuristique, illustrée par la figure 1, est communément utilisée dans de nombreuses techniques hybrides comme celle effectuée par Tadunfock et Fotso[1]ou celle de Dorigo [5]. Seulement, c'est un inconvénient de spécifier k en avance car il est difficile de savoir quelle valeur de k utiliser pour accomplir le meilleur compromis entre temps courant et qualité de solution.

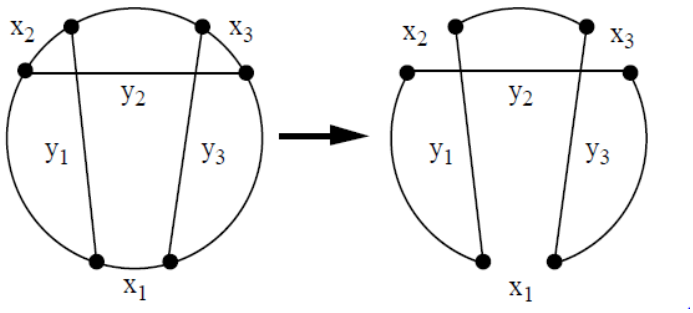


Figure 1. *3-opt*

3.1.3. Procédure de Lin et Kernighan(LK)

Cette procédure consiste à échanger itérativement un certain nombre d'arêtes à partir d'une solution donnée pour trouver une solution de meilleur coût.C'est une généralisation simple du principe du k-opt décrit dans la section précédente[6]. En effet, dans l'algorithme, k n'est pas fixé à une valeur précise. Mais, il varie de manière croissante jusquà une valeur de k qui n'améliore plus une solution déjà trouvée par la précédente. De plus, une règle centrale dans l'algorithme original du LK est celle qui restreint l'inclusion de liens dans la visite aux cinq voisins les plus proches à une ville donné [7]. Il a été prouvé par les auteurs [6] que cela permet de diriger la recherche vers une visite plus courte et réduit substantiellement l'effort de recherche mais aussi que le LK est le plus efficace pour résoudre le TSP symétrique actuellement[7].

3.1.4. Algorithme de Lin-Kergnighan- Helsgaun (LKH)

La conception et la mise en uvre d'un algorithme basé sur la procédure de Lin et Kernighan(LK) est pas trivial. En effet, elles font appel à beaucoup de décisions de conception et de mise en uvre. La plupart des décisions ont un une grande influence sur les performances finales. Helsgaun dans [6] propose une version modifiée et étendue de l'algorithme de LK [4, 7]. Son mise en oeuvre s'est avére très efficace et est appelée LKH. De ce fait, elle apporte une amélioration considérable à la procédure de Lin-Kernighan. Cette nouvelle version de l'algorithme implementé par Helsgaun est capable de trouver des solutions optimales pour tous les instances de problèmes testés, y compris le fameux problème de 7397 villes [4, 7, 8].

3.2. Les métaheuristiques

Une méta-heuristique est une classe d'heuristiques adaptable et applicable à une large classe de problèmes. C'est une méthode générique pour la résolution de problèmes combinatoires NP-difficiles. On distingue les approches perturbatives et les approches constructives[10].

Les approches perturbatives explorent l'espace des combinaisons en perturbant itérativement des combinaisons déja construites en partant d'une ou plusieurs combinaisons initiales (généralement prises aléatoirement dans l'espace des combinaisons). L'idée est de générer à chaque étape une ou plusieurs nouvelles combinaisons en modifiant une ou plusieurs combinaisons générées précédemment.

Les approches constructives construisent une ou plusieurs combinaisons de façon incrémentale, c'est-à dire, en partant d'une combinaison vide, et en ajoutant des composants de combinaison jusqu'à obtenir une combinaison complète[10].

3.2.1. Algorithme de colonie de fourmis

Basé sur l'approche constructive, un algorithme de colonies de fourmis (Ant Colony System, en Anglais) est une méthode itérative à population où tous les individus partagent un savoir commun qui leur permet de guider leurs futurs choix et d'indiquer aux autres individus des directions à suivre ou au contraire à éviter [5, 9, 13]. Cette méthode inspirée du déplacement des groupes de fourmis, a pour but de construire les meilleures solutions à partir des éléments qui ont été explorés par d'autres individus. Chaque fois qu'un individu découvre une solution au problème, bonne ou mauvaise, il enrichit la connaissance collective de la colonie. Ainsi, chaque fois qu'un nouvel individu aura à faire des choix, il pourra s'appuyer sur la connaissance collective pour pondérer ses choix. Les individus sont ici des fourmis qui vont se déplacer à la recherche de solutions et qui vont sécréter

des phéromones[1] pour indiquer à leurs congénères si un chemin est intéressant ou non. Si un chemin se retrouve fortement phéromoné, cela signifiera que beaucoup de fourmis l'ont jugé comme faisant partie d'une solution intéressante et que les fourmis suivantes devront la considérer avec intérêt[5, 13].

3.2.2. Mise en oeuvre du ACS pour la résolution du PVC

L'ACS décrit dans ce paragraphe est un Ant System orienté pour résoudre le problème du voyageur de commerce. Chaque noeud du graphe représente une ville. Soit d_{ij} la distance entre les villes i et j. (i,j), l'arête entre ces deux villes. Initialement (au temps t=0), l'algorithme positionne m fourmis sur n villes. À chaque unité de temps, chaque fourmi k choisit la prochaine ville à visiter parmi l'ensemble des villes V_k à l'aide de la règle de transition définie à l'équation 1. L'ensemble de villes V_k contient les villes que la fourmi k n'a pas encore visitées. L'équation 1 représente la probabilité qu'une fourmi k se déplace de la ville i à la ville j en considérant la distance à parcourir pour atteindre cette ville en fonction de la visibilité locale entre i et j (η_{ij}) et la quantité de phéromone (τ_{ij}) présente entre i et j.

$$\eta_{ij} = \frac{1}{d_{ij}}$$

$$p_{ij}^{k} = \frac{[\eta_{ij}]^{\beta}[\tau_{ij}(t)]^{\alpha}}{\Sigma u \in v_k [\eta_{iu}]^{\beta}[\tau_{iu}(t)]^{\alpha}} \quad (1)$$

α est le paramètre contrôlant l'importance accordée à la trace de phéromone et β celui contrôlant l'importance accordée à la visibilité. Une fois la tournée construite, chaque fourmi laisse une trace de phéromone sur les arêtes empruntées en mettant à jour la matrice de phéromone selon l'équation 2 ci dessous.

$$\tau_{ij}(t+1) = \rho \, \mathbf{P}_{ij}^{k}(t) + \triangle \tau_{ij} \quad \textbf{(2)}$$

On a $0 < \rho < 1$ qui est la persistance de la trace et (1 - ρ) représente l'évaporation des phéromones. $\Delta \tau_{ij} = \Sigma_{k=1}^{m} \Delta_{ij^k}$ cumule la quantité de phéromone Δ_{ij^k} laissée par unité de longueur sur l'arête (i,j) par l'ensemble des m fourmis.
En particulier, $\Delta_{ij^k} = \frac{Q}{L_k}$ si la k^{ieme} fourmi est passée sur l'arête (i, j) dans sa tournée, $\Delta_{ij^k} = 0$ dans le cas contraire. Q est une constante (généralement égale à 1) et L_k est la longueur de la tournée de la k^{ieme} fourmi calculée à la fin de chaque cycle. Ceci complète un cycle de l'algorithme.

Différents paramètres sont fixés à l'avance, tels que le nombre de fourmis, le nombre de cycles, les constantes Q, α et β. La principale particularité du Ant Colony System (ACS) est la mise-à-jour de la matrice de phéromone à l'aide de la meilleure fourmi de chaque cycle.

3.2.3. Algorithme génétique

Les algorithmes génétiques aussi appelés évolutionnaires, ont été introduits par Holland dès 1975. Ils sont issus d'une comparaison entre certains problèmes d'optimisation et la théorie de l'évolution de Darwin.
Cette théorie, complétée récemment par la biologie moderne, met en scène des populations d'êtres vivants soumis à une sélection naturelle et confrontés à des conditions de vie plus ou moins favorables. Les individus sont alors contraints d'évoluer de génération en génération, de façon à s'adapter au milieu sous peine d'extinction de l'espèce. Ainsi, ces

1. une substance odorante

algorithmes s'inspirent de l'évolution génétique des espèces. Leurs techniques reposent toutes sur l'évolution d'une population de solutions qui, sous l'action de règles précises, optimisent un comportement donné exprimé sous forme d'une fonction dite fonction coût, caractérisant l'adaptation à l'environnement. Le principe d'évolution des algorithmes génétiques est le suivant [11] :

– créer une population aléatoire pour chaque génération ;

– déterminer une liste dindividus à muter ;

– faire muter ces individus ;

– déterminer une liste d'individus à croiser ;

– croiser ces individus ;

– injecter ces 2 nouvelles listes d'individus dans la population et choisir les individus pour la génération suivante ;

– choisir les individus pour la génération suivante.

4. Nouvelles approches hybrides de résolution du Problème du Voyageur de Commerce

4.1. Justification et motivation

Actuellement,les approches hybrides gagnent en popularité car ce type d'algorithme produit généralement de meilleurs résultats pour plusieurs problèmes d'optimisation combinatoire[7]. En effet,les approches hybrides permettent d'obtenir de bons résultats dans une grande variété de problèmes théoriques d'optimisation combinatoire tels le problème du voyageur de commerce [1].

Etant donné que les heuristiques de construction se limitent généralement à une solution aux problèmes d'optimisation combinatoire difficiles,et que les heuristiques d'amélioration sont spécialistes pour améliorer une solution déjà connue, nous pensons que l'hybridation de ces métaheuristiques peut devenir une alternative très intéressante aux TSP symétrique. Les deux méthodes ont des particularités bien différentes qui peuvent être associées pour produire de meilleurs résultats.

L'hybridation des méthodes peut permettre de bénéficier des points forts de chacune de ces méthodes et de surmonter leurs limites.

4.2. Approche d'hybridation LKH-ACS

4.2.1. Principe et stratégie d'hybridation

Nous supposons que l'on dispose de m fourmis. Chaque fourmi k utilise l'heuristique de colonies de fourmis (ACS) pour produire une tournée. Une fois que toutes les fourmis ont construit leurs tournées, l'algorithme de Lin-Kernighan-Helsgaun (LKH) est appliqué à chacune de ces tournées en l'améliorant par des opérations de λ-opt move $\lambda \in \{2,3,4,5\}$ jusqu'à ce qu'aucune amélioration ne soit plus possible. Ensuite on effectue la régle de mise à jour globale aux solutions pour avoir la solution optimale. Etant donné que le fonctionnement interne de ACS n'est pas en relation avec celui de LKH, et que le second

algorithme fait suite au premier, nous parlons d'une hybridation de haut niveau à relais. L'algorithme proposé est le suivant :

Algorithme 4.1 : Approche d'hybridation LKH-ACS.

```
1  Début
2      Initialisation;
3      répéter
4          Chaque fourmi est positionnée à un noeud (ville de départ) ;
5          répéter
6              i. Chaque fourmi applique la règle de transition d'état pour se déplacer
               d'une ville à l'autre et construit ainsi une solution.;
7              ii. Chaque fourmi applique également la règle de mise à jour locale.
8          jusqu'à chaque fourmi achève sa tournée;
9          - Appliquer la procédure de LKH aux solutions (tournées) obtenues par
           chaque fourmi avec ACS;
10         - Appliquer la règle de mise à jour globale;
11     jusqu'à condition finale;
12 Fin
```

4.3. Approche d'hybridation LKH-AG

On débute avec l'algorithme génetique : Une population de solutions du problème est d'abord initialisée aléatoirement, puis évaluée : c'est la genèse. Certaines solutions de la population sont ensuite sélectionnées pour former la population de parents. Ces parents sont ensuite recombinées et modifiés pour produire une nouvelle population (enfants) en appliquant des opérateurs génétiques : c'est la phase de reproduction. Lors de cette phase, les opérateurs génétiques perturbent les parents afin d'explorer l'espace de recherche. Expliquons cela en détail :

Il existe deux types principaux d'opérateurs génétiques : **croisement et mutation**.

Croisement

A partir de deux parents (solutions) choisis aléatoirement en fonction de leurs évaluations (fitness)[2,10], nous essayons de générer un fils (une solution) qui soit réalisable(qui respecte les contraintes du problème).

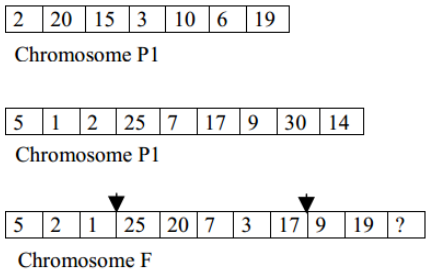


Figure 2. *Exemple de croisement entre deux parents*

Pour choisir les individus (solutions) qui seront en mesure de contribuer à la création de la nouvelle population, nous avons adopté un mode de sélection qui consiste à attribuer à chaque individu une probabilité de sélection proportionnelle à son évaluation (fitness)(fonction objectif) et à la somme des évaluations des individus en reference aux travaux de [**?**].

Si nous appelons f(i) la force de l'individu i, alors la fonction S(i), décrite dans l' équation (3) représente la probabilité de sélectionner ce chromosome i.

$S(i) = \frac{f(i)}{\sum_{j=1}^{n} f(j)}$ i∈ [1,N] (3)

avec N : la taille de la population.

Les enfants sont évalués via la mise à jour de leurs valeurs de fonction objectif S(i) et le chromosome les étapes de parcours.

Mutation

Au cours du processus d'évolution,la mutation effectue une exploration plus large de l'espace de recherche afin d'éviter toute convergence prématurée ou disparition de la diversité en apportant de l'innovation dans la population. La procédure consiste à inverser une chaîne de sommets du chromosome et les extrémités sont choisies aléatoirement.

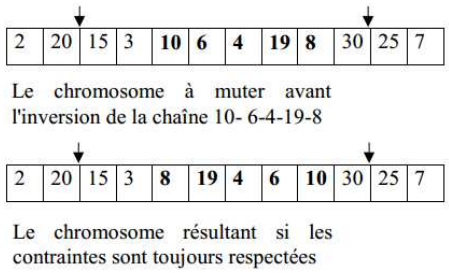


Figure 3. *Exemple de mutation*

Enfin, un sous ensemble de solutions est choisi parmi les parents et les enfants, pour remplacer la population courante par une nouvelle population pour la génération suivante. Ce processus est répété jusqu'a ce qu'une condition d'arret soit satisfaite.

L'algorithme retourne la (ou les) meilleure(s) solution(s) qu'il a identifié(s), qui est supposée etre une solution proche de l'optimale ou optimale. Une fois qu'on évalue les individus, l'algorithme de Lin-Kernighan - Helsgaun (LKH) est appliqué à l'individu le mieux adapté afin de l'améliorer. La méthode hybride résultante est l'algorithme ci-dessous.

5. Résultats expérimentaux

Nous avons implémenté nos algorithmes en langage C. Nous les avons exécutés sur plusieurs instances de TSP choisies dans la bibliothèque TSPLIB[12]. L'environnement de programmation présente les caractéristiques suivantes : Système d'exploitation : Débian, Processeur : 1.8 GHz core-Dio, RAM : 4 GHZ, DD : 500 GHz. Chaque résultat

Algorithme 4.2 : Algorithmes hybride LKH-AG.

```
Début
    Données : t : Taille de la population initiale;
    n : Nombre de générations ;
    Résultats : Cycle hamiltonien (Solution);
    t ← 0 ;
    I ← CreerPopulation(t) ;
    Tantque (t < n) Faire
        S ← Sélection(I) ;
        M ← Mutations (S) ;
        P1 ← 1ereMoitié(S) ;
        P2 ← 2emeMoitié(S) ;
        F ← Croisements(P1 ;P2) ;
        N ← Mutation(F) ;
        I ← I ∪ M ∪ F ∪ N ;
        I ← SupprimerDoublons(I) ;
        t ← t+1 ;
    Fintantque
    I' ← Renvoyer le meilleur individu de I ;
    Sol ← Appliquer la procédure de LKH à la solutions(I') obtenues;
Fin
```

obtenu a été testé cent (100) fois. Les tableaux 1, 2 et 3 recapitulent ces résultats. Les paramètres, Nbre Villes, Coût Op ACS-LKH, Tps de calcul, Nbre Réussite répresentent respectivement le nom du problème considéré, le nombre de villes du problème, le Coût optimal obtenu par notre algorithme, le temps mis pour obtenir cette solution exprimé en secondes, et le nombre de réussite sur 100 tests.

6. Interprétations et discussions

Parmi les techniques d'hybridation, on constate aussi que l'hybridation de haut niveau (LKH-ACS) semble être la meilleure technique actuellement. La figure 2 illustre clairement la nouveauté des algorithmes proposés par rapport à l'état de l'art particulièrement celle du *LKH-ACS*. En effet, la comparaison de la vitesse de convergence de LKH-ACS

Problème	Nbre Villes	Coût Op LKH-ACS	Temps de calcul	Nbre Réussite/100
eil51 (426)	51	426	0.1 Sec	100
lin 105 (14379)	105	14379	0.3	100
pr124 (59030)	124	59030	0.8	100
Pr144 (58537)	144	58537	9.2	100
att532 (276787)	532	27687.7	52.6	100
ALi535 (202339)	535	202339	19.4	100
rat783 (8806)	783	8806.00	2.8	100
std1655(62128)	1655	62128.6	732.8	100
Vm1748 (336557)	1748	336557	718.6	100
pr2392 (378032)	2392	378032	21.9	100

Tableau 1. *Résultat optimal obtenu par l'hybridation LKH-ACS*

Problèmes	Nbre Villes	Coût Op LKH-AG	Temps de calcul	Nbre Réussite/100
eil51 (426)	51	426	0.1	100
lin 105 (14379)	105	14379.02	0.3	100
pr124 (58537)	124	59030	4.6	100
Pr144 (58537)	144	58537	10.2	100
att532 (276787)	532	27691	137.4	100
aLi535 (202339)	535	202339	382.6	100
rat783 (8806)	783	8806.012	3.2	100
d1655 (62128)	1655	62129	111.28	97
Vm1748 (336557)	1748	336557	1170.70	99
pr2392 (378032)	2392	378032.8	317.01	98

Tableau 2. *Résultat optimal obtenu par l'hybridation LKH-AG*

et du LKH-AG avec les meilleurs algorithmes connus pour la résolution du TSP au moment des tests (LKH, LK-AG, LK-3opt, LK-ACS,) comme le montre la figure 2 a révélé que tous ces algorithmes aboutissaient certes à la solution optimale pour un grand nombre d'instances du problèmes. Mais en plus de cella, le LKH-ACS à réussir à surpasser ces algorithmes au niveau du temps mis pour obtenir cette solution optimale avec une différence très remarquable. Ce qui était notre objectif à savoir améliorer l'efficacité du LK-ACS , LK-AG, et LKH meilleurs algorithmes de l'heure pour la résolution du TSP. Sur cette courbe(figure2) nous remarquons que pour le probleme Pr2392 le LKH-ACS parvient à la solution optimale en 1.19s contre 1114.4s pour LK-AG, 317s pour LKH-AG, 60s pour LKH, 212.9 s pour ACS-3opt et 239.5s pour LK-ACS. Les tests sur d'autres instances ont donné d'aboutir aux mêmes conclusions.

7. Conclusion

Dans ce travail, nous nous sommes intéressés aux métaheuristiques hybrides de résolution du célèbre problème du voyageur de commerce. La revue de littérature faite, nous a permi de recenser une multitude d'approches de résolution du TSP. Parmi celles qui ont fait leur preuve, nous avons hybridé un algorithme génétique et un algorithme de colonie de fourmis avec l'algorithme de Lin-Kernighan modifié par Helsgaun (LKH) [6]. Cela nous à donné deux nouvelles métaheuristiques : LKH-AG, et LKH-ACS. Ces nouvelles méthodes ont été testées sur 10 instances de problèmes TSP choisies au hasard de la librairie TSPLIB. Les résultats expérimentaux nous ont permis de mettre en evidence la supériorité de LKH-ACS ainsi que celle de LKH-AG sur les autres metaheuristiques d'optimisation connues. Dans les travaux futurs, nous envisageons :

Problème	Meilleur Coût	Meilleur Coût LKH-ACS	Temps de calcul	Nbre Réussite/10
rl5934	548447.6	556136	41.03 Sec	2
fl3795	27487.9	28921	1220.36 Sec	0
usa13509	19849705.9	19983330	1864.74 Sec	0
vm1748	332049.8	336556	40.55 Sec	7
lin318	41881.1	41882	0.83 Sec	10
d1655	61456	62128	28.86 Sec	9

Tableau 3. *Résultats d'hybridation LKH-ACS sur de grandes instances*

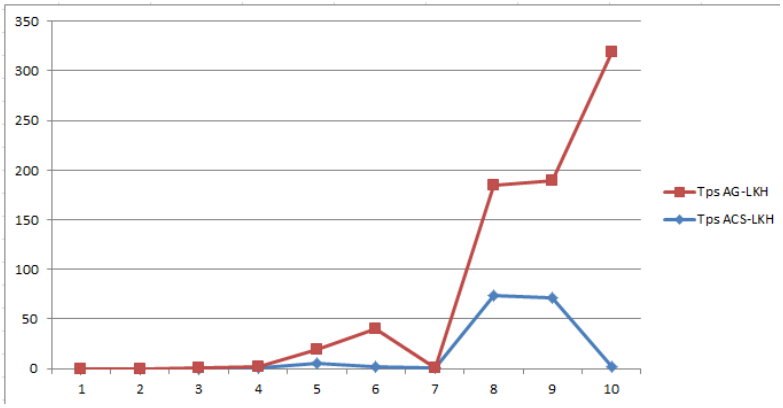


Figure 4. *Comparaison de la vitesse de convergence de LKH-ACS et du LKH-AG*

1) nous intéresser à la structure de la trace de pheromone ou au comportement de construction des fourmis pour d'avantage renforcer sa capacité a donner la solution optimale avant de l'améliorer avec LKH.

2) paralléliser en fonctions des paramétres AG, ACS ces méthodes.

8. Bibliographie

[1] B.TADUNFOCK TETI, L. P.FOTSO, « Heuristiques du problème du voyageur de commerce », *Proceedings CARI06*, vol. 1, n° 1-8, 2006.

[2] J. GREFENSTETE, R. GOPAL, B. ROSIMAITA, D.V. GUCHT, « Genetic Algorithms for the Traveling Salesman Problem », *Proceedings of an International Conference on Genetic Algo-*

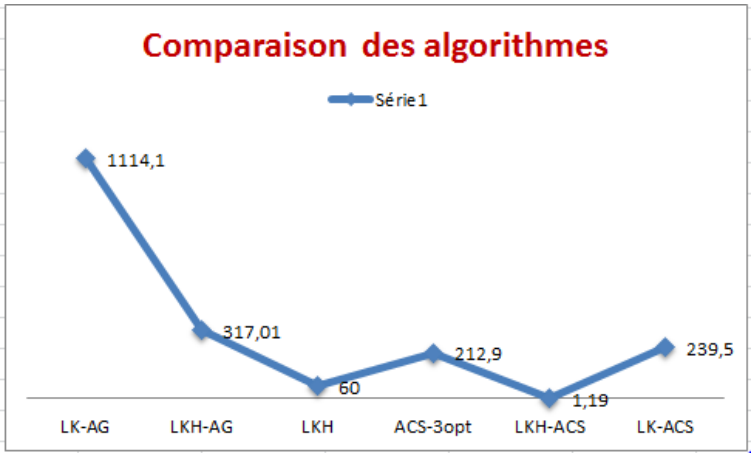


Figure 5. *Comparaison de la vitesse de convergence de LKH-ACS et du LKH-AG Avec les meilleurs algorithmes connus pour le PVC au moment des tests en fonction du Temps sur le Pr2392*

rithms and their Applications, n° Carnegie Mellon publishers, 1985.

[3] N.CAHON MELAB, « Designing cellular networks using a parallel hybrid metaheuristic », *Journal of Computer Communications*, n° pp 698-713, 2007.

[4] K. T. MAK, MORTON, « A modified Lin-Kernighan traveling-salesman heuristic », *Operations Research Letters*, n° pp 127-13, 1999.

[5] M. DORIGO, L.M. GAMBARDELLA, « AntColony System : A cooperative learning approach to the traveling salesman problem », *IEEE Transactions on Evolutionary Computation*, n° pp 53-66, 1997.

[6] K. HELSGAUN, « An effective implementation of the Lin-Kernighan traveling salesman heuristic », *European Journal of Operations Research*, vol. 12, n° :pp106-130, 2000.

[7] S. LIN, B. W.KERNIGHAN, « An Effective Heuristic Algorithm for the Traveling-Salesman Problem », *Operations Research*, vol. 21, n° 2 Pages498516, 1973.

[8] T. STUETZLE, « The Traveling Salesman Problem : State of the Art », *TUD SAP AG Workshop on Vehicle Routing*, vol. 12, n° July 10, 2003.

[9] M. DORIGO, T. STÜTZLE, « Ant Colony Optimization », *MIT Press, Cambridge, MA, USA*, vol. , n° 2004.

[10] L.SAID, « Méthodes bio-inpirées hybrides pour la résolution de problèmes complexes », *Thèse Doctorat en Sciences en Informatique, Université Constantine 2, Tunisie*, vol. , n° Avril 2013.

[11] O.MORIN « Résolution dun problème du voyageur de commerce avec des méta-heuristiques hybrides », *Thèse Doctorat en Sciences en Informatique, Université Constantine 2, Tunisie*, vol. , n° décembre, 2010.

[12] TSPLIB : « http ://www.iwr.uni-heidelberg.de/iwr/comopt/soft/TSPIB95/TSPLIB.html. », , vol. , n° december, June 2015.

[13] M.DORIGO, L. M.GAMBARDELLA « AntColony System : A cooperative learning approach to the traveling salesman problem », *IEEE Transactions on Evolutionary Computation*, vol. , n° 53-66, 1997.

A systematic approach to derive navigation model from data model in web information systems

Mohamed Tahar Kimour[a], Yassad-Mokhtari Safia[b]
[a]Département d'informatique
Université Badji Mokhtar-Annaba
Laboratoire des systèmes embarques (LASE)
BP 12, Annaba
ALGERIE
kimour@yahoo.com

[b]Département d'informatique
Université Badji Mokhtar-Annaba
Laboratoire des systèmes embarques (LASE)
BP 12, Annaba
ALGERIE
safiay03@yahoo.fr

RÉSUMÉ. Les méthodologies de conception de systèmes d'information web présentent le modèle de navigation comme étant un élément très critique dans le processus de développement. Ce dernier est un moyen efficace permettant de représenter la structure et le chemin selon lesquels les données sont présentées à l'utilisateur. Cependant, ces méthodologies ne traitent pas l'aspect comportemental lors de la modélisation de la navigation, où les services et l'interaction avec l'utilisateur ne sont pas présentés. Dans cet article, nous proposons une approche alternative de nature à élaborer un modèle de navigation plus complet et mieux structuré. Il est basé sur l'utilisation de la notion de cas d'utilisation atomique et la combinaison entre le modèle conceptuel de données et le modèle de cas d'utilisation. Ce faisant, notre approche fournit une vue unifiée des aspects structurels et de comportement d'une application Web.

ABSTRACT. The design methodologies of web information systems present the navigation model as a very critical element in the development process. It is considered as an efficient means to represent the structure and the path according to which data is shown to the user. However, such methodologies do not deal with the behavior aspect in the navigation modeling, and services and interaction with the user are not represented. In this paper, we present an alternative approach to build a more complete and better structured navigation model. It is based on the use of the atomic use case concept and the combination between the conceptual data model and the use case model. In doing so, our approach provides a unified view of the structural and behavior aspects of a web application.

MOTS-CLÉS: Système d'information web, modèle de navigation, UWE, Ingénierie du Web.

KEYWORDS: Web Information Systems, Navigation Model, UWE, Web Engineering.

CARI 2016, Tunis, Tunisie, Octobre 2016

1. Introduction

Nowadays, web information systems are increasingly adopted due to the ubiquity of the client and also because user experience is becoming each time more interactive [1, 2]. The most notable methods for web application design [4, 5, 8, 9, 10, 11, 12] support the design of Web applications building conceptual, navigation and presentation models. Conceptual modeling of Web applications does not differ from conceptual modeling of other applications.

One of the possible debatable concepts used in the web community is the concept of navigation. Navigation is an important aspect widely studied by a lot of researchers, such as comparison between requirements of the methods in [12], comparison business process development in [13], comparison between UWE, WebML and OOH in [14], Requirements Engineering In current web engineering methodologies [12], and comparison study describe advantages and disadvantages of some selected methods [15].

Navigation design is a critical step in the design of web applications, and the navigation model is one of the important models in the process of the developing web applications [7]. However, a navigation model based on a domain model is relatively rigid when faced to new, often unpredictable, use contexts. The reason is that the OO paradigm is specially suited to encapsulate data concerns into classes, but it is not so well suited to represent other concern types, such as business-related or functional.

Usually, navigation model is considered as a means to structure the information to show to the user, without any reference to the user-view behaviour. The navigation model is much more than this because it should integrate the system user-view services the web application should provide to the user.

In this paper, we propose an alternative approach to build a more complete, but also, better structured navigation model. It is based on the use of the atomic use case concept and the combination between the conceptual data model and the use case model. In doing so, our approach provide a unified view of the structural and behavior aspects of a web application.

The paper is structured as follows: Section 2 explains the background work for typical information system web engineering processes, especially the derivation process of the navigation model. Section 3 describes our approach throughout the presentation of underlying concepts and the method. Finally, Section 4 presents some concluding remarks and an overview of future work.

2. Navigation modeling

In the past few years, some web engineering methods have suggested an operation to the development of Web Applications. The significance of the navigation between the application nodes is the meaning of the navigational model which is one of the fields [13].

Both in the UML-based Web Engineering (UWE) [4] and the OO-H [5] methods the navigation model is derived in part from the content or conceptual model respectively. UWE is an approach that allows the modeling of the architecture, the navigation space and the interfaces of web systems using UML with some extensions [4]. It defines a UML profile including stereotypes which denote new modeling elements. The modeling process proposed by UWE is composed by four steps:

- Requirement Analysis with Use Cases.
- Conceptual Model.
- Navigation Model.
- Presentation Model

Based on the standard UML, the UWE methodology [4] is an object-oriented approach, where the notation and diagrams are restricted to those provided by UML. UWE presents a new approach for improving the navigation model. In the navigation space model, a stereotyped class diagram is used,

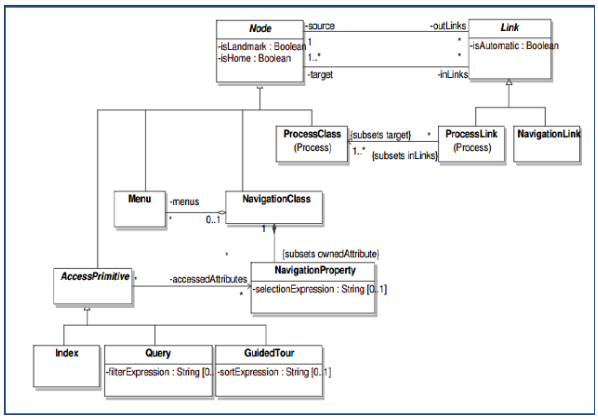


Fig. 1. UWE navigation package [3].

including the classes of those objects, which can be visited during navigation. The <<navigation class>> and the <<navigation link>> stereotypes are used to model nodes and links.

As a refinement of the navigation space model, the navigation structure model includes stereotypes such as: <<menu>>, <<query>>, <<index>> and <<guided tour>>. Modeled in UWE by a composite object, index means direct access to instances of a navigation class. Each index item is in turn an object, which has a name that identifies the instance and owns a link to an instance of a navigation class.

3. The proposed approach

Usually, the most notable current approaches to model hypermedia depart from any kind of domain model to define the navigation design model of the system under

development. Our approach, on the contrary, is different. It builds the navigation model from both the conceptual data model and the use case model (Fig. 2).

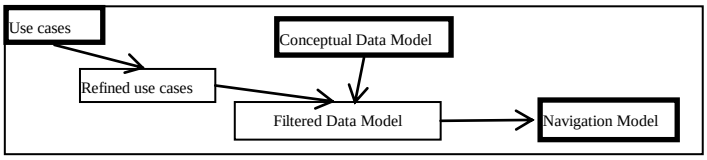


Fig. 2 The flow model of our approach.

We is illustrate our approach by taking the case study of a conference management Web system, which hosts multiple users and conferences, allowing the creation of new users and conferences at any time. Any user can apply for a new conference. After approval from the Supervisor, the applicant becomes a conference chair. He can add new chairs and new PC members. An author can list the conferences awaiting submissions. He can submit a paper, upload new versions, or indicate other users as co-authors thereby granting them reading and editing rights. PC members are allowed to view the submitted papers. PC Chair can assign papers to PC members for reviewing either manually or automatically based on some rules. Reviewers can download papers they are concerned with and upload their reviews. The authors can read the reviews and the accept/reject decision made by the PC chair.

A. The use case model

Use case modeling is widely used in modern software development methods as an approach for describing a system's software requirements [4]. A use case represents how a system interacts with its environment and who are the actors involved in such interactions.

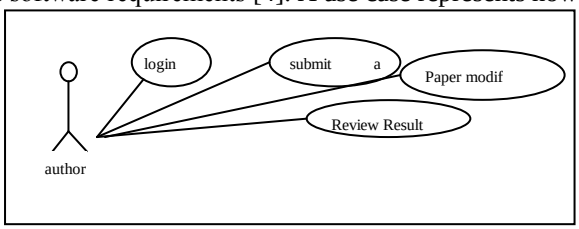


Fig.3 A use case model.

However, to deal with web pages and navigation, we need to break use cases into more reusable units.

Breaking use case down into their smallest size allows breaking them into their most reusable, most common elements. From there, structuring, planning, and designing become much more predictable.

To this end, we use the concept of atomic use cases [16]. An atomic use case is used to decompose a use case in order to identify units of functional behaviour a system should offer to the user. Such units of functional behaviour will be transformed into navigation structures.

Atomic use case is defined as an atomic functionality that the system offers to the user. For instance, the use case "buying a book" may be broken down into the atomic use cases: "viewing a book catalogue", "register as a new customer", deleting an item from the shopping cart.

We identify two types of atomic use cases: *structural atomic use case* and *functional atomic use case*. An atomic use case can be structural, when it provides a data view (i.e. viewing a catalogue, viewing the customer's data, etc.).

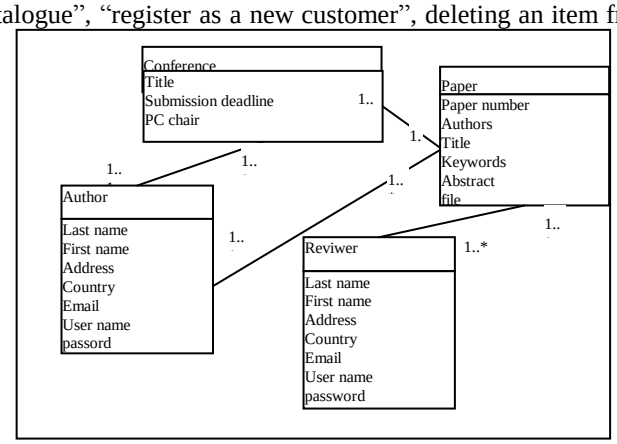


Fig.4 An excerpt of a conceptual data model of a conference management system.

A functional *atomic use case* implies some interaction with the user, generally requiring some input data (i.e. searching a book, adding a product to the shopping cart, etc.).

B. The conceptual data model

The conceptual data model of the Web application is built with UML classes models. This model is the input artifact to the derivation process of navigation and presentation models. Fig4 depicts an excerpt of a conceptual data model for a conference management web system.

C. Refining the use case model

To identify meaningful interaction units we refine the use case model using the atomic use case concept as defined above.

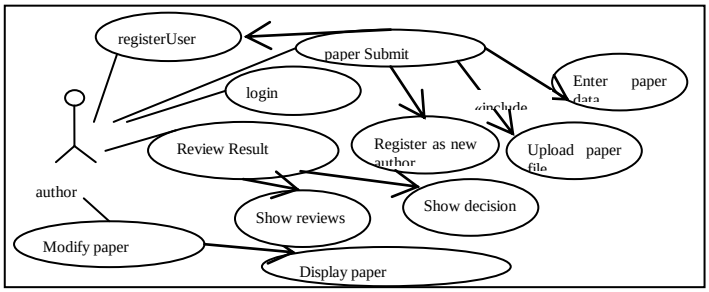


Fig.5 A refined use case model..

Refining the use case model consists of decomposition each use case into atomic use cases and hierarchically structure theme using the conceptual data model. For example in the use case "login", we identify two atomic use cases: *successful login* and *forgot password*.

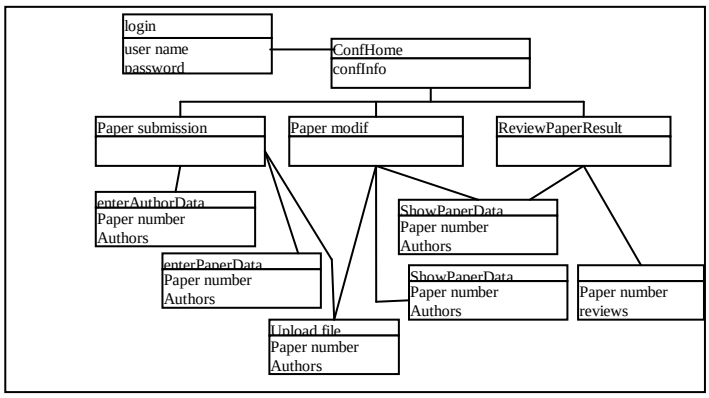


Fig.6 the refined conceptual data model

The use case "submit a paper" is decomposed into the following atomic use cases: *register new author data, register new paper data, and upload the paper*. Figure 5 depicts the refined use cases model.

D. Deriving the navigation model

In most current web engineering approaches, navigation model is created from navigation classes, a set of guided tours, indexes, queries and links. Also the navigation classes and links are parts of conceptual classes.

To build the navigation model, we use as input to our approach both the refined use case model (Fig.5) and the conceptual data model (Fig.4). Using the refined use case model, we filter the conceptual data model where only data elements and links that are relevant to the refined use case model remains.

In a use case, an atomic use case may be organized using include and extend relationships defined by UML. "Include" and "extend" relationships have the same semantics that in the use case model: An include relationship specifies the location in a flow of events in which the base use case includes the behavior of another one that the behavior of the base use case may be optionally extended by the behavior of another use case [20g]. UML defines two stereotypes to mark these relationships: the <<include>>,<<extend>>.

Taking into account the definition of structural atomic use cases, we refine the conceptual data model in order to represent data and links invoked by such atomic use cases. In addition, we enrich the obtained conceptual model by adding corresponding classes invoked by the functional atomic use cases. We start by adding a home class to represent common data.

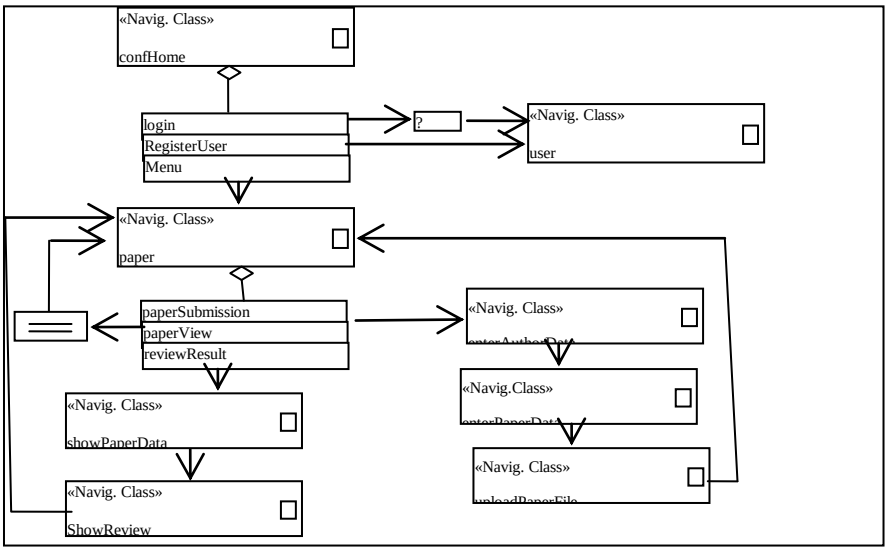


Fig.7 the navigation model..

For an atomic use case we define a class that contains invoked data in the initial conceptual data model. Figure 6 depicts. A navigation model will be derived from the two above-mentioned models, while adding the navigational structures such as menus, indexes, queries, guided tours, etc. Thus, we apply the mapping rules of UWE to derive the navigation model from our refined conceptual data model. Our navigation model includes not only navigation between data nodes and behavioral nodes (Fig. 7.).

4. Conclusion

In web engineering, navigation model is defined as an important model in all the hypermedia design. In this paper we have presented an approach to systematic derivation of the navigation model from use cases and the conceptual data model. Use cases are refined and restructured based on the concept of atomic use case. Then, the conceptual data model is refined with regard the refined use case model. In doing so, we have facilitated the derivation process of the navigation model, which incorporates not only, data and their links but also the user-view related behavioral units. Our approach allows systematizing the derivation of the navigation model, throughout the definition of the process, the modeling techniques and the mapping rules. As a future work, we plan to investigate other case studies to measure the usability of our approach, and also to integrate it into model-driven design environment.

5. Bibliographie

[1] Valeria de Castro, Esperanza Marcos, Paloma Cáceres, A User Service Oriented Method to model Web Information Systems, WISE 2004, LNCS 3306, pp. 41-52, Springer-Verlag, 2004.

[2] Ingrid O. Nunes, Uirá Kulesza, Camila Nunes, Elder, Extending web-based applications to incorporate autonomous behavior, Proceedings of the 14th Brazilian Symposium on Multimedia and the Web, WebMedia '08, ACM, New York, NY, USA, 2008.

[3] N. Koch and A. Kraus, "Towards a common metamodel for the development of web applications", Web Engineering, (2003), pp. 419-422.

[4] N. Koch, A. Knapp, G. Zhang, and H. Baumeister, "Uml-based web engineering," Web Engineering: Modelling and Implementing Web Applications, pp. 157-191, 2008.

[5] Schwabe, D and Rossi, G.: An Object -Oriented Approach to Web-Based Application Design. Theory and Practice of Object Systems (TAPOS), Vol 4 (1998) 207-225.

[6] Lars Bækgaard, Event-Based Activity Modeling (2004), ALOIS'04 – in proceedings of Action in Language, Organisation and Information Systems, Linköping, Sweden, 2004.

[7] Karzan Wakil, Amirhossein Safi and Dayang. N. A. Jawawi, Enhancement of UWE Navigation Model: Homepage Development Case Study, International Journal of Software Engineering and Its Applications, Vol.8, No.4 (2014), pp.197-212.

[8] G. Rossi, "Web engineering: modelling and implementing web applications", Springer, vol. 12, (2008).

[9] K. Vlaanderen, F. Valverde and O. Pastor, "Improvement of a web engineering method applying situational method engineering", ICEIS (3-1), (2008), pp. 147-154.

[10] J. Conallen, "Building Web applications with UML", Addison-Wesley Professional, (2003).

[11] Karzan Wakil, Dayang N. A. Jawawi, and Amirhossein Safi , A Comparison of Navigation Model between UWE and WebML: Homepage Development Case Study , International Journal of Information and Education Technology, Vol. 5, No. 9, September 2015

[12] R. Jeyakarthik, "Requirements engineering in current web engineering methodologies," International Journal, vol. 2, 2011.

[13] T. Bosch, "A web engineering approach for the development of business process-driven web applications," Ph.D. dissertation, Dept. Information Systems and Computation Technical, Univ. of Valencia, 2008.

[14] R. Gustavo, O. Pastor, D. Schwabe, and L. Olsina, "Web engineering: modelling and implementing web applications," Human–Computer Interaction Series, vol. 12, Springer, 2008.

[15] A. L. D. S. Domingues et al., "A comparison study of web development methods," 2008

[16] K. Nguyen, Th. Dillon, Atomic Use Case: A Concept for Precise Modelling of Object-Oriented Information Systems, Lecture Notes in Computer Science, Vol. 2817, 2003,pp 400-411.

Réconciliation par consensus des mises à jour des répliques partielles d'un document structuré

Maurice TCHOUPÉ TCHENDJI *, Milliam M. ZEKENG NDADJI*

* Département de Maths-Informatique
Faculté des Sciences, Université de Dschang
BP 67, Dschang-Cameroun
ttchoupe@yahoo.fr
ndadjimaxime@yahoo.fr

RÉSUMÉ. Dans un workf ow d'édition coopérative asynchrone d'un document structuré, chacun des co-auteurs reçoit dans les différentes phases du processus d'édition une copie du document pour y insérer sa contribution. Pour des raisons de conf dentialité, cette copie peut n'être qu'une réplique partielle ne contenant que les parties du document (global) qui sont d'un intérêt avéré pour le co-auteur considéré. Remarquons que certaines parties peuvent être d'un intérêt avéré pour plus d'un co-auteur ; elles seront par conséquent accessibles en concurrence. Quand vient le moment de la synchronisation (à la f n d'une phase du processus d'édition par exemple), il faut fusionner toutes les contributions de tous les co-auteurs en un document unique. Du fait de l'asynchronisme de l'édition et de l'existence potentielle des parties offrant des accès concurrents, des conf its peuvent surgir et rendre les répliques partielles non fusionnables dans leur entièreté: elles sont incohérentes ou en conf its. Nous proposons dans ce papier une approche de fusion dite par consensus de telles répliques partielles à l'aide des automates d'arbre. Plus précisément, à partir des mises à jour des répliques partielles, nous construisons un automate d'arbre dit du consensus qui accepte exactement les documents du consensus. Ces documents sont les préf xes maximums ne contenant pas de conf it des répliques partielles fusionnées.

ABSTRACT. In an asynchronous cooperative editing workf ow of a structured document, each of the co-authors receive in the different phases of the editing process a copy of the document to insert its contribution. For conf dentiality reasons, this copy may be only a partial replica containing only parts of the (global) document which are of demonstrated interest for the considered co-author. Note that some parts may be a demonstrated interest over a co-author; they will therefore be accessible concurrently. When it's synchronization time (for eg. at the end of a phase of the process), we want to merge all contributions of all authors in a single document. Due to the asynchronism of edition and to the potential existence of the document parts offering concurrent access, conf icts may arise and make partial replicas unmergeable in their entirety: they are inconsistent or in conf ict. We propose in this paper a merging approach said by consensus of such partial replicas using tree automata. Specif cally, from the partial replicas update, we build a tree automaton that accepts exactly the consensus documents. These documents are the maximum pref xes containing no conf ict of partial replicas merged.

MOTS-CLÉS : Documents Structurés, Workf ow d'Edition Coopérative, Fusion des Répliques Partielles, Conf its, Consensus, Automates d'Arbre, Produit d'automates, Evaluation Paresseuse.

KEYWORDS : Structured Documents, Worf ow of Cooperative Edition, Merging Partial Replicas, Conf ict, Consensus, Tree Automata, Automata Product, Lazy Evaluation.

1. Introduction

Une proportion importante des documents manipulés et/ou échangés par les applications présente une structure régulière définie par un modèle grammatical (DTD-*Document Type Definition*-, schéma, ...) : on les appelle des *documents structurés*. La puissance toujours croissante des réseaux de communication en terme de débit et de sûreté ainsi que le souci d'efficacité a révolutionné la façon d'éditer de tels documents : au modèle classique d'un auteur éditant en local et de façon autonome son document, s'est adjoint l'édition coopérative (asynchrone) dans laquelle, plusieurs auteurs situés sur des sites géographiquement éloignés se coordonnent pour éditer de façon asynchrone un même document structuré (fig. 1). Dans de tels processus d'édition coopérative asynchrone, les phases d'éditions désynchronisées dans lesquelles chaque co-auteur édite sur son site sa copie du document, alternent avec les phases de synchronisation-redistributions dans lesquelles les différentes contributions (répliques locales) sont fusionnées en un unique document, qui est par la suite redistribué aux différents co-auteurs pour la poursuite de l'édition.

Pour des raisons de confidentialité, il est parfois souhaitable qu'un co-auteur n'ait accès qu'à certaines informations c-à-d, à des parties du document appartenant à des types donnés (des *sortes*) du modèle du document. Ainsi, la réplique t_i éditée par le co-auteur c_i depuis le site i n'est donc qu'une *réplique partielle* du document (global[1]) t ; elle est obtenue via une *opération de projection* qui supprime convenablement du document global t les parties qui ne sont pas accessibles au co-auteur considéré. Nous appelons *vue* d'un co-auteur, l'ensemble des *sortes* qui lui sont accessibles [1].

Quand les éditions locales asynchrones se font sur des répliques partielles, on peut supposer que chaque co-auteur possède sur son site un modèle local de document le guidant dans son édition à l'aide duquel on peut garantir que toute mise à jour d'une réplique partielle valide par rapport à ce modèle local est cohérente vis-à-vis du modèle global du document[2]. Ainsi, du fait de l'asynchronisme de l'édition, les seules incohérences qu'on puisse avoir quand arrive le moment de la synchronisation sont celles issues de l'édition concurrente du même noeud (du point de vue du document global) par plusieurs co-auteurs : on dit que les répliques partielles concernées sont en conflits. Ce papier propose une approche de détection et de résolution de tels conflits par *consensus* pendant la phase de synchronisation-redistribution, en se servant d'un automate d'arbre dit du consensus, pour représenter l'ensemble des documents qui sont les consensus d'éditions concurrentes réalisées sur les répliques partielles.

Un document structuré t est représentable intentionnellement par un arbre possèdant éventuellement des bourgeons[3] [1]. Intuitivement, synchroniser ou fusionner consensuellement les m-à-j $t_1, \ldots t_n$ de n répliques partielles d'un document t, consiste à trouver un document t_c conforme au modèle global, intégrant tous les noeuds des t_i non en conflits et dans lequel, tous les noeuds en conflits sont remplacés par des bourgeons. L'algorithme de fusion consensuelle présenté dans ce papier est une adaptation de celui de fusion présenté dans [1] qui ne gère pas les conflits. Techniquement, la démarche permettant d'obtenir les documents faisant partie du consensus est la suivante : (1) pour chaque m-à-j t_i^{maj} d'une réplique partielle t_i, nous associons un automate d'arbre $\mathcal{A}^{(i)}$ reconnaissant les arbres (conforme au modèle global) dont t_i^{maj} est une projection. (2) L'automate consensuel $\mathcal{A}_{(sc)}$ engendrant les documents du consensus est obtenu en effectuant

1. Nous désignons par *document global* ou tout simplement *document* quand il n'y a pas d'ambiguïté, le document comprenant toutes les parties.
2. Intuitivement, un modèle local de document sera dit cohérent vis-à-vis du modèle global si tout document partiel t_i qui lui est conforme est la réplique partielle d'au moins un document (global) t conforme au modèle global.
3. Un *bourgeon* est un noeud feuille d'un arbre indiquant qu'une édition doit être effectuée à ce niveau dans l'arbre. Editer un bourgeon revient à le remplacer par un sous arbre en se servant des productions de la grammaire du document.

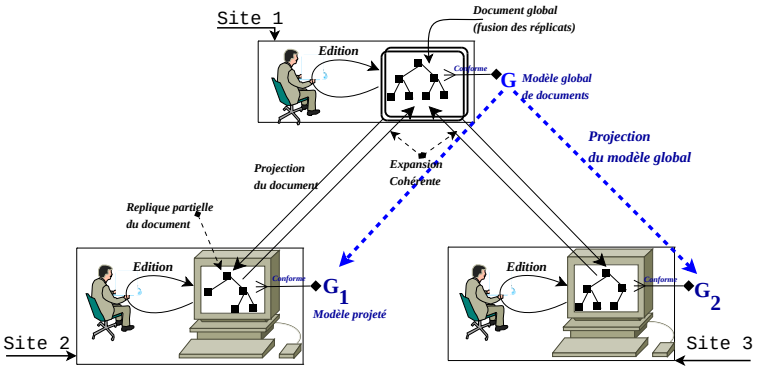


Figure 1. *L'édition coopérative désynchronisée ; site1 : édition+fusion du document (global) conformément au modèle (global) G de documents ; sites 2 et 3 : éditions des répliques partielles conformément aux modèles projetés de documents G_1 et G_2 obtenus à partir du modèle global G.*

un *produit synchrone* des automates $\mathcal{A}^{(i)}$ au moyen d'un opérateur commutatif et associatif noté $\otimes^{\Omega}$: $\mathcal{A}_{(sc)} = \otimes_i^{\Omega} \mathcal{A}^{(i)}$ que nous définissons. Il suffit alors de générer l'ensemble des arbres (ou ceux les plus représentatifs) acceptés par l'automate $\mathcal{A}_{(sc)}$, pour avoir les documents du consensus.

Dans ce qui suit, après la présentation (sec. 2) de quelques concepts et définitions relatifs à l'édition coopérative et aux automates d'arbre, nous exposons (sec. 3) le processus de construction de l'opérateur $\otimes_i^{\Omega}$. La section 4 est consacrée à la conclusion. En annexe, nous déroulons complètement l'exemple introduit dans la section 3 en mettant en exergue les concepts manipulés dans ce papier.

2. Edition coopérative structurée et notion de réplique partielle

2.1. Document structuré, édition et conformité

Il est usuel de représenter la structure abstraite d'un document structuré par un arbre et son modèle par une grammaire algébrique abstraite ; un document structuré valide est alors un arbre de dérivation pour cette grammaire. Une grammaire algébrique définit la structure de ses instances (les documents qui lui sont conformes) au moyen des productions. Une production, généralement notée $p : X_0 \to X_1 \dots X_n$ est assimilable dans ce contexte à une règle de structuration présentant comment le symbole X_0 situé en partie gauche de la production se décompose en une séquence d'autre symboles $X_1 \dots X_n$ situés dans sa partie droite. Plus formellement,

Définition 1 *Une **grammaire algébrique abstraite** est la donnée $\mathbb{G} = (\mathcal{S}, \mathcal{P}, A)$ d'un ensemble fini $\mathcal{S}$ de **symboles grammaticaux** ou **sortes** qui correspondent aux différentes **catégories syntaxiques** en jeu, d'un symbole grammatical $A \in \mathcal{S}$ particulier, appelé **axiome**, et d'un ensemble fini $\mathcal{P} \subseteq \mathcal{S} \times \mathcal{S}^*$ de **productions**. Une production $P = \left(X_{P(0)}, X_{P(1)} \cdots X_{P(n)}\right)$ est notée $P : X_{P(0)} \to X_{P(1)} \cdots X_{P(n)}$ et $|P|$ désigne la longueur de la partie droite de P.*

Pour certains traitements sur les arbres (documents) il est nécessaire de désigner précisément un noeud particulier. Plusieurs techniques d'indexation existent parmi lesquelles celle dite de numérotation dynamique par niveau [6] basée sur des identificateurs à longueur variable inspirés

de la classification décimale de *Dewey*. Suivant ce système d'indexation, on peut définir un arbre comme suit :

Définition 2 *Un **arbre** dont les noeuds sont étiquetés dans un alphabet $\mathcal{S}$ est une fonction $t : \mathbb{N}^* \to \mathbf{S}$ dont le domaine $Dom(t) \subseteq \mathbb{N}^*$ est un ensemble clos par préfixe tel que pour tout $u \in Dom(t)$ l'ensemble $\{i \in \mathbb{N} \mid u \cdot i \in Dom(t)\}$ est un intervalle d'entiers $[1, \cdots, n] \cap \mathbb{N}$ non vide ($\varepsilon \in Dom(t)(racine)$) ; l'entier n est l'**arité** du noeud d'adresse u. $t(w)$ est la valeur (étiquette) du noeud de t d'adresse w. Si $t_1, \cdots, t_n$ sont des arbres et $a \in \mathbf{S}$, on note $t = a[t_1, \ldots, t_n]$ l'arbre t de domaine $Dom(t) = \{\varepsilon\} \cup \{i \cdot u \mid 1 \leq i \leq n,\ u \in Dom(t_i)\}$ avec $t(\varepsilon) = a$ et $t(i \cdot u) = t_i(u)$.*

Soient t un document et $\mathbb{G} = (\mathcal{S}, \mathcal{P}, A)$ une grammaire. t est un arbre de dérivation pour $\mathbb{G}$ si sa racine est étiquetée par l'axiome A de $\mathbb{G}$, et si pour tout noeud interne n_0 étiqueté par le sorte X_0, et dont les fils $n_1, \ldots n_n$, sont étiquetés respectivement par les sortes $X_1, \ldots X_n$, il existe une production $P \in \mathcal{P}$ telle que, $P : X_0 \to X_1 \cdots X_n$ et $|P| = n$. On dit aussi dans ce cas que t appartient au langage engendré par $\mathbb{G}$ à partir du symbole A et on note $t \in \mathscr{L}(\mathbb{G},\ A)$ ou encore $t \therefore \mathbb{G}$.

Un document structuré en cours d'édition est représenté par un arbre contenant des *bourgeons* (ou *noeuds ouverts*) qui indiquent dans un arbre, les seuls lieux où des éditions (mises à jour) sont possibles[4]. Les bourgeons sont typés ; un *bourgeon de sorte X* est un noeud feuille étiqueté X_ω : il ne peut être édité (étendu en un sous-arbre) qu'en utilisant une *X-production* (production ayant X en membre gauche). Ainsi donc, un document structuré en cours d'édition et ayant la grammaire $\mathbb{G} = (\mathcal{S}, \mathcal{P}, A)$ pour modèle est un arbre de dérivation pour la grammaire étendue $\mathbb{G}_\Omega = (\mathcal{S}, \mathcal{P} \cup \mathcal{S}_\Omega, A)$ obtenue de $\mathbb{G}$ en ajoutant à l'ensemble $\mathcal{P}$ des productions et pour tout sorte $X \in \mathcal{S}$, une nouvelle ε-production $X_\Omega : X \to \varepsilon$; $\mathcal{S}_\Omega = \{X_\Omega : X \to \varepsilon,\ X \in \mathcal{S}\}$.

Pour décider de la conformité d'un document complètement édité ou en cours d'édition, on peut utiliser comme outil formel les automates d'arbre. Comme nous le verrons ci-dessous, il est facile de construire un automate d'arbre (un reconnaisseur/générateur) à partir d'une grammaire donnée. En effet, quand on regarde les productions d'une grammaire, on peut remarquer que chaque sorte est associé à un ensemble de productions. On peut donc de ce point de vue considérer une grammaire comme une application $gram : symb \to [(prod,\ [symb])]$qui associe à chaque sorte une liste de couples formés d'un nom de production et de la liste des sortes en partie droite de cette production. Une telle observation suggère qu'une grammaire peut être interprétée comme un automate d'arbre (descendant) utilisable pour la reconnaissance ou pour la génération de ses instances.

Définition 3 *Un **automate d'arbre** (descendant) défini sur Σ est la donnée $\mathcal{A} = (\Sigma, Q, R, q_0)$ d'un ensemble Σ de symboles ; ses éléments sont les étiquettes des noeuds des arbres à reconnaître, d' un ensemble Q d'états, d'un état particulier $q_0 \in Q$ appelé état initial, et d'un ensemble fini $R \subseteq Q \times \Sigma \times Q^*$ de transitions.*

– Un élément de R est noté $q \to (\sigma, [q_1, \cdots, q_n])$: intuitivement, il s'agit de la liste des états $[q_1, \cdots, q_n]$ accessibles à partir d'un état q donné en franchissant une transition étiquetée σ.

– Si $q \to (\sigma_1, [q_1^1, \cdots, q_{n_1}^1]), \cdots, q \to (\sigma_k, [q_1^k, \cdots, q_{n_k}^k])$ désigne l'ensemble des transitions associées à l'état q, on note next $q = [(\sigma_1, [q_1^1, \cdots, q_{n_1}^1]), \cdots, (\sigma_k, [q_1^k, \cdots, q_{n_k}^k])]$ la liste formée des couples $(\sigma_i, [q_1^i, \cdots, q_{n_i}^i])$. Dans le cas où q est un état terminal (c-à-d. aucune transition n'est franchissable depuis l'état q), next $q = []$.

4. Nous nous intéressons dans ce papier qu'à l'*édition positive* basée sur une réplication optimiste [5] partielle des documents édités ; en effet, les documents édités ne font que croitre : il n'y a pas d'effacement possible dès qu'une synchronisation a été effectuée.

On peut interpréter une grammaire $\mathbb{G} = (\mathcal{S}, \mathcal{P}, A)$ comme un automate d'arbre (descendant) [2] $\mathcal{A} = (\Sigma, Q, R, q_0)$ en considérant que : (1) $\Sigma = \mathcal{S}$ est le type des étiquettes des noeuds de l'arbre à reconnaitre. (2) $Q = \mathcal{S}$ est le type des états (tous considérés comme finals) et, (3) $q \rightarrow (\sigma, [q_1, \cdots, q_n])$ est une transition de l'automate lorsque la paire $(\sigma, [q_1, \cdots, q_n])$ apparait dans la liste $(gram\ \ q)$[5]. On notera $\mathcal{A}_{\mathbb{G}}$ l'automate d'arbre dérivé à partir de $\mathbb{G}$.

Pour reconnaître un arbre à l'aide d'un automate d'arbre à partir d'un état initial, il suffit de : (1) Associer l'état initial à la racine de l'arbre. (2) Si un noeud étiqueté A est associé à l'état q, alors A doit être égal à q. Si ce noeud possède n successeurs non encore associés à des états, et que la transition $q \rightarrow (A, [q_1, \cdots, q_n])$ est une transition de l'automate, alors on associe les états de q_1 jusqu'à q_n à chacun de ces n successeurs. (3) L'arbre est reconnu si on a ainsi réussi à associer un état à chacun des noeuds de l'arbre. On note $\mathcal{L}(\mathcal{A}, q)$ (langage d'arbre) l'ensemble des arbres acceptés par l'automate $\mathcal{A}$ à partir de l'état initial q.

Comme pour les automates sur les mots, on peut définir un produit synchrone sur les automates d'arbre pour obtenir l'automate reconnaissant l'intersection, l'union, ... des langages réguliers d'arbre [2]. Nous introduisons ci-dessous la définition du produit synchrone de k automates d'arbre dont une adaptation sera utilisée dans la section suivante pour la dérivation de l'automate du consensus.

Définition 4 ***Produit synchrone de*** *k* ***automates :*** *Soient* $\mathcal{A}_1 = (\Sigma, Q^{(1)}, R^{(1)}, q_0^{(1)}), \ldots, \mathcal{A}_k = (\Sigma, Q^{(k)}, R^{(k)}, q_0^{(k)})$ *k automates d'arbre. Le produit synchrone de ces k automates* $\mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_k$ *noté* $\otimes_i \mathcal{A}^{(i)}$*, est l'automate* $\mathcal{A}_{(sc)} = (\Sigma, Q, R, q_0)$ *défini comme suit :* ***(a)*** *Ses états sont les vecteurs d'états :* $Q = Q^{(1)} \times \cdots \times Q^{(k)}$ *;* ***(b)*** *Son état initial est formé par le vecteur des états initiaux des différents automates :* $q_0 = (q_0^{(1)}, \cdots, q_0^{(k)})$ *;* ***(c)*** *Ses transitions sont données par :* $(q^{(1)}, \ldots, q^{(k)}) \xrightarrow{a} \left((q_1^{(1)}, \ldots, q_1^{(k)}), \ldots, (q_n^{(1)}, \ldots, q_n^{(k)})\right) \Leftrightarrow \left(q^{(i)} \xrightarrow{a} (q_1^{(i)}, \ldots, q_n^{(i)}) \quad \forall i,\ 1 \leq i \leq k\right)$

2.2. Notions de vue, de projection, de projection inverse et de fusion

2.2.1. Vue, projection associée et fusion

L'arbre de dérivation donnant la représentation (globale) d'un document structuré édité de façon coopérative, rend visible l'ensemble des symboles grammaticaux de la grammaire ayant participé à sa construction. Comme déjà mentionné à la section 1, pour des raisons de confidentialité (degré d'accréditation), un co-auteur manipulant un tel document n'aura pas accès nécessairement à l'ensemble de tous ces symboles grammaticaux ; seul un sous-ensemble d'entre eux peut être jugé pertinent pour lui : c'est sa *vue*. Une vue $\mathcal{V}$ est donc un sous-ensemble de symboles grammaticaux ($\mathcal{V} \subseteq \mathcal{S}$).

Une réplique partielle de t suivant la vue $\mathcal{V}$, est une copie partielle de t obtenue en supprimant dans t tous les noeuds étiquétés par des symboles n'appartenant pas à $\mathcal{V}$. La figure 2 présente un document t (au centre) ainsi que deux répliques partielles t_{v_1} (à gauche) et t_{v_2} (à droite) obtenues respectivement par projections à partir des vues $\mathcal{V}_1 = \{A, B\}$ et $\mathcal{V}_2 = \{A, C\}$. Pratiquement, une réplique partielle est obtenue via une opération de *projection* notée π. On note donc $\pi_{\mathcal{V}}(t) = t_{\mathcal{V}}$ le fait que $t_{\mathcal{V}}$ est une réplique partielle obtenue par projection de t suivant la vue $\mathcal{V}$.

Notons $t_{\mathcal{V}_i} \leq t_{\mathcal{V}_i}^{maj}$ le fait que le document $t_{\mathcal{V}_i}^{maj}$ soit une m-à-j du document $t_{\mathcal{V}_i}$, c-à-d que $t_{\mathcal{V}_i}^{maj}$ est obtenu de $t_{\mathcal{V}_i}$ en remplaçant certains de ses bourgeons par des arbres. Dans un processus d'édition coopérative asynchrone, il existe des points de synchronisations[6] au cours desquels

5. Rappel : *gram* est l'application obtenue par abstraction de $\mathbb{G}$ et a pour type : $gram : symb \rightarrow [(prod,\ [symb])]$.
6. Un point de synchronisation peut être déf ni statiquement ou déclenché par un co-auteur dès que certaines propriétés sont satisfaites.

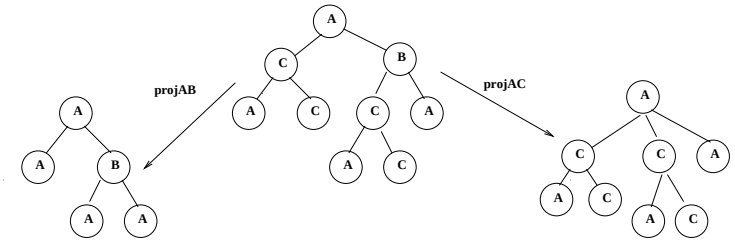


Figure 2. *Exemple de projections effectuées sur un document.*

on essaye de fusionner toutes les contributions $t_{\mathcal{V}_i}^{maj}$ des différents co-auteurs pour obtenir un document global unique t_f[7]. Un algorithme de fusion n'intégrant pas la gestion des conflits et s'appuyant sur une solution au problème de la *projection inverse* est donné dans [1].

2.2.2. Réplique partielle et projection inverse (expansion)

La *projection inverse* d'une réplique partielle mise à jour $t_{\mathcal{V}_i}^{maj}$ relativement à une grammaire $\mathbb{G} = (\mathcal{S}, \mathcal{P}, A)$ donnée, c'est l'ensemble $T_{i\mathcal{S}}^{maj}$ des documents conformes à $\mathbb{G}$ qui admettent $t_{\mathcal{V}_i}^{maj}$ comme réplique partielle suivant la vue $\mathcal{V}_i$: $T_{i\mathcal{S}}^{maj} = \left\{ t_{i\mathcal{S}}^{maj} \therefore \mathbb{G} \mid \pi_{\mathcal{V}_i}(t_{i\mathcal{S}}^{maj}) = t_{\mathcal{V}_i}^{maj} \right\}$.

Une solution à ce problème utilisant les automates d'arbre est donnée dans [1]. Dans celle ci, on se sert des productions de la grammaire $\mathbb{G}$, pour associer à une vue $\mathcal{V}_i \subseteq \mathcal{S}$ un automate d'arbre $\mathcal{A}^{(i)}$ tel que les arbres qu'il reconnait à partir d'un état initial construit à partir de $t_{\mathcal{V}_i}^{maj}$ sont exactement ceux qui ont cette réplique partielle comme projection suivant la vue $\mathcal{V}_i$: $T_{i\mathcal{S}}^{maj} = \mathcal{L}(\mathcal{A}^{(i)},\ q_{t_{\mathcal{V}_i}})$. Le lecteur interessé peut consulter [1] pour une description plus détaillée du processus d'association d'un automate d'arbre à une vue et l'annexe pour une illustration.

3. Réconciliation par consensus

3.1. Problématique et principe de la solution de la réconciliation par consensus

Dans un processus d'édition coopérative asynchrone de plusieurs répliques partielles d'un document, quand on atteint un point de synchronisation, on peut se retrouver avec des répliques non fusionnables dans leur entièreté car, contenant des m-à-j non compatibles[8] : il faut les réconcilier. On peut le faire en remettant en cause (annulation) certaines actions d'édition locales afin de lever les conflits et aboutir à une version globale cohérente dite de consensus.

Les études portant sur la réconciliation des versions d'un document reposent sur des heuristiques [4] dans la mesure où il n'y a pas de solution générale à ce problème. Dans notre cas, étant donné que toutes les actions d'édition sont réversibles[9] et qu'il est facile de localiser les conflits lors de la tentative de fusion des répliques partielles (voir section 3.2), nous disposons d'une méthode canonique pour éliminer les conflits : nous remplaçons lors de la fusion tout noeud (du document global) dont les répliques sont en conflits par un bourgeon. On élague donc au niveau

7. Il peut arriver que l'édition doit être poursuivie après la fusion (c'est le cas s'il existe encore des bourgeons dans le document fusionné) : on doit redistribuer à chacun des n co-auteurs une réplique (partielle) $t_{\mathcal{V}_i}$ de t_f telle que $t_{\mathcal{V}_i} = \pi_{\mathcal{V}_i}(t_f)$ pour la poursuite du processus d'édition.

8. C'est notamment le cas s'il existe au moins un noeud du document global accessible par plus d'un co-auteur et édité par au moins deux d'entre eux en utilisant des productions différentes.

9. Rappel : Les actions d'éditions effectuées sur une réplique partielle peuvent être annulées tant qu'elles n'ont pas encore été intégrées dans le document gobal.

des noeuds où un conflit apparaît, en remplaçant le sous arbre correspondant par un bourgeon du type approprié, indiquant que cette partie du document n'est pas encore éditée : les documents obtenus sont appelés les consensus. Ce sont les préfixes maximaux sans conflits de la fusion des documents issus des différentes expansions des diverses répliques partielles m-à-j.

Le problème de la fusion consensuelle de n répliques patielles dont le modèle global est donné par une grammaire $\mathbb{G} = (\mathcal{S}, \mathcal{P}, A)$ peut donc s'énoncer comme suit :

Problème de la fusion consensuelle : Etant donné n vues $(\mathcal{V}_i)_{1\leq i\leq n}$ et n répliques partielles $(t^{maj}_{\mathcal{V}_i})_{1\leq i\leq n}$, fusionner consensuellement la famille $(t^{maj}_{\mathcal{V}_i})_{1\leq i\leq n}$ consiste à rechercher les plus grands documents $t^{maj}_{\mathcal{S}} \therefore \mathbb{G}$ satisfaisant : $\forall i \in 1,\ldots,n \ \left(\exists t_i \therefore \mathbb{G},\ t_i \leq t^{maj}_{\mathcal{S}},\ \pi_{\mathcal{V}_i}(t_i) \leq t^{maj}_{\mathcal{V}_i}\right)$

La solution que nous proposons à ce problème découle d'une instrumentalisation de celle proposée pour l'expansion (section 2.2.2). En effet, nous nous servons d'un opérateur associatif et commutatif noté $\otimes^{\Omega}$ pour synchroniser les automates d'arbre $\mathcal{A}^{(i)}$ construits pour réaliser les différentes expansions afin de générer l'automate d'arbre de la fusion consensuelle. En notant $\mathcal{A}_{(sc)}$ cet automate, les documents du consensus sont les arbres du langage engendré par l'automate $\mathcal{A}_{(sc)}$ à partir d'un état initial construit à partir du tuple formé des états initiaux des automates $(\mathcal{A}^{(i)}) : \mathscr{L}(\mathcal{A}_{(sc)},\ (q_{t^{maj}_{\mathcal{V}_i}})) = consensus\{\mathscr{L}(\mathcal{A}^{(i)},\ q_{t^{maj}_{\mathcal{V}_i}})\}$. $\mathcal{A}_{(sc)}$ est obtenu en procédant de la façon suivante : (1) Pour chaque vue $\mathcal{V}_i$, construire l'automate $\mathcal{A}^{(i)}$ qui réalisera l'expansion de la réplique partielle $t^{maj}_{\mathcal{V}_i}$ comme indiqué précédemment (sec. 2.2.2) : $\mathscr{L}(\mathcal{A}^{(i)},\ q_{t^{maj}_{\mathcal{V}_i}}) = T^{maj}_{i\mathcal{S}}$. (2) Au moyen de l'opérateur $\otimes^{\Omega}$, calculer l'automate générant le langage du consensus $\mathcal{A}_{(sc)} = \otimes^{\Omega}_i \mathcal{A}^{(i)}$.

3.2. Calcul du consensus

Avant de présenter l'algorithme du calcul du consensus, précisons en utilisant les notions introduites à la section 2.1 que deux documents t_1 et t_2 sont en conflits (et on note $t_1 \Yup t_2$) s'il existe une adresse $w \in Dom(t_1) \cap Dom(t_2)$ telle que l'étiquette du noeud n_1 situé à l'adresse w dans t_1 est différente de celle du noeud n_2 située à la même adresse dans t_2 ; n_1 et n_2 ne sont pas des bourgeons. c-à-d. $(t_1 \Yup t_2) \Leftrightarrow (\exists w \in Dom(t_1) \cap Dom(t_2),\ t_1(w) \neq X_\omega,\ t_2(w) \neq X_\omega,\ t_1(w) \neq t_2(w)))$ [10]

3.2.1. Consensus entre plusieurs (deux) documents

Soient $t_1, t_2 : \mathbb{N}^* \to \mathbf{A}$ deux arbres (documents) en conflits de domaines respectifs $Dom(t_1)$ et $Dom(t_2)$. L'arbre consensuel $t_c : \mathbb{N}^* \to \mathbf{A}$ issu de t_1 et t_2 est tel que : (1) Le domaine de t_c est l'union des domaines des deux arbres auquel on soustrait les éléments appartenant aux domaines des sous-arbres issus des noeuds en conflits (on élague au niveau des noeuds en conflits).

(2) Quand un noeud n_1 de t_1 est en conflit avec un noeud n_2 de t_2, ils apparaissent dans l'arbre consensuel t_c sous la forme d'un (unique) bourgeon. Ainsi,

$$\forall w \in Dom(t_c),\ t_c(w) = \begin{cases} t_1(w) & si \quad t_1(w) = t_2(w) \\ t_1(w) & si \quad t_2(w) = X_\omega \\ t_2(w) & si \quad t_1(w) = X_\omega \\ t_1(w) & si \quad w \notin Dom(t_2) \text{ et } \exists u, v \in \mathbb{N}^* \ tq\ w = u.v,\ t_2(u) = X_\omega \\ t_2(w) & si \quad w \notin Dom(t_1) \text{ et } \exists u, v \in \mathbb{N}^* \ tq\ w = u.v,\ t_1(u) = X_\omega \\ X_\omega & si \quad t_1(w) \neq X_\omega \text{ et } t_2(w) \neq X_\omega \text{ et } t_1(w) \neq t_2(w) \end{cases}$$

3.2.2. Construction de l'automate du consensus

La prise en compte des documents avec des bourgeons nécessite le réaménagement de certains modèles utilisés. Par exemple, dans ce qui suit, nous manipulerons les *automates d'arbre avec états de sortie* en lieu et place des automates d'arbre introduits dans la définition 3. Intuiti-

10. Rappel : on note X_ω l'étiquette d'un bourgeon : $(t(w) = X_\omega) \Leftrightarrow$ le noeud situé à l'adresse w est un bourgeon.

vement, un état q d'un automate est qualifié d'*état de sortie* s'il ne lui est associé qu'une unique transition $q \rightarrow (X_\omega, [])$ permermetant de reconnaitre un arbre réduit à un bourgeon de type $X \in \Sigma$. Ainsi, un automate d'arbre à états de sortie $\mathcal{A}$ est un quintuplet $(\Sigma, Q, R, q_0, exit)$ où Σ, Q, R, q_0 désignent les mêmes objets que ceux introduits dans la définition 3, et *exit* est un prédicat défini sur les états (*exit* : $Q \rightarrow Bool$). Tout état q de Q pour lequel *exit* q est *Vrai* est un état de sortie.

Dans la section 3.2.1 ci-dessus, nous avons dit que, "*quand deux noeuds sont en conflits, ils apparaissent dans l'arbre consensuel sous la forme d'un (unique) noeud ouvert*". Du point de vue de la synchronisation d'automates, la notion de "noeuds en conflits" se traduit par la notion "d'états en conflits" (que nous précisons ci-dessous) et l'extrait précédent se traduit par "*quand deux états sont en conflits, ils apparaissent dans l'automate du consensus sous la forme d'un (unique) état de sortie*". Ainsi, si on considère deux familles de transitions $q_o^1 -> [(a_1^1, qs_1) \dots (a_{n_1}^1, qs_{n_1})]$ et $q_o^1 -> [(a_1^2, qs'_1) \dots (a_{n_2}^2, qs'_{n_2})]$ associées aux états q_0^1 et q_0^2 de deux automates d'arbres $auto_1$ et $auto_2$, on dira que les états q_0^1 et q_0^2 (qui ne sont pas des états de sortie) sont en conflits (et on note $q_0^1 \curlyvee q_0^2$) s'il n'existe pas de transition partant de chacun d'eux et portant la même étiquette, c-à-d. il n'existe pas d'étiquette de transition a^3 telle que (a^3, qs) appartient à $[(a_1^1, qs_1) \dots (a_{n_1}^1, qs_{n_1})]$ et (a^3, qs') appartient à $[(a_1^2, qs'_1) \dots (a_{n_2}^2, qs'_{n_2})]$. Il est alors évident que deux automates donnés admettent un automate consensuel si leurs états initials ne sont pas en conflit.

L'automate synchronisé consensuel $\mathcal{A}_{(sc)} = \otimes_i^\Omega \mathcal{A}^{(i)}$ dont le processus de construction est décrit par l'algorithme 1 est un automate à états de sortie. Il est obtenu en considérant lors du calcul du produit synchrone des automates $\mathcal{A}^{(i)}$ que : (1) quand un automate $\mathcal{A}^{(j)}$ quelconque a atteint un état de sortie[11], il ne contribue plus au comportement mais, ne s'oppose pas à la synchronisation des autres automates : on dit qu'il est *endormi* (algo. 1 ligne 9). (2) un état $q = (q^1, \cdots, q^k)$ est un état de sortie si : (a) tous les états composites q^i sont endormis (algo. 1 ligne 17) ou (b) s'il existe deux états quelconques q^i et q^j, $i \neq j$, composants de l'état q qui sont en conflits (algo. 1 ligne 18) $\Big(exit\ (q^{(1)}, \dots, q^{(k)})\Big) \Leftrightarrow \Big(\Big(exit\ q^{(i)}, \forall i \in \{1 \dots k\}\Big)$ ou $\Big(\exists i, j,\ \ i \neq j,\ q^{(i)} \curlyvee q^{(j)}\Big)\Big)$.

Ainsi, l'automate synchronisé consensuel $\mathcal{A}_{(sc)} = \otimes_i^\Omega \mathcal{A}^{(i)}$ est construit comme suit (algo. 1) :

– Ses états sont les vecteurs d'états : $Q = Q^{(1)} \times \cdots \times Q^{(k)}$;

– Son état initial est formé par le vecteur des états initiaux des différents automates : $q_0 = (q_0^{(1)}, \cdots, q_0^{(k)})$ (algo. 1 ligne 2) ;

– Ses transitions sont données par :

- Si $(exit\ q)$ alors $q \rightarrow (X_\omega, [])$ est une transition de $\mathcal{A}_{(sc)}$ (algo. 1 ligne 18),
- sinon $(q^{(1)}, \dots, q^{(k)}) \xrightarrow{a} \Big((q_1^{(1)}, \dots, q_1^{(k)}), \dots, (q_n^{(1)}, \dots, q_n^{(k)})\Big) \Leftrightarrow \forall i,\ 1 \leq i \leq k$ ou bien
  - $*$ *exit* $q^{(i)}$ et $\Big(q_j^{(i)} = q^{(i)}, \forall j,\ 1 \leq j \leq n\Big)$ /* $q^{(i)}$ *est endormi* */ , sinon
  - $*$ $q^{(i)} \xrightarrow{a} \Big(q_1^{(i)}, \dots, q_n^{(i)}\Big)$

$\otimes_i^\Omega$ est donc une relaxation de la synchronisation d'automates que nous avions introduite dans la définition 4 et $\mathscr{L}(\mathcal{A}_{(sc)},\ (q_{t_{q'_i}^{maj}})) = consensus(\mathscr{L}(\mathcal{A}^{(i)},\ q_{t_{q'_i}^{maj}}))$. Il ne reste plus qu'à appliquer un générateur[13] à $\mathcal{A}_{(sc)}$ comme dans [1] pour obtenir les documents les plus simples qu'il accepte, c-à-d. ceux du consensus.

11. Le noeud correspondant dans le document de la projection inverse est un bourgeon et traduit le fait que l'auteur correspondant ne l'a pas édité. Dans le cas où il serait partagé avec un autre co-auteur l'ayant édité dans sa réplique, c'est l'édition réalisée par ce dernier qui sera retenue lors de la fusion.

13. Il est facile d'écrire un générateur (voir dans [1]) qui à partir d'un automate d'arbre énumère ses arbres acceptés les plus simples, c-à-d., dans aucune branche, un état n'est utilisé plus d'une fois pour la génération des noeuds de la branche.

entrée: - k automates $\mathcal{A}_i = (\Sigma, Q_i, R_i, q_0^{(i)}, exit_i)$ $1 \leq i \leq k$, associés aux k répliques ;
- La grammaire globale $\mathbb{G} = (\mathcal{S}, \mathcal{P}, A)$;

sortie : L'Automate $\mathcal{A}_{(sc)} = (\Sigma, Q, R, q_0, exit) = \otimes_i^{\Omega} \mathcal{A}^{(i)}$ qui accepte les documents du consensus

$R \leftarrow \emptyset$ et $N_i \leftarrow \emptyset$, $\forall i, 1 \leq i \leq k$; /* N_i est un ensemble de transitions */
Au départ, $q_0 = (q_0^{(1)}, \cdots, q_0^{(k)})$ est l'unique état de Q et il est non marqué;
si $\forall i, exit_i\ q_0^{(i)}$ *et* $q_0^{(i)} \rightarrow (X_\omega, [\,]) \in R_i$ **alors**
    Ajouter à R la transition $q_0 \rightarrow (X_\omega, [\,])$; positionner $(exit_i\ q_0)$ à $TRUE$; Retourner $\mathcal{A}_{(sc)}$;
tantque *il existe un état non marqué* $q = (q^{(1)}, \cdots, q^{(k)})$ *dans* Q **faire**
    marquer q;
    si *not* $(exit\ q)$ **alors** /* q n'est pas un état de sortie */
        pourTout $P : X_0 \rightarrow X_1 \ldots X_n \in \mathcal{P}$ *telle que*
        $\left[q^{(i)} \rightarrow \left(P,\ \left[q_1^{(i)}, \ldots, q_n^{(i)}\right]\right) \in R_i \text{ ou } \left(q^{(i)} \rightarrow (X_{0_\omega}, [\,]) \in R_i \cup N_i\right)\right]$, $\forall i, 1 \leq i \leq k$ **faire**
            Ajouter à R la transition $q \rightarrow \left(P, \left[(q_1^{(1)}, \ldots, q_1^{(k)}), \ldots, (q_n^{(1)}, \ldots, q_n^{(k)})\right]\right)$ dans laquelle
            si $exit_i\ q^{(i)}$ **alors** /* $q^{(i)}$ est endormi */
                $q_l^{(i)} = q^{(i)}$, $\forall l,\ 1 \leq l \leq n$
                pourTout $P' : X_i \rightarrow X_1' \ldots X_{m'}' \in \mathcal{P}$ **faire** /* Forward état endormi */
                    Ajouter à N_i la transition $q^{(i)} \rightarrow \left(X_{l_\omega}', [\,]\right), \forall l,\ 1 \leq l \leq m'$;
            pourTout $q_j = (q_j^{(1)}, \ldots, q_j^{(k)}),\ q_j \notin Q,\ 1 \leq j \leq n$ **faire**
                Ajouter q_j à Q[12] ; /* q_j est un nouvel état */;
                si $\left(\forall i,\ exit_i\ q_j^{(i)}\right)$ *ou* /* q_j est un état de sortie */
                $\left(\nexists P'' : X_j \rightarrow X_1'' \ldots X_{m''}'' \in \mathcal{P},\ \left(q^{(i)} \rightarrow \left(P'',\ \left[q_1^{(i)}, \ldots, q_{m''}^{(i)}\right]\right) \in R_i \text{ ou } \left(q^{(i)} \rightarrow (X_{j_\omega}, [\,]) \in R_i \cup N_i\right)\right)\right)$, $\forall i$
                /* les états composites de q_j sont en conflits */ **alors**
                Ajouter à R la transition $q_j \rightarrow \left(X_{j_\omega}, [\,]\right)$ et positionner $(exit\ q_j)$ à $TRUE$;
Retourner $\mathcal{A}_{(sc)}$;

Algorithm 1: Algorithme de construction de l'automate du consensus

3.3. Illustration

La figure 3 est une illustration d'un processus d'édition coopérative asynchrone engendrant des répliques partielles (fig. 3c et fig. 3e) en conflits [14] par rapport à la grammaire constituée des productions suivantes : $P_1 : A \rightarrow C\,B$ $P_2 : A \rightarrow \varepsilon$ $P_3 : B \rightarrow C\,A$
$P_4 : B \rightarrow B\,B$ $P_5 : C \rightarrow A\,C$ $P_6 : C \rightarrow C\,C$ $P_7 : C \rightarrow \varepsilon$

Initialement dans ce processus, deux répliques partielles (fig. 3b et fig. 3d) sont obtenues par projections du document global (fig. 3a). Suite à leur m-à-j (fig. 3c et fig. 3e) un point de synchronisation est atteint et par application de la démarche décrite dans la section 3.1 c-à-d., association des automates d'arbres $\mathcal{A}^{(1)}$ et $\mathcal{A}^{(2)}$ respectivement aux répliques partielles *tv1* et *tv2*, leur synchronisation consensuelle en l'automate $\mathcal{A}_{(sc)} = \mathcal{A}^{(1)} \otimes^{\Omega} \mathcal{A}^{(2)}$ puis, calcul du langage accepté par cet automate, et enfin, génération des documents du consensus (fig. 3h). Rappelons que le présent exemple est explicitement repris en annexe (sec. A) et la figure (fig. 4) de cet annexe donne l'ensemble des documents les plus simples du consensus.

14. En réalisant les expansions respectives de chacune des répliques, on obtient les documents des f gures respectives f g. 3f et f g. 3g sur lesquels on peut observer aisément un conf it mis en exergue par des zones grisées.

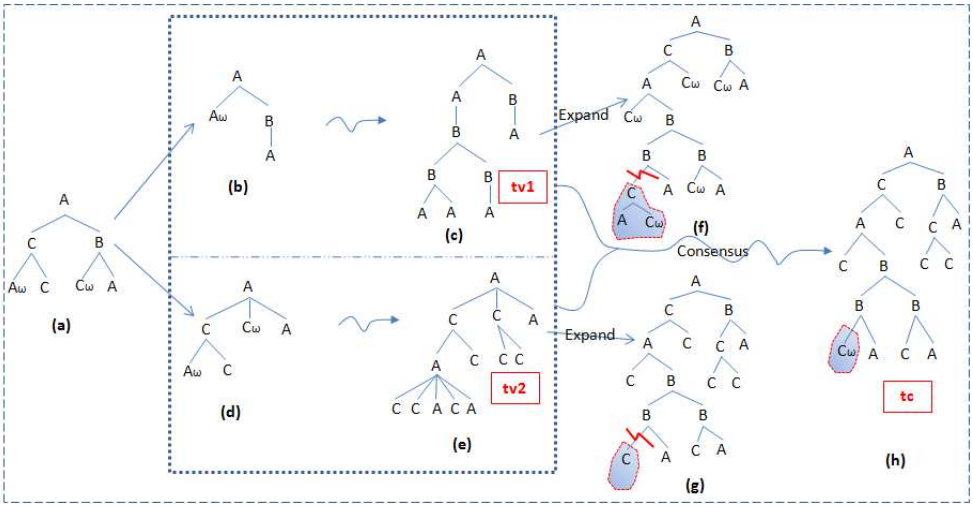


Figure 3. *Exemple de workflow d'une édition avec conflit et consensus correspondant.*

4. Conclusion

Nous avons présenté dans ce papier une approche de réconciliation dite par *consensus*, des répliques partielles d'un document soumis à un processus d'édition coopérative asynchrone. L'approche proposée s'appuie sur une relaxation du produit synchrone d'automates pour construire un automate pouvant générer les documents du consensus.

Les algorithmes présentés dans ce papier ont été implémentés en haskell [3] et expérimentés sur bien des exemples (dont celui explicité en annexe (sec. A)) avec des résultats probants. Nous nous investissons actuellement à la production d'un prototype d'expérimentation via une interface graphique des algorithmes proposés dans un environnement véritablement distribué.

5. Bibliographie

[1] E. Badouel and M. Tchoupé, *«Merging hierarchically structured documents in workflow systems »*, Proceedings of the Ninth Workshop on Coalgebraic Methods in Computer Science (CMCS 2008), Budapest. Electronic Notes in Theoretical Computer Science, 203(5), pp. 3-24, 2008.

[2] H. Comon, M. Dauchet, R. Gilleron, D. Lugiez, S. Tison and M. Tommasi, *«Tree automata techniques and applications »*, Draft, Available at http ://www. grappa. univ-lille3. fr/tata/, 2005.

[3] *Haskell*, A Purely Functional Language. http ://www.haskell.org.

[4] T. Mens, *«A State-of-the-Art Survey on Software Merging »*, Journal of IEEE Transactions on Software Engineering, 28(5), pp. 449-462, 2002.

[5] Y. Saito and M. Shapiro, *«Optimistic replication »*, In ACM Computing Surveys, Vol. V. No. N. 3, pp. 1-44, 2005.

[6] B. Timo and R. Erhard, *«Supporting Efficient Streaming and Insertion of XML Data in RDBMS »*, In Proc. Int. Workshop Data Integration over the Web (DIWeb), pp. 70-81, 2004.

[7] A. VAN DEURSEN, P. KLINT, AND J. VISSER, *«Domain-specific languages : An annotated bibliography »*, ACM SIGPLAN Notices, 35(6), pp. 36-36, June 2000.

A. Annexe

Nous illustrons l'algorithme de fusion consensuelle avec la grammaire de la section 3.3. Nous associons les automates $\mathcal{A}^{(1)}$ et $\mathcal{A}^{(2)}$ respectivement aux répliques partielles m-à-j $tv1$ et $tv2$ (fig. 3c et fig. 3e), puis nous construisons l'automate du consensus $\mathcal{A}_{(sc)} = \mathcal{A}^{(1)} \otimes^{\Omega} \mathcal{A}^{(2)}$ par application de la démarche décrite dans la section 3.2.2 et enfin, présentons les documents les plus simples du consensus (fig. 4).

Linéarisation d'un document structuré

Afin de simplifier la présentation, nous représentons dans ce qui suit les arbres par leur linéarisation sous la forme d'un mot de Dyck. Pour ce faire, nous associons une paire (différentes) de parenthèses à chaque symbole grammaticale et la linéarisation d'un arbre est obtenue en effectuant un parcours en profondeur d'abord de l'arbre résultant.

Les schémas de transition pour la vue $\{A,B\}$

Une liste d'arbres (forêts) est représentée par la concaténation de leurs linéarisations. Nous utilisons la parenthèse ouvrante '(' et fermante ')' pour représenter les symboles de Dyck associés au symbole visible A et le crochet ouvrant '[' et fermant ']' pour représenter ceux associés au symbole visible B. Chaque transition des automates associés aux répliques partielles suivant la vue $\{A,B\}$ est conforme à l'un des schémas de règles suivants :

$$
\begin{array}{llll||llll}
\langle A,w_1\rangle & \longrightarrow & (P_1,[\langle C,u\rangle,\langle B,v\rangle]) & \text{si } w_1=u[v] & \langle A,w_2\rangle & \longrightarrow & (P_2,[\,]) & \text{si } w_2=\varepsilon \\
\langle B,w_3\rangle & \longrightarrow & (P_3,[\langle C,u\rangle,\langle A,v\rangle]) & \text{si } w_3=u(v) & \langle B,w_4\rangle & \longrightarrow & (P_4,[\langle B,u\rangle,\langle B,v\rangle]) & \text{si } w_4=[u][v] \\
\langle C,w_5\rangle & \longrightarrow & (P_5,[\langle A,u\rangle,\langle C,v\rangle]) & \text{si } w_5=(u)v & \langle C,w_6\rangle & \longrightarrow & (P_6,[\langle C,u\rangle,\langle C,v\rangle]) & \text{si } w_6=uv\neq\varepsilon \\
\langle C,w_7\rangle & \longrightarrow & (C_\omega,[\,]) & \text{si } w_7=\varepsilon & & & &
\end{array}
$$

Ces schémas de règles sont obtenus à partir des productions de la grammaire [1] et les couples $\langle X,w_i\rangle$ sont des états dans lesquels X est un symbole grammatical et w_i une forêt codée dans le langage de Dyck. Le premier schéma par exemple, stipule que les arbres de syntaxe abstraite (AST) générés à partir de l'état $\langle A,w_1\rangle$ sont ceux obtenus en utilisant la production P_1 pour créer un arbre de la forme $P_1[t_1,t_2]$; t_1 et t_2 étant générés respectivement à partir des états $\langle C,u\rangle$ et $\langle B,v\rangle$ tel que $w_1=u[v]$. L'état $\langle C,w_7\rangle$ avec $w_7=\varepsilon$ étant un état de sortie [1], la règle $\langle C,w_7\rangle \longrightarrow (C_\omega,[\,])$ liée à la production P_7 stipule que l'AST généré à partir de l'état $\langle C,w_7\rangle$ est réduit à un bourgeon de type C (C est le symbole situé en partie gauche de P_7).

Construction de l'automate $\mathcal{A}^{(1)}$ associé à $tv1$

Ayant associé les symboles de Dyck '(' et ')' (resp. '[' et ']') au symbole grammatical A (resp. B), la linéarisation de la réplique partielle $tv1$ (fig. 3.c) donne $(([[()()][()]])[()])$. A étant l'axiome de la grammaire, l'état initial de l'automate $\mathcal{A}^{(1)}$ est $q_0^1=\langle A,([[()()][()]])[()]\rangle$. En ne considérant que les états accessibles à partir de q_0^1 et en appliquant les schémas de règles présentés précédemment, nous obtenons l'automate d'arbre suivant pour la réplique $tv1$ (fig. 3.c) :

$$
\begin{array}{llll}
q_0^1 & \longrightarrow & (P_1,[q_1^1,q_2^1]) & \text{avec } \ q_1^1=\langle C,([[()()][()]])\rangle \text{ et } q_2^1=\langle B,()\rangle \\
q_1^1 & \longrightarrow & (P_5,[q_3^1,q_4^1]) & \text{avec } \ q_3^1=\langle A,[[()()][()]]\rangle \text{ et } q_4^1=\langle C,\varepsilon\rangle \\
q_1^1 & \longrightarrow & (P_6,[q_4^1,q_1^1]) \mid (P_6,[q_1^1,q_4^1]) & \\
q_2^1 & \longrightarrow & (P_3,[q_4^1,q_5^1]) & \text{avec } \ q_5^1=\langle A,\varepsilon\rangle \\
q_3^1 & \longrightarrow & (P_1,[q_4^1,q_6^1]) & \text{avec } \ q_6^1=\langle B,[()()][()]\rangle \\
q_4^1 & \longrightarrow & (C_\omega,[\,]), & q_5^1 \longrightarrow (P_2,[\,]) \\
q_6^1 & \longrightarrow & (P_4,[q_7^1,q_2^1]) & \text{avec } \ q_7^1=\langle B,()()\rangle \\
q_7^1 & \longrightarrow & (P_3,[q_8^1,q_5^1]) & \text{avec } \ q_8^1=\langle C,()\rangle \\
q_8^1 & \longrightarrow & (P_5,[q_5^1,q_4^1]), & q_8^1 \longrightarrow (P_6,[q_8^1,q_4^1]) \mid (P_6,[q_4^1,q_8^1])
\end{array}
$$

L'état $q_4^1=\langle C,\varepsilon\rangle$ est le seul état de sortie de $\mathcal{A}^{(1)}$. Il est aisé de vérifier que le document de la figure 3f issu de la projection inverse de $tv1$ appartient au langage accepté par l'automate $\mathcal{A}^{(1)}$.

Construction de l'automate $\mathcal{A}^{(2)}$ associé à $tv2$

Comme précédemment, en associant au symbole grammatical C (resp. A) les symboles de Dyck

'[' et ']' (resp. '(' et ')'), on obtient les schémas des transitions pour les automates associés aux répliques partielles suivant la vue $\{A,C\}$.

La linéarisation de la réplique partielle *tv*2 (fig. 3e) donne $([([\,][\,](\,)[\,](\,))[\,]][[\,][\,]](\,))$. L'automate $\mathcal{A}^{(2)}$ associé à cette réplique a pour état initial $q_0^2 = \langle A, [([\,][\,](\,)[\,](\,))[\,]][[\,][\,]](\,)\rangle$ et pour transitions :

$q_0^2 \longrightarrow (P_1, [q_1^2, q_2^2])$ avec $q_1^2 = \langle C, ([\,][\,](\,)[\,](\,))[\,]\rangle$ et $q_2^2 = \langle B, [[\,][\,]](\,)\rangle$

$q_1^2 \longrightarrow (P_5, [q_3^2, q_4^2])$ avec $q_3^2 = \langle A, [\,][\,](\,)[\,](\,)\rangle$ et $q_4^2 = \langle C, \varepsilon\rangle$

$q_2^2 \longrightarrow (P_3, [q_5^2, q_6^2])$ avec $q_5^2 = \langle C, [\,][\,]\rangle$ et $q_6^2 = \langle A, \varepsilon\rangle$

$q_3^2 \longrightarrow (P_1, [q_4^2, q_7^2])$ avec $q_7^2 = \langle B, [\,](\,)[\,](\,)\rangle$

$q_4^2 \longrightarrow (P_7, [\,])$

$q_5^2 \longrightarrow (P_6, [q_4^2, q_4^2])$

$q_6^2 \longrightarrow (P_2, [\,])$

$q_7^2 \longrightarrow (P_4, [q_8^2, q_7^2]) \mid (P_4, [q_9^2, q_{10}^2]) \mid (P_4, [q_{11}^2, q_{11}^2]) \mid (P_4, [q_{12}^2, q_{13}^2]) \mid (P_4, [q_7^2, q_8^2])$ avec $q_8^2 = \langle B, \varepsilon\rangle$, $q_9^2 = \langle B, [\,]\rangle$, $q_{10}^2 = \langle B, (\,)[\,](\,)\rangle$, $q_{11}^2 = \langle B, [\,](\,)\rangle$, $q_{12}^2 = \langle B, [\,](\,)[\,]\rangle$ et $q_{13}^2 = \langle B, (\,)\rangle$

$q_8^2 \longrightarrow (B_\omega, [\,])$

$q_9^2 \longrightarrow (P_4, [q_8^2, q_9^2]) \mid (P_4, [q_9^2, q_8^2])$

$q_{10}^2 \longrightarrow (P_4, [q_8^2, q_{10}^2]) \mid (P_4, [q_{13}^2, q_{11}^2]) \mid (P_4, [q_{14}^2, q_{13}^2]) \mid (P_4, [q_{10}^2, q_8^2])$ avec $q_{14}^2 = \langle B, (\,)[\,]\rangle$

$q_{11}^2 \longrightarrow (P_3, [q_4^2, q_6^2])$

$q_{12}^2 \longrightarrow (P_4, [q_8^2, q_{12}^2]) \mid (P_4, [q_9^2, q_{14}^2]) \mid (P_4, [q_{11}^2, q_9^2]) \mid (P_4, [q_{12}^2, q_8^2])$

$q_{13}^2 \longrightarrow (P_4, [q_8^2, q_{13}^2]) \mid (P_4, [q_{13}^2, q_8^2])$

$q_{14}^2 \longrightarrow (P_4, [q_8^2, q_{14}^2]) \mid (P_4, [q_{13}^2, q_9^2]) \mid (P_4, [q_{14}^2, q_8^2])$

L'état $q_8^2 = \langle B, \varepsilon\rangle$ est le seul état de sortie de l'automate $\mathcal{A}^{(2)}$.

Construction de l'automate du consensus $\mathcal{A}_{(sc)}$

Par application du produit synchrone de plusieurs automates d'arbres décrit dans la section 3.2.2 aux automates $\mathcal{A}^{(1)}$ et $\mathcal{A}^{(2)}$, l'automate du consensus $\mathcal{A}_{(sc)} = \mathcal{A}^{(1)} \otimes^{\Omega} \mathcal{A}^{(2)}$ a pour état initial $q_0 = (q_0^1, q_0^2)$. $\mathcal{A}^{(1)}$ possède une transition de q_0^1 vers $[q_1^1, q_1^1]$ étiquetée P_1. De même, $\mathcal{A}^{(2)}$ possède une transition de q_0^2 vers $[q_1^2, q_2^2]$ étiquetée P_1. On a donc dans $\mathcal{A}_{(sc)}$ une transition etiquetée P_1 permettant d'accéder aux états $[q_1 = (q_1^1, q_1^2), q_2 = (q_2^1, q_2^2)]$ à partir de l'état initial $q_0 = (q_0^1, q_0^2)$. Suivant ce principe, on construit l'automate consensuel suivant :

$q_0 = (q_0^1, q_0^2)$

$q_0 \longrightarrow (P_1, [q_1, q_2])$ avec $q_1 = (q_1^1, q_1^2)$ et $q_2 = (q_2^1, q_2^2)$

$q_1 \longrightarrow (P_5, [q_3, q_4])$ avec $q_3 = (q_3^1, q_3^2)$ et $q_4 = (q_4^1, q_4^2)$

$q_2 \longrightarrow (P_3, [q_5, q_6])$ avec $q_5 = (q_4^1, q_5^2)$ et $q_6 = (q_5^1, q_6^2)$

$q_3 \longrightarrow (P_1, [q_4, q_7])$ avec $q_7 = (q_6^1, q_7^2)$

$q_4 \longrightarrow (P_7, [\,])$,

$q_5 \longrightarrow (P_6, [q_4, q_4])$,

$q_6 \longrightarrow (P_2, [\,])$

$q_7 \longrightarrow (P_4, [q_8, q_9]) \mid (P_4, [q_{10}, q_{11}]) \mid (P_4, [q_{12}, q_{13}]) \mid (P_4, [q_{14}, q_{15}]) \mid (P_4, [q_{16}, q_{17}])$ avec $q_8 = (q_7^1, q_8^2)$, $q_9 = (q_2^1, q_7^2)$, $q_{10} = (q_7^1, q_9^2)$, $q_{11} = (q_2^1, q_{10}^2)$, $q_{12} = (q_7^1, q_{11}^2)$, $q_{13} = (q_2^1, q_{11}^2)$, $q_{14} = (q_7^1, q_{12}^2)$, $q_{15} = (q_2^1, q_{13}^2)$, $q_{16} = (q_7^1, q_7^2)$ et $q_{17} = (q_2^1, q_8^2)$

$q_8 \longrightarrow (P_3, [q_{18}, q_{19}])$ avec $q_{18} = (q_8^1, q_8^2)$ et $q_{19} = (q_5^1, q_8^2)$

$q_{12} \longrightarrow (P_3, [q_{20}, q_6])$ avec $q_{20} = (q_8^1, q_4^2)$

$q_{13} \longrightarrow (P_3, [q_4, q_6])$

$q_{17} \longrightarrow (P_3, [q_{21}, q_{19}])$ avec $q_{21} = (q_4^1, q_8^2)$

$q_{18} \longrightarrow (P_5, [q_{19}, q_{21}]) \mid (P_6, [q_{18}, q_{21}]) \mid (P_6, [q_{21}, q_{18}])$

$q_{19} \longrightarrow (P_2, [\,])$, $q_9 \longrightarrow (B_\omega, [])$ $q_{10} \longrightarrow (B_\omega, [])$ $q_{11} \longrightarrow (B_\omega, [])$ $q_{14} \longrightarrow (B_\omega, [])$ $q_{15} \longrightarrow (B_\omega, [])$ $q_{16} \longrightarrow (B_\omega, [])$ $q_{20} \longrightarrow (C_\omega, [])$ $q_{21} \longrightarrow (C_\omega, [])$

Les états $\{q_9, q_{10}, q_{11}, q_{14}, q_{15}, q_{16}, q_{20}, q_{21}\}$ sont les états de sortie de l'automate $\mathcal{A}_{(sc)}$. Ce sont des états dont les états composites sont soit en conflits (par exemple $q_9 = (q_2^1, q_7^2)$ et $q_2^1 \between q_7^2$), soit sont tous des états de sortie (par exemple $q_{21} = (q_4^1, q_8^2)$).

L'utilisation de la fonction de génération des AST (avec bourgeons) les plus simples d'un langage d'arbre à partir de sont automate [1] sur l'automate $\mathcal{A}_{(sc)}$, produit *quatre* AST dont les arbres de dérivations (les consensus) sont schématisés sur la figure 4.

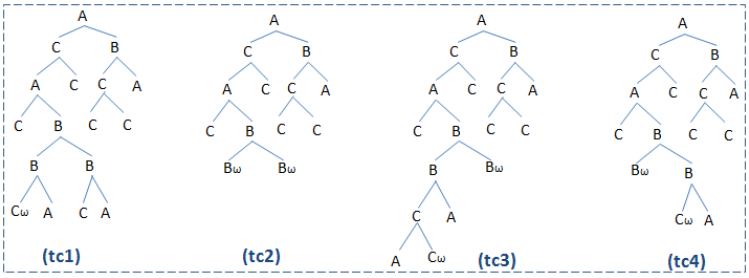


Figure 4. *Arbres consensuels générés à partir de l'automate* $\mathcal{A}_{(sc)}$

Un dépliage par processus pour calculer le préfixe complet des réseaux de Petri

Médésu Sogbohossou — Antoine Vianou

Département Génie Informatique et Télécommunications
École Polytechnique d'Abomey-Calavi, 01 BP 2009 Cotonou, BENIN
{medesu.sogbohossou,antoine.vianou}@epac.uac.bj

RÉSUMÉ. La technique d'ordre partiel du dépliage représente implicitement l'espace d'état d'un réseau de Petri (RdP), en conservant notamment les relations de concurrence entre les événements. Cela permet de contenir le phénomène de l'explosion combinatoire en cas de forte concurrence. Un préfixe complet de dépliage sert à couvrir tout l'espace d'état d'un RdP borné: son calcul suivant l'approche classique se base sur le concept d'ordre adéquat, ne prenant directement en compte que les RdP saufs. Dans cet article, une nouvelle approche indépendante du concept d'ordre adéquat et fidèle à la sémantique d'ordre partiel, consiste à créer les événements du dépliage dans le contexte d'un unique processus à la fois. Les résultats des tests sont concluants pour les RdP saufs et non saufs. Pour améliorer la compacité du préfixe obtenu, deux solutions sont présentées.

ABSTRACT. The partial-order technique of the unfolding implicitly represents state-space of a Petri net (PN), by in particular preserving the concurrency relations between the events. That makes it possible to contain state-space explosion problem in case of strong concurrency. A complete prefix of unfolding is used to cover all the state-space of a bounded PN: its computation according to the classical approach is based on the concept of adequate order, taking directly into account only safe PN. In this paper, a new approach independent of the concept of adequate order and faithful to the partial-order semantics, consists in creating the events of the unfolding in the context of a single process at the same time. The results of the tests are conclusive for safe and nonsafe PN. To improve compactness of the prefix obtained, two solutions are presented.

MOTS-CLÉS : réseaux de Petri bornés, préfixe complet de dépliage, ordre adéquat, processus alternatifs

KEYWORDS : bounded Petri nets, complete prefix of unfolding, adequate order, alternative processes

1. Introduction

Les réseaux de Petri (RdP) [11] constituent un des formalismes bien connus pour modéliser de manière compacte et explicite la concurrence et la synchronisation entre composantes dynamiques des systèmes à événement discret. Le modèle établi permet alors de conduire des vérifications de propriétés sur le système représenté, en passant généralement par la construction de l'espace d'état. Toutefois, l'énumération exhaustive des états globaux, sous forme d'un graphe d'état, est exponentielle avec la taille du modèle en cas de concurrence : on parle d'explosion combinatoire. Les techniques dites d'*ordre partiel* constituent un des moyens utilisés pour endiguer ce problème. Ainsi, les réductions d'ordre partiel [15, 16, 17] visent à générer un espace d'état réduit, en économisant les entrelacements des événements concurrents au cours de la construction du graphe d'état. La technique d'ordre partiel du dépliage [2] est une alternative qui préserve une représentation des états globaux, mais de manière implicite en conservant notamment les relations de concurrence entre les états locaux des composantes et entre les événements. Des travaux récents [12, 1, 13] montrent que ces différentes techniques sont toujours en cours d'amélioration. Par exemples, l'article [12] propose un compromis entre rapidité [5] et moindre coût mémoire [8] du dépliage, et l'article [13] intègre certains atouts des dépliages des RdP aux techniques de réduction d'ordre partiel dites dynamiques.

Le calcul d'un préfixe fini du dépliage permet de capturer l'espace d'état du réseau de Petri : ce préfixe est alors dit *complet* [10, 6]. L'approche classique [6] (et ses généralisations dans [9, 1]) de calcul d'un préfixe complet se base sur le concept d'*ordre adéquat* qui exclut la catégorie des RdP non saufs. Dans cet article, un nouvel algorithme se passant du concept d'ordre adéquat est défini : il donne des résultats satisfaisants pour les réseaux saufs. De plus, cet algorithme prend en compte le dépliage des RdP bornés non saufs, avec le souci de préserver la concurrence. En effet, pour cette classe de réseaux, l'approche actuelle [6] consiste à passer par une conversion vers un modèle sauf, ce qui fait perdre l'expression des relations de concurrence.

La particularité du nouvel algorithme consiste à créer les événements du dépliage dans le contexte d'un unique processus à la fois, à l'instar de travaux précédents [14] qui sont valables pour une classe restreinte de réseaux temporels ; ici, aucune restriction ne s'applique à la forme des processus générés pour obtenir un préfixe complet qui soit fini. Ainsi, les événements ne sont plus créés en permettant le développement simultané de plusieurs processus en conflit, ce qui évite le recours au concept d'ordre adéquat. Pour améliorer la compacité du préfixe obtenu, nous esquissons des solutions en envisageant d'une part la détection et la suppression de redondance entre processus alternatifs, et d'autre part en éliminant les auto-conflits apparaissant dans les réseaux non saufs.

La section 2 rappelle les définitions sur les RdP et le dépliage. Ensuite, la section 3 présente le nouvel algorithme, son principe, les résultats de sa mise en œuvre, ainsi que les améliorations qui peuvent être intégrées à l'algorithme. Enfin, la section 4 présente la synthèse des résultats et énonce les perspectives à plus long terme.

2. Rappels

2.1. Réseaux de Petri

Définition 1. *Un* réseau de Petri *(ou* RdP*) est un triplet* $N \stackrel{\text{def}}{=} < P, T, W >$ *:*

– P et T sont resp. les ensembles des places et des transitions : $P \cap T = \emptyset$;

– $W \subseteq P \times T \cup T \times P$ est la relation de flux.

Pour un RdP destiné au calcul d'un espace d'état fini, P et T sont finis. L'ensemble des nœuds prédécesseurs (resp. successeurs) d'un nœud $x \in P \cup T$ est noté ${}^\bullet x \stackrel{\text{def}}{=} \{y \in P \cup T \mid (y,x) \in W\}$ (resp. $x^\bullet \stackrel{\text{def}}{=} \{y \in P \cup T \mid (x,y) \in W\}$).

Un *marquage* est une application $m : P \to \mathbb{N}$: il est interprété comme un état global du système. Le doublet $< N, m_0 >$ représente le RdP N de marquage initial m_0.

Une transition t est *sensibilisée* par un marquage m, ce qui est noté $m \xrightarrow{t}$, si ${}^\bullet t \subseteq m$. Le *tir* de t conduisant au marquage $m' = m \backslash {}^\bullet t \cup t^\bullet$ est noté $m \xrightarrow{t} m'$. Soit $m_0 \xrightarrow{\sigma} m$ t.q. $\sigma = t_1 t_2 ... t_n \in T^*$: σ désigne une séquence de tirs à partir de m_0.

L'*ensemble d'accessibilité* du RdP marqué $< N, m_0 >$ est défini par : $A(N, m_0) \stackrel{\text{def}}{=} \{m \mid \exists \sigma \in T^*, m_0 \xrightarrow{\sigma} m\}$. $< N, m_0 >$ est *borné* si $\exists n \in \mathbb{N}$ t.q. pour tout marquage $m \in A(N, m_0)$, $m(p) \leq n$, $\forall p \in P$. L'ensemble d'accessibilité est fini *ssi* le RdP est borné. Un marquage m *sauf* signifie $m(p) \leq 1, \forall p \in P$: pour un RdP sauf, tous les marquages accessibles sont saufs.

$A(N, m_0)$ fini se représente sous la forme d'un *graphe des marquages* (ou *graphe d'état*) : les nœuds sont les marquages et les arcs représentent les tirs de transition entre couples de marquages directement accessibles.

2.2. Dépliage

Un dépliage prend la forme d'un RdP $O \stackrel{\text{def}}{=} < B, E, F >$ acyclique, dénommé *réseau d'occurrence*, t.q. : $\forall b \in B, |{}^\bullet b| \leq 1$, $\forall e \in E, {}^\bullet e \neq \emptyset$ et $e^\bullet \neq \emptyset$, et F^+ (la fermeture transitive de F) est une relation d'ordre strict.

B (resp. E) est dénommé ensemble des *conditions* (resp. ensemble des *événements*). Pour $e \in E$, ${}^\bullet e$ (resp. $e^\bullet$) forme les *pré-conditions* (resp. *post-conditions*) de e.

On définit : $Min(O) \stackrel{\text{def}}{=} \{b \in B \mid {}^\bullet b = \emptyset\}$ et $Max(O) \stackrel{\text{def}}{=} \{b \in B \mid b^\bullet = \emptyset\}$.

Trois types de relations sont définis entre deux nœuds quelconques de O :

– la causalité ($\prec$) : $\forall x, y \in B \cup E, x \prec y$ ssi $(x, y) \in F^+$;

– le conflit ($\sharp$) : $\forall e_1, e_2 \in E\ (e_1 \neq e_2),\ e_1 \sharp e_2$ si ${}^\bullet e_1 \cap {}^\bullet e_2 \neq \emptyset$. De plus, si $e_1 \sharp e_2$, alors $\forall x, y \in B \cup E,\ e_1 \preceq x \wedge e_2 \preceq y \ \Rightarrow\ x \sharp y$;

– et la concurrence ($\wr$) : $\forall x, y \in B \cup E\ (x \neq y), x \wr y$ ssi $\neg((x \prec y) \vee (y \prec x) \vee (x \sharp y))$.

Soit $B' \subseteq B$ t.q. $\forall b, b' \in B', b \neq b' \Rightarrow b \wr b'$: B' est appelé une *coupe*.

Soient le réseau d'occurrence $O_F \stackrel{\text{def}}{=} < B_F, E_F, F_F >$ et la fonction d'étiquetage $\lambda_F : B_F \cup E_F \to P \cup T$ t.q. $\lambda(B_F) \subseteq P$ et $\lambda(E_F) \subseteq T$.

Définition 2. Le dépliage (exhaustif) *[14]* $Unf_F \stackrel{\text{def}}{=} < O_F, \lambda_F >$ *de* $< N, m_0 >$ *est donné par :*

1) *$\forall p \in P$, si $m_0(p) \neq \emptyset$, alors $B_p \stackrel{\text{def}}{=} \{b \in B_F \mid \lambda_F(b) = p \wedge {}^\bullet b = \emptyset\}$ et $m_0(p) = |B_p|$;*

2) *$\forall B_t \subseteq B_F$ t.q. B_t est une coupe, si $\exists t \in T, \lambda_F(B_t) = {}^\bullet t \wedge |B_t| = |{}^\bullet t|$, alors :*

- ***a)*** *$\exists! e \in E_F$ t.q. ${}^\bullet e = B_t \wedge \lambda_F(e) = t$;*
- ***b)*** *si $t^\bullet \neq \emptyset$, alors $B'_t \stackrel{\text{def}}{=} \{b \in B_F \mid {}^\bullet b = \{e\}\}$ est t.q. $\lambda_F(B'_t) = t^\bullet \wedge |B'_t| = |t^\bullet|$;*
- ***c)*** *si $t^\bullet = \emptyset$, alors $B'_t \stackrel{\text{def}}{=} \{b \in B_F \mid {}^\bullet b = \{e\}\}$ est t.q. $\lambda_F(B'_t) = \emptyset \wedge |B'_t| = 1$;*

3) *$\forall B_t \subseteq B_F$, si B_t n'est pas une coupe, alors $\nexists e \in E_F$ t.q. ${}^\bullet e = B_t$.*

La définition 2 exprime succinctement l'algorithme d'un dépliage exhaustif. Les articles de Engelfriet [2] et Esparza et al. [6] par exemples en donnent une définition plus explicite.

Soit $E \subset E_F$. Le réseau d'occurrence $O \stackrel{\text{def}}{=} < B, E, F >$ associé à E tel que $B \stackrel{\text{def}}{=} \{b \in B_F \mid \exists e \in E, b \in {}^\bullet e \cup e^\bullet\}$, $F \stackrel{\text{def}}{=} \{(x,y) \in F_F \mid x \in E \vee y \in E\}$ et $Min(O) = Min(O_F)$ est *un préfixe* de O_F. Par extension, $Unf \stackrel{\text{def}}{=} < O, \lambda >$ (avec λ, la restriction de λ_F à $B \cup E$) est un préfixe du dépliage Unf_F.

Si les événements E du préfixe de dépliage Unf sont tels que $\forall (e, e') \in E \times E$, on a $\neg(e \sharp e')$, alors E constitue un *processus*.

Soit E_i un processus fini. Le réseau $C_i \stackrel{\text{def}}{=} < B_i, E_i, F_i >$ associé est appelé *réseau causal*. Il vérifie : $\forall b \in B_i, |b^\bullet| \leq 1$. $Max(C_i)$ est l'état final de C_i : il correspond au marquage final du comportement exprimé par E_i, à savoir le multi-ensemble de jetons du RdP résultant de $\lambda(Max(C_i))$, et qui est noté $Mark(E_i)$.

La *configuration locale* d'un événement $e_i \in E_i$ est le processus $E_{e_i} \stackrel{\text{def}}{=} \{e_j \in E_F \mid e_j \prec e_i \vee e_j = e_i\}$. Le marquage $Mark(E_{e_i})$ sera appelé *marquage propre* de e_i.

A l'instar d'un graphe d'état, un préfixe fini de dépliage peut capturer l'espace d'état du RdP : le préfixe est alors dit *complet*.

Définition 3. *Un préfixe* $Unf \stackrel{\text{def}}{=} << B, E, F >, \lambda >$ *de* Unf_F *est* complet *lorsque, pour tout marquage accessible* m *de* $< N, m_0 >$*, il existe un processus* $E_i \subseteq E$ *t.q. :*

1) $m = Mark(E_i)$*,*

2) si $\exists t \in T$ *t.q.* ${}^\bullet t \subseteq m$*, alors* $\exists e \in E$ *t.q.* ${}^\bullet e \subseteq Max(C_i) \wedge \lambda(e) = t$*.*

En pratique, pour représenter tous les marquages de $A(N, m_0)$ sans énumération exhaustive de l'espace d'état, la création de chaque événement du dépliage est soumise à une comparaison entre son marquage propre et ceux des événements déjà produits [10, 6].

2.3. Principe des algorithmes de calcul de préfixe complet

Au cours du calcul d'un dépliage complet, un événement à produire est candidat *cut-off* lorsque son marquage propre est équivalent à celui d'un événement (que nous qualifions de *référence*) précédemment ajouté dans le dépliage : ceci autorise à ne pas calculer les événements successeurs d'un événement cut-off. Mais contrairement à un graphe d'état où une comparaison de marquage suffit, avoir deux marquages propres identiques n'est qu'une des conditions nécessaires pour identifier effectivement un cut-off.

L'algorithme de Esparza *et al.* [6], plus général et plus optimal que celui proposé par McMillan [10], se base sur le concept d'*ordre adéquat*. Une relation d'ordre adéquat $<$ identifie les événements cut-off en comparant les configurations locales des événements produits au cours du dépliage : c'est l'élément (une configuration locale) *plus grand* qui peut être cut-off. Bien entendu, le calcul du dépliage complet consiste à produire les événements possibles un à un suivant l'ordre $<$ de leurs configurations locales, afin de toujours produire en premier les événements de référence potentiels. L'adéquation de la relation $<$ devra être préservée par toute extension en événement d'une configuration : ceci n'est garanti que pour les réseaux saufs (cf. annexe A pour plus de détails).

Les travaux subséquents [7, 8, 9, 12, 1] sur le dépliage reposent également sur ce concept. La prise en compte des RdP non saufs passe par une conversion en RdP sauf [6], avec pour conséquence une perte de concurrence qui peut nuire à la compacité du résultat.

3. Le nouvel algorithme

3.1. Processus alternatifs

Soit un dépliage $<< B, E, F >, \lambda >$. Ses événements peuvent toujours être décomposés en un ensemble $\overline{E}$ de processus tel que :

– $\bigcup_{E_i \in \overline{E}} E_i = E$ et,

– $\forall (E_i, E_j) \in \overline{E} \times \overline{E}$, si $E_i \neq E_j$ alors les deux processus sont en conflit ; ce qui signifie qu'il existe $(e_i, e_j) \in E_i \times E_j$ t.q. $e_i \sharp e_j$ et ${}^\bullet e_i \cup {}^\bullet e_j \subseteq B_i \cap B_j$.

Les processus de $\overline{E}$ sont ainsi qualifiés de *processus alternatifs*.

L'ensemble $\overline{E}$ sur un préfixe complet E est à rapprocher de l'ensemble des séquences maximales dont le calcul produit le graphe d'état. Dans le contexte de la sémantique d'ordre partiel, un processus alternatif E_i est caractérisé par un état final qui, soit est sans successeur (marquage mort), soit constitue la répétition d'un marquage interne d'un certain processus de E. Dans ce dernier cas, les éléments de E_i sans événement successeur dans E sont évidemment tous cut-off. Un processus alternatif est *maximal* s'il n'a pas d'événement successeur dans E.

3.2. Dépliage par processus : principe et algorithme

Dans un graphe d'état, le calcul des successeurs d'un état global ne dépend pas des événements qui y ont conduit : cet état est une convergence de toutes les séquences de tir possibles pouvant y conduire.

Par contre, dans un dépliage, même si des séquences (ou entrelacements) de processus alternatifs aboutissent à un même état global de RdP, leur représentation est distincte (sous forme de coupes). La décision de calculer les successeurs d'un seul des états globaux équivalents implique que toutes ses extensions possibles sont à ajouter à l'espace d'état, tandis que celles des autres états équivalents ne devraient pas l'être pour éviter des redondances. Les algorithmes actuels de dépliage n'énumérant pas des états globaux, le risque est que des extensions nécessaires soient ajoutées de manière dispersée à partir de divers états équivalents, forçant ainsi à calculer les dérivations de plusieurs états au lieu d'un seul. Un exemple de Esparza *et al.* (reproduit à la figure 3 en annexe A), qui a servi à introduire le concept d'ordre adéquat, a montré comment des cut-offs dispersés sur la succession de plusieurs états équivalents a occasionné l'arrêt prématuré d'un dépliage.

A la différence du concept d'ordre adéquat, la solution adoptée ici est d'éviter de générer simultanément des états globaux équivalents et incompatibles (du fait de conflit). L'idée est de ne pas produire un dépliage mêlant les productions concomitantes d'événements appartenant à des processus alternatifs. Le dépliage est obtenu en calculant les processus alternatifs dans un ordre total, ce qui permet d'éviter le scénario décrit précédemment. Dans ce contexte, la définition d'un événement cut-off est simplifiée :

Définition 4. *Un événement* cut-off e *est tel que* $\exists e' \in E, Mark(E_e) = Mark(E_{e'}) \wedge (e' \prec e \vee e' \sharp e)$.

Bien entendu, cette définition exclut la concurrence qui traduit que le processus $E_e \cup E_{e'}$ induit un état global à représenter, avec des dérivations potentielles.

En nous affranchissant du besoin d'ordre adéquat entre les événements, tel que défini dans [6], l'avantage immédiat apporté par ce nouveau principe est qu'on ne se restreint plus aux RdP saufs. L'algorithme 1 est l'implémentation proposée.

1 Créer B_0 ; $B \leftarrow B_0$;
2 $E \leftarrow \{e_0\}$; $Ext \leftarrow \emptyset$; $CurrentProcess \leftarrow \emptyset$; $\overline{E} \leftarrow \emptyset$;
3 $NewExt \leftarrow \{e \in E_F \mid {}^\bullet e \subseteq B_0\}$;
4 $Ext \leftarrow NewExt$;
5 **tant que** $Ext \neq \emptyset$ **faire**
6 **si** $\exists e \in Ext \mid \forall e' \in CurrentProcess, \neg(e \sharp e')$ **alors**
7 $CurrentProcess \leftarrow CurrentProcess \cup \{e\}$;
8 **sinon**
9 $\overline{E} \leftarrow \overline{E} \cup \{CurrentProcess\}$;
10 Choisir $e \in Ext$;
11 $CurrentProcess \leftarrow E_e$;
12 **fin**
13 $Ext \leftarrow Ext \backslash \{e\}$;
14 $E \leftarrow E \cup \{e\}$;
15 $Post\_e \leftarrow e^\bullet$; $B \leftarrow B \cup Post\_e$;
16 **si** *e n'est pas cut-off* **alors**
17 $NewExt \leftarrow \{e \in E_F \mid ({}^\bullet e \subseteq B) \wedge ({}^\bullet e \cap Post\_e \neq \emptyset)\}$;
18 $Ext \leftarrow Ext \cup NewExt$;
19 **fin**
20 **fin**

Algorithm 1. *Préfixe complet de dépliage par processus*

$NewExt$ contient les nouvelles extensions possibles dues aux post-conditions créées après la production de chaque nouvel événement e. L'état initial[1] B_0 est créé par l'événement fictif e_0. $CurrentProcess$ contient les événements du processus maximal en cours de calcul, par ajout un à un des extensions possibles et compatibles de Ext (lignes 6 et 7). Le développement du processus est arrêté lorsque plus aucun ajout d'extension compatible n'est pas possible (ligne 8). Un nouveau processus alternatif est alors chargé dans $CurrentProcess$ (lignes 8 à 11), après la production d'une des extensions de Ext (toutes incompatibles avec les processus alternatifs précédents) et en intégrant sa configuration locale. Ainsi, les différents processus alternatifs sont dépliés un à un. Le dépliage est complet quand il n'existe plus d'extension possible (Ext devient vide).

En gros, l'algorithme 1 ne diffère essentiellement des précédents que par le fait que les extensions possibles ne s'ajoutent pas suivant un ordre adéquat entre les configurations locales, mais suivant un ordre imposé par les processus alternatifs générés un à un. Avec les méthodes classiques de calcul de préfixe complet, les extensions choisies permettent au contraire de développer des processus alternatifs simultanément.

En matière de complexité, l'algorithme ne remet pas en cause l'affirmation selon laquelle le facteur dominant est le calcul des extensions possibles [6] (cf. lignes 3 et 17). En effet, les spécificités de l'algorithme sont d'une complexité qui reste d'ordre polynomial. Ainsi, le choix d'une extension (ligne 6) nécessite un test de conflit avec chacun des événements du processus courant, comparé au choix d'un élément minimal ($<$) de [6]. Et le test d'événement cut-off (ligne 16) semble globalement plus simple que celui de [6], qui peut nécessiter de comparer des configurations locales en plusieurs étapes (cf. annexe

1. On admet que m_0 est produit par l'événement fictif e_0, qui est pris en compte par les implémentations disponibles de [6] : en effet, e_0 peut constituer un événement de référence.

A). Enfin, le chargement d'un nouveau processus alternatif (la configuration locale d'une extension ajoutée) est une opération spécifique à l'algorithme, et qui a aussi un certain coût.

3.3. Résultats expérimentaux

L'annexe A illustre le principe du calcul par processus et montre sur quelques exemples de réseaux la spécificité de l'algorithme comparé aux autres algorithmes.

Des résultats de tests sur un bon nombre de RdP bornés, saufs ou non, sont présentés en annexe B. Ils sont comparés avec ceux obtenus avec l'algorithme classique (pour les RdP saufs) : les deux algorithmes ne donnent pas toujours le même résultat structurel. Nos tests ont toutefois montré que l'espace d'état (généré par un graphe d'état) est toujours couvert, y compris pour les réseaux non saufs.

3.4. Discussions

Le principe du dépliage par processus permet d'intégrer aisément le test pour identifier un réseau non borné (en assurant ainsi la terminaison de l'algorithme comme pour le calcul d'un graphe d'état) : il suffira de mémoriser les états globaux visités par chaque processus au cours du dépliage, leur nombre restant proportionnel au nombre d'événements créés. En effet, un dépliage par processus est similaire au calcul en profondeur d'un graphe d'état.

Avec les réseaux non saufs, le résultat du dépliage peut être parfois moins efficace qu'une conversion préalable en RdP sauf [6] (ce qui fait perdre l'expression de la concurrence), en raison des auto-conflits (cf. annexe B.2). Nous préconisons la suppression des répétitions de mêmes instances de transition en conflit et sous forme d'événements cut-off, afin d'avoir un résultat toujours au moins aussi efficace que celui obtenu via une conversion en réseau sauf.

L'algorithme s'appuie sur le calcul de processus alternatifs pour aboutir au dépliage complet. Nos tests ont révélé que plusieurs processus alternatifs pouvaient converger vers le même marquage final (avec au besoin leurs développements in extenso, par simulation du dépliage, pour les rendre maximaux). Dans certains cas, cela offre la possibilité de réduire le préfixe généré au cours du dépliage, par raccourcissement des processus alternatifs avec un suffixe redondant. Cela peut permettre de minimiser un préfixe complet à l'instar des travaux de [7]. Des illustrations sont données en annexe C.

4. Conclusion et perspectives

Dans cet article, nous avons proposé un nouvel algorithme de calcul du préfixe complet de dépliage des réseaux de Petri bornés, valable pour les réseaux non saufs. Il consiste à obtenir les processus alternatifs nécessaires pour composer un préfixe complet. Sans pour autant s'appuyer sur le concept d'ordre adéquat de [6], il produit un espace d'état complet comme l'attestent de nombreux exemples. Faute de place ici, les preuves de finitude et de complétude ne sont pas fournies.

Une suite immédiate des travaux consisterait à s'intéresser à l'optimisation du coût du dépliage [12], comme évoqué dans [1] avec la stratégie du calcul en profondeur.

Plus généralement, une perspective serait le développement d'algorithmes de vérification des propriétés sur les RdP basée sur les processus alternatifs maximaux. Vérifier par exemple les propriétés génériques à l'aide du dépliage devrait être moins coûteux qu'avec

un graphe d'état généré en entrelaçant les tirs de transitions concurrentes. Ceci pourra être comparé avec des travaux similaires [3]. Une autre perspective est de se servir du dépliage pour obtenir une réduction d'ordre partiel, à l'instar des travaux dans [13].

5. Bibliographie

[1] Blai Bonet, Patrik Haslum, Victor Khomenko, Sylvie Thiébaux, and Walter Vogler. Recent advances in unfolding technique. *Theoretical Computer Science*, 551 :84–101, 2014.

[2] Joost Engelfriet. Branching processes of Petri nets. *Acta Informatica*, 28(6) :575– 591, june 1991.

[3] Javier Esparza and Keijo Heljanko. Unfoldings : A Partial-Order Approach to Model Checking. *Monographs in Theoretical Computer Science. An EATCS Series.. Springer Publishing Company*, Incorporated, 1 edition, 2008.

[4] Javier Esparza, Pradeep Kanade, and Stefan Schwoon. A negative result on depth-first net unfoldings. *International Journal on Software Tools for Technology Transfer (STTT)*, 10(2) :161–166, 2008.

[5] Javier Esparza and Stefan Römer. An unfolding algorithm for synchronous products of transition systems. In *CONCUR'99 Concurrency Theory*, pages 2–20. Springer, 1999.

[6] Javier Esparza, Stefan Römer, and Walter Vogler. An improvement of McMillan's unfolding algorithm. *Formal Methods in System Design*, 20(3) :285–310, 2002.

[7] Keijo Heljanko. Minimizing finite complete prefixes. *Proceedings of the Workshop Concurrency, Specification & Programming 1999*, pages 83–95, Warsaw, Poland, September 1999. Warsaw University.

[8] Victor Khomenko and Maciej Koutny. Towards an efficient algorithm for unfolding Petri nets. In *CONCUR 2001 - Concurrency Theory*, pages 366–380. Springer, 2001.

[9] Victor Khomenko, Maciej Koutny, and Walter Vogler. Canonical prefixes of Petri net unfoldings. *Acta Informatica*, 40(2) :95–118, 2003.

[10] Kenneth L. McMillan. A technique of state space search based on unfolding. *Form. Methods Syst. Des.*, 6(1) :45–65, 1995.

[11] Tadao Murata. Petri nets : Properties, analysis and applications. In *ICATPN*, volume 77, pages 541–580. IEEE, April 1989.

[12] César Rodríguez and Stefan Schwoon. An improved construction of Petri net unfoldings. *Proc. of the French-Singaporean Workshop on Formal Methods and Applications (FSFMA'13)*, volume 31 of *OASICS*, pages 47–52. Leibniz-Zentrum für Informatik, july 2013.

[13] César Rodríguez, Marcelo Sousa, Subodh Sharma, and Daniel Kroening. Unfolding-based partial order reduction. *arXiv preprint arXiv :1507.00980*, 2015.

[14] Médésu Sogbohossou and David Delfieu. Dépliage des réseaux de Petri temporels à modèle sous-jacent non sauf. *ARIMA*, volume 14, pages 185–203, 2011.

[15] Antti Valmari. Stubborn sets for reduced state space generation. In *Proceedings of the 10th International Conference on Application and Theory of Petri Nets, 1989, Bonn, Germany ; Supplement*, pages 1–22, 1989.

[16] François Vernadat, Pierre Azéma, and François Michel. Covering step graph. In *ICATPN*, pages 516–535. Springer-Verlag, 1996.

[17] Pierre Wolper and Patrice Godefroid. Partial-order methods for temporal verification. In Eike Best, editor, *CONCUR*, volume 715 of *Lecture Notes in Computer Science*, pages 233–246. Springer, 1993.

A. Illustrations des deux différents algorithmes

Nous reprenons trois figures[2] dans la littérature (respectivement les figure 1 de [4], figure 3 de [6] et figure 4 de [1]) pour comparer la méthode standard [6] (notée ERV) qui est basée sur le concept d'ordre adéquat, avec notre algorithme proposé (noté DPP).

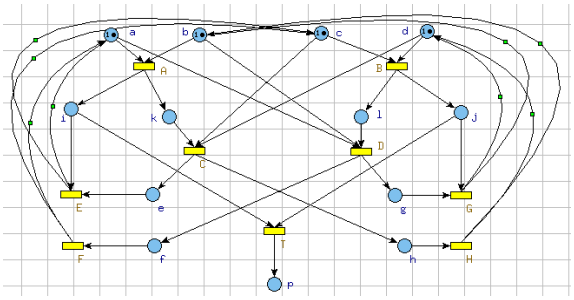


Figure 1. *Exemple 1 (fig. 1 de [4])*

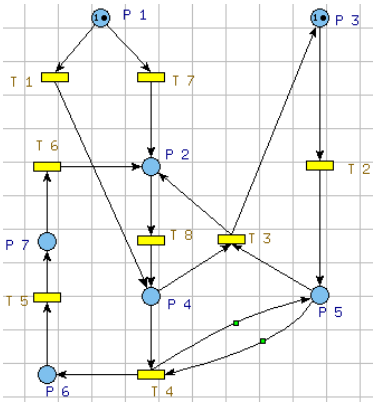


Figure 2. *Exemple 3 (fig. 4 de [1])*

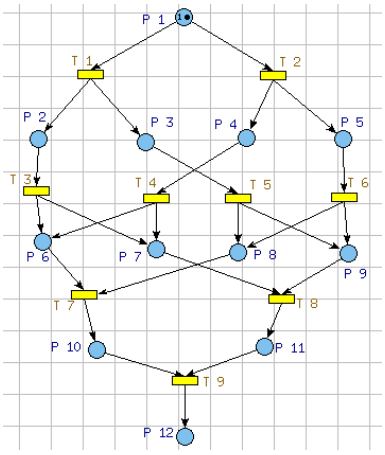


Figure 3. *Exemple 2 (fig. 3 de [6])*

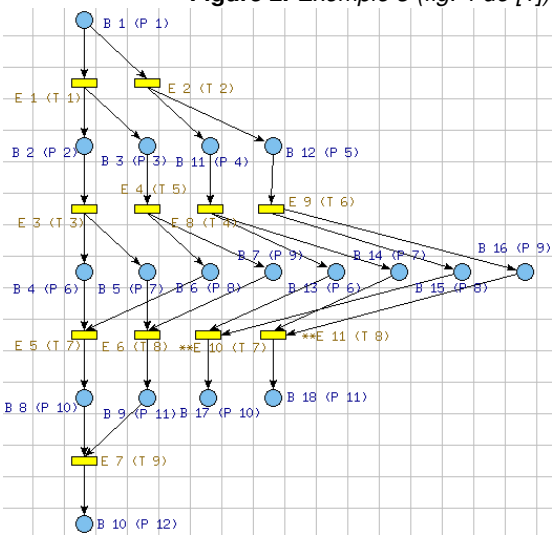


Figure 4. *Dépliage de l'exemple 2*

Il est montré que, selon les alternatives basées sur les stratégies *depth-first search* (DFS) [4] ou *breadth-first search* (BFS) [1, 6], sans associer un concept d'ordre adéquat (ou de relation bien fondée), le dépliage généré n'est pas toujours complet. C'est les cas par exemples de la figure 1 pour DFS et de la figure 3 pour BFS. Notre approche n'est complètement assimilable à aucune de ces deux stratégies.

Selon [6], une extension possible est choisie pour l'intégrer au dépliage s'il est mini-

2. Les réseaux sont édités avec le logiciel Romeo : `http://romeo.rts-software.org`

mal selon l'ordre adéquat $<$. La relation d'ordre $<$ est définie selon le critère de comparaison :

– de la taille de la configuration (locale), et en cas d'égalité,

– de l'ordre lexicographique des transitions associées du RdP N à la configuration locale, et en cas d'égalité,

– de l'ordre basé sur la forme normale de Foata[3] d'une configuration locale.

L'adéquation de la relation $<$ devra être préservée par toute extension en événement d'une configuration : en appliquant la forme normale de Foata, ceci n'est garanti que pour les réseaux saufs.

Le tableau 1 résume les résultats obtenus.

Tableau 1. *Résultats selon les deux algorithmes*

	ERV					DPP				
Petri nets	$\|E\|$	$\|B\|$	cut-offs	$\|BE\|$	$\|EB\|$	$\|E\|$	$\|B\|$	cut-offs	$\|BE\|$	$\|EB\|$
fig. 1	11	18	2	16	17	11	18	2	16	17
fig. 3	13	29	4	26	25	13	29	4	26	25
fig. 2	11	17	4	15	15	9	13	3	11	11

Pour illustration, le dépliage[4] de l'exemple 2 (fig. 4) est obtenu à partir de 2 processus : $\{e_1, e_3, e_4, e_5, e_6, e_7\}$ et $\{e_2, e_8, e_9, e_{10}, e_{11}\}$. Les événements cut-off sont e_{10} et e_{11}, en référence resp. à e_5 et e_6.

Bien que notre approche produise parfois un nombre d'événements différent de celui de [6], le dépliage produit est complet. Le principe utilisé pour tester la complétude sur le dépliage obtenu consiste à énumérer tous les marquages couverts et toutes les transitions entre ses marquages : à partir des labels des nœuds du dépliage, on traduit les éléments (marquages et transitions) de son graphe d'état en éléments du graphe d'état du RdP initial, puis on compare le résultat avec le graphe d'état obtenu directement à partir du RdP initial.

B. Présentation des résultats expérimentaux

B.1. Réseaux saufs

Les exemples de réseaux testés proviennent du logiciel de dépliage Mole[5]. Le tableau 2 permet de comparer les résultats de notre implémentation et ceux du dépliage classique.

Les résultats coïncident souvent, mais il existe des cas de dépliage pour lesquels ERV est plus favorable (gasnq3, over3-5) et d'autres pour lesquels notre algorithme est plus favorable (gasnq2, mmgt3, ring3-7).

3. Elle prend la forme d'une partition des événements de la configuration, partition obtenue en détachant itérativement l'ensemble des événements minimaux de la configuration locale [6].
4. Nos résultats (DPP) sont convertis en fichiers Romeo.
5. `http://www.lsv.ens-cachan.fr/~schwoon/tools/mole/`

Tableau 2. *Résultats du dépliage de réseaux saufs.*

	RdP				ERV			DPP		
	$\|T\|$	$\|P\|$	Marquages	Transitions	$\|E\|$	$\|B\|$	cut-offs	$\|E\|$	$\|B\|$	cut-offs
cyclic6	35	47	638	2176	50	112	7	50	112	7
cyclic9	53	71	7422	36608	77	172	10	77	172	10
cyclic12	71	95	77822	501760	104	232	13	104	232	13
dac6	34	42	640	2144	53	92	0	53	92	0
dac9	52	63	7424	35968	95	167	0	95	167	0
dac12	70	84	77824	493568	146	260	0	146	260	0
dme2	98	135	538	1036	122	487	4	122	487	4
dme3	147	202	6795	18312	321	1210	9	321	1210	9
dme4	196	269	76468	265868	652	2381	16	652	2381	16
dpfm2	5	7	4	5	5	12	2	5	12	5
dpfm5	41	27	12	31	31	67	20	31	67	20
gasnq2	85	71	192	373	169	338	46	165	330	46
gasnq3	223	143	1769	4587	1205	2409	401	1219	2437	401
gasq1	21	28	18	23	21	43	4	21	43	4
gasq2	97	78	180	357	173	346	54	173	346	54
mmgt1	58	50	72	144	58	118	20	58	118	20
mmgt2	114	86	816	2047	645	1280	260	641	1272	260
mmgt3	172	122	7702	22449	5841	11575	2529	5800	11493	2515
over2	32	33	64	133	41	83	10	41	83	10
over3	53	52	518	1563	187	369	53	286	566	83
over4	74	71	4174	16502	783	1536	237	1208	2378	410
over5	95	90	33506	163618	3697	7266	1232	5829	11490	2136
ring3	33	39	86	191	47	97	11	46	95	11
ring5	55	65	1289	4299	167	339	37	145	295	37
ring7	77	91	16999	75919	403	813	79	325	657	80

B.2. Réseaux non saufs

Nous considérons des réseaux non saufs qui modélisent un système multiprocesseur, puis un système producteur-consommateur : ils ont la particularité de contenir un grand nombre de conflits de transitions au cours de l'exécution.

Tableau 3. *Résultats du dépliage des réseaux non saufs.*

	RdP originel		DPP			
	Marquages	Transitions	$\|E\|$	$\|B\|$	cut-offs	Processus
Multiproc (5 proc, 2 bus)	45	107	91	172	67	49
Multiproc (10 proc, 5 bus)	250	845	1600	4036	1516	874
ProdCons (5 buf)	48	90	90	169	49	43
ProdCons (20 buf)	183	360	810	1429	649	613

Les résultats de dépliage (tableau 3) sont confirmés par le calcul des graphes d'état. On notera, en comparaison avec le nombre de transitions des graphes d'état, le plus grand nombre d'événements (dominés par les événements cut-off), et également le grand nombre de processus alternatifs nécessaires. Ceci est dû manifestement aux auto-conflits, i.e. le fait qu'un conflit concerne deux mêmes transitions du réseau originel avec des pré-conditions en concurrence : dans un graphe d'état, on ne considère jamais plusieurs instances de la même transition à partir d'un marquage. La solution évidente serait donc d'éliminer au cours du dépliage les auto-conflits qui sont des événements cut-off, afin que le nombre possible d'événements n'excède jamais le nombre de transitions du graphe d'état.

Les auteurs [6], dans la section 7.2 de leur article, comparent les résultats de la sémantique d'ordre partiel et de la sémantique d'exécution (i.e. par conversion préalable en réseau sauf) pour un certain nombre de réseaux non saufs. Notre implémentation donne des résultats toujours au moins aussi compacts que ceux prévus par la sémantique d'ordre partiel. Comparativement à la sémantique d'exécution, le résultat défavorable de la figure 9(c) de [6] pourra être corrigé en utilisant la solution préconisée au paragraphe précédent.

C. Détection de redondance et perspective de réduction du préfixe complet de dépliage

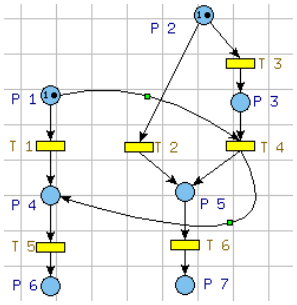


Figure 5. *Exemple 1 avec redondance*

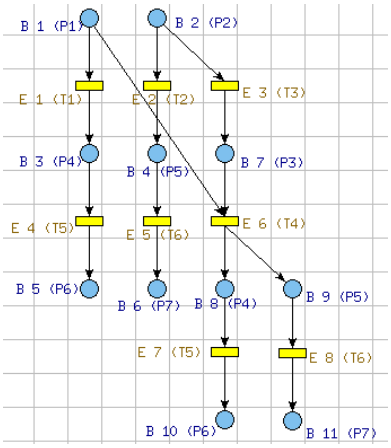


Figure 6. *Dépliage de l'exemple 1*

Le dépliage du RdP de la figure 5 est donné à la figure 6 (absence d'événement cut-off), soit la même structure que l'implémentation Mole (aux numéros des nœuds près). Il est occasionné par deux processus maximaux, $E_1 = \{e_1, e_2, e_4, e_5\}$ et $E_2 = \{e_3, e_6, e_7, e_8\}$, tombant sur le même marquage final $\{P_6, P_7\}$. Les deux processus ont ainsi un suffixe commun, les tirs des transitions T_5 et T_6, ce qui signifie des expressions de redondances dans le dépliage.

La question est de savoir si ses redondances peuvent être toujours supprimées, afin de réduire le préfixe complet. L'exemple nous suggère qu'en supprimant les descendants de b_8 et b_9 dans le processus E_2, la complétude est maintenue. Par contre, l'alternative de supprimer les descendants de b_3 et b_4 dans le processus E_1 ne conviendrait pas, parce qu'on perdrait la représentation des marquages intermédiaires P_2P_6 (b_2b_5) et P_1P_7 (b_1b_6), et des transitions qui en résultent. En somme, un préfixe réduit à $\{e_1, e_2, e_3, e_4, e_5, e_6\}$ est envisageable, mais la réduction à $\{e_1, e_2, e_5, e_6, e_7, e_8\}$ serait un dépliage incomplet.

De ce qui précède, pour qu'un des processus maximaux convergents vers le même marquage final soit réductible, l'événement (dit charnière) immédiatement en amont du suffixe commun doit être un point de convergence pour tous les événements plus antérieurs, et doit être une origine pour tous les événements constituants le suffixe commun : cela évite ainsi que des jetons (conditions) constituant un état intermédiaire soient dispersés sur le suffixe et le préfixe dont l'événement charnière constitue le point de ralliement.

Les travaux de Heljanko [7] ne traitent pas explicitement la question des redondances nécessaires, qui affectent la complétude du préfixe calculé.

L'extension de notre algorithme consistera à comparer l'état final du dernier processus maximal généré avec ceux des processus maximaux antérieurs, et en cas d'identité de marquage final, de voir si une réduction est possible sur l'un d'eux (par exemple, par une co-simulation inverse des deux processus) avant de poursuivre le dépliage.

User Interactions in Dynamic Processes

Modeling User Interactions in Dynamic Collaborative Processes using Active Workspaces

Nsaibirni Robert Fondze Jr* — Gäetan Texier**

* LIRIMA, University of Yaounde 1
PO Box 812, Yaounde, Cameroon
Centre Pasteur of Cameroon
nsairobby@gmail.org

** Centre d'épidémiologie et de santé publique des armées (CESPA)
UMR 912 - SESSTIM - INSERM/IRD/Aix-Marseille Université
gaetan.texier@univ-amu.fr

ABSTRACT. Flexibility and change at both design- and run-time are fast becoming the Rule rather than the Exception in Business Process Models. This is attributed to the continuous advances in domain knowledge, the increase in expert knowledge, and the diverse and heterogeneous nature of contextual variables. In such processes, several users with possibly heterogeneous profiles collaborate to achieve set goals on a processes mostly designed on-the-fly. A model for such processes should thus natively support human interactions. We show in this paper how the Active Workspaces model proposed by Badouel et al. for distributed collaborative systems supports these interactions.

RÉSUMÉ. La flexibilité et la changement pendant la conception et l'éxécution sont de plus en plus centrale dans les modèles des Business Process. Ceci est dû aux avancées continues des connaissances dans divers domaines, à l'augmentation des connaissances des experts, et de la nature hétérogènes et multiple des variables contextuelles. Dans ces processus, plusieurs utilisateurs ayant des profiles hétérogènes collaborent à des fins communs sur un processus défini progressivement. Un modèle pour de tels processus doit donc supporter nativement les interactions utilisateur. Nous montrons dans ce papier comment le modèle des Active Workspaces proposé par Badouel et al. pour la modélisation des tels processus support les interactions utilisateurs.

KEYWORDS : Collaborative Business Process, Human Interactions Patterns, Active Workspaces

MOTS-CLÉS : Processus Collaboratif, Interactions Utilisateurs, Active Workspaces

1. Introduction

Flexibility and change are fast becoming the Rule rather than the Exception in Business Process Models. As domain knowledge advances and expert knowledge increases, data and process definitions are prone to change. The need for dynamic process models is continuously being felt. Moreover, it is safe to say that dynamic process models increase user satisfaction and motivation at work, and positively influence productivity.

In [16] processes are classified as tightly-framed, loosely-framed, adhoc-framed, or unframed, depending on their predictable and repetitive nature, and on the degree of dynamism they require. The move from tightly-framed process models to unframed process models is characterized by the increasing facilities to manage uncertainty and exceptions, and the increasing influence of users and expert-knowledge in process design and enactment.

We focus on adhoc-framed and/or unframed domains, where users carry out processes in a fair degree of uncertainty[2][16] because processes cannot be completely modelled at design time either due to their large numbers or because they are highly data-centric and will have to be discovered as data is produced and as the environment evolves. In these domains, users (knowledge workers) are central to the different processes. They perform various interconnected knowledge intensive tasks and have to make complex rapid decisions on process models defined on-the-fly[2].

An example of such a domain is the disease surveillance process in public health. The process usually goes through a continuous cycle of collecting, analyzing, and dissipating information about a health condition of interest with the aim of detecting and handling unwanted events in the general population[8]. Disease surveillance is characterized as being multi-user, multi-organizational, knowledge-intensive, and time-bound[8][5]. Users and/or organizations need to collaborate and make complex rapid (timely) decisions on a semi-structured process model[2].

Like most organizational structures, a majority of national disease surveillance systems place users in a hierarchical pyramid[8]. In each level of the pyramid, users are grouped into Roles to carry out related work. Communication between the different levels of the pyramid and between the different Roles is usually through the asynchronous exchange of messages.

Our objective in this paper is to illustrate how user interactions (collaboration) in dynamic processes is supported by the Active Workspaces model[1]. We start by presenting key forms of human interactions found in business processes, then we present a purely distributed and informal specification of the Active-Workspaces model and show how it supports these interactions.

2. User Interaction Patterns

By user interaction, we mean any form of communication between a user and a computer or between two or more users via a computer[14]. Users interact in protean ways to have work done on a variety of task categories. Tasks are seen as work to be done and either originate from service calls or from work-(re)distribution in a team (work transfer and work delegation). In the following paragraphs, we describe the different ways users can interact. Our descriptions are inspired from the IBM's Business Spaces [14] that

define a human workflow attached to the underlying process model and on observations from concrete disease surveillance scenarios at the Centre Pasteur of Cameroon.

2.1. User interactions

In dynamic processes in general, users collaborate in the context of resolving specific cases. A case is a concrete instance of a business process[1]. For example, a case can comprise all tasks that will be invoked due to the arrival of a patient at a hospital or due to some outbreak alarm produced by some automated disease surveillance algorithm. One of the participating users initiates the case by instantiating the main task and providing the initially needed information. He then proceeds with the initial assignments and orchestration of tasks (work) to the other participants.

A simple description of a user's working environments could be: each of the participating users possess a work-basket which contains pending pieces of work that have been assigned to the user. In like manner, team-baskets are used to share work among a group of individuals. Task definitions contain information about the roles that have the ability to carry them out.

2.1.1. Work assignment or service request

Though users collaborate on processes in a peer-to-peer fashion, there is always a coordinating user who besides doing work is charged with initiating processes, assigning work to users, and coordinating the orchestration of the entire process. Such users exist throughout the entire process hierarchy, each managing the coordination of work that originates from him/her. Assigning work to some user (respectively to a group of users) consists in placing the work description in the user's work-basket (respectively in a group's work-basket).

2.1.2. Claiming/Releasing work

Users claim and carry out work placed in their work-baskets either based on the work-priorities or on the availability of the required input. A user can on the other hand release work placed on his/her work-basket when for some reason he is unable to carry it out.

2.1.3. Completing work

When a user claims work from his work-basket, he can either use an existing process definition to carry out the work or define a new process to do so. In both cases, he explicitly choses the method to use and provides the required input data. For certain routine tasks, he uses a rule-based approach to define a default method to always apply.

2.1.4. Handling situations

One of the following situations may arise: a user might want to rollback and change the method he applied to resolve a task, or a user might become overbooked or unavailable or unable to complete work due to the unavailability of some input data. Such situations are handled in one or more of the following ways: undo, redo, release work, transfer work, delegate work, re-prioritize work. These strategies are applied to take into account new constraints and/or facilitate and quicken decision making.

3. User Interactions in Active-Workspaces

Explicitly described in [1], the Active-Workspaces (AW) model uses attribute grammars to represent tasks and their decomposition into sub-tasks. Inherited attributes are

used to pass data from the parent to the sub-tasks while synthesized attributes are used to return results from subtasks to parent tasks. Attributes are terms over an ordered alphabet and task triggering and execution is guarded by conditions on the inherited attributes (using First Order Logic formulas and Pattern Matching). Hence the name Guarded Attribute Grammars (GAGs) given to the underlying grammar on which Active-Workspaces are built. In this section, we will show how the Active-Workspace model supports the major aspects of user interactions presented in the previous section.

3.1. Active-Workspace: User-roles, Users, and Services

The main building block in the Active Workspaces model is the user (identified by his Active-Workspace) and collaboration between users is materialized by the exchange of services. Each user can play several roles. Services are attached to Roles and users only offer services that are attached to the Roles they play. An Active-Workspace contains:

– Guarded Attribute Grammars: A (minimal) GAG is defined for each new service in the system and copied into the workspaces of the users that offer the service (that is, users that play the role to which the service is attached). The axiom of the GAG specifies the name of the service and the productions (Business Rules) describe how this service is decomposed into subtasks. A service definition contains a unique sort s (the axiom), input variables t_i (eventually with guards), and output variables y_i.

$$s(t_1, \ldots, t_n)\langle y_1, \ldots, y_m\rangle$$

– Artifacts: These are process execution trees corresponding to concrete cases (work carried out by a user in his workspace). They hold data and computations pertaining to cases from their inception to their completion. The tree contains two types of nodes: Closed nodes corresponding to resolved tasks or tasks for which a resolution method has been assigned, and Open nodes corresponding tasks that await to be assigned a resolution method. Visually, an artifact is a tree with sorted nodes $X :: s$, where s is the sort of node X.

– Input Buffer: A mail box in which any service requests made to a user as well as local variables whose values are produced in distant locations are placed. In practical situations, it is divided into two; a personal inbox (work-basket) and a role-inbox (team-basket). The former contains task requests made to the user directly and the latter contains tasks made to a role the user offers which he can pick-up and execute.

– Output Buffer: Contains information produced locally and used elsewhere in the system. This includes information about distant calls to services offered by the active-workspace and distant synthesized attributes whose values will have to be produced locally in the active-workspace.

A *task* is therefore simply a guarded attribute grammar production (Business Rule). It is identified by its name (*sort*), its inherited attributes eventually with guards, its synthesized attributes, and a decomposition into subtasks showing how synthesized attributes are produced from inherited attributes. BR1 below is an example of a Business Rule.

$$\begin{aligned}
&\text{BR1} :: \mathbf{caseAnalysis}(patient, symps, antecedents, checkRes, labResult) = \\
&\quad \mathbf{do}\,(todo, alarm, alert) \leftarrow \text{manageAlarm}(patient, symps, antecedents, \\
&\qquad\qquad labResult, checkRes) \\
&\qquad () \leftarrow \text{manageAlert}(alert, patient, symps, checkResult) \\
&\qquad \mathbf{return}(todo, alarm)
\end{aligned}$$

The above task **caseAnalysis**, extracted from the disease surveillance scenario for the monitoring of cases of Ebola[7] depicts what an Epidemiologist does when he receives a suspect case declaration (an Ebola outbreak alarm). This task receives as input information about the patient, the different checks carried out on him, and his laboratory results. It is decomposed into two subtasks manageAlarm and manageAlert, and returns two synthsized attributes *todo* and *alarm*. In like manner, we give an example of an Active-Workspace system description.

$$
\begin{array}{l}
\text{diseaseSurveillance} :: \langle \\
\qquad \text{consultPatient}[\mathit{clinician}], \\
\qquad \text{laboratoryAnalysis}[\mathit{biologist}], \\
\qquad \text{caseAnalysis}[\mathit{epidemiologist}] \\
\rangle \\
\textbf{where} \\
\mathit{clinician} \;=\; \text{Alice} \mid \text{Bob} \\
\mathit{epidemiologist} \;=\; \text{Ann} \mid \text{Paul} \\
\mathit{biologist} \;=\; \text{Frank} \mid \text{Mary} \mid \text{Alice} \\
\text{diseaseSurveillance} :: \quad \% \text{ Modelled system} \\
\text{consultPatient} :: \quad \% \text{ Service offered by clinicians} \\
\text{laboratoryAnalysis} :: \quad \% \text{ Service offered by biologists} \\
\text{caseAnalysis} :: \quad \% \text{Service offered by epidemiologists}
\end{array}
$$

Three services (consultPatient, laboratoryAnalysis, and caseAnalysis) are modeled in this system each offered by a distinct role (*clinician*, *biologist*, and *epidemiologist* respectively). A total of six (6) active workspaces will be generated corresponding to each of the users in the different roles. Parametric Business Rules are used in specifying Business Rules that are service calls. These simply tag the rules with the attached roles.

3.2. Requesting a service and Resolving a case

3.2.1. Requesting a service

As mentioned earlier, whatever the organizational structure, users communicate essentially by rendering and requesting services. Communication is enhanced in the Active Workspaces model using *variable subscriptions*. *Subscriptions* are equations of the form $x = u$ used to model variables x whose values u are produced at a distant site. Thus when a user calls a distant service, the synthesized attributes in the service call become subscriptions to values that will be returned by the call. Each variable has a unique defined occurrence in some workspace and may have several used occurrences elsewhere. This is enhanced using name generators that produce unique identifiers for newly created variables in each workspace.

More formally, let us consider two users: a local user identified by his active workspace AW_1 and a distant user identified by his active workspaces AW_2. When a service call is made from AW_1 to AW_2, the following takes place:

– $X = s(t_1, \ldots, t_n)\langle y_1, \ldots, y_m\rangle$ is added to the output buffer of AW_1 indicating the distant service call. This is distinguished from local calls in that there exist no defining rule for task s in AW_1.

– $Y = s(t_1, \ldots, t_n)\langle y_1, \ldots, y_m\rangle$ is added to the input buffer of AW_2, indicating that a distant service call has been made at node Y. This automatically creates a local node X and adds $Y = X$ to input buffer of AW_2 indicating where this service call is rooted in the the distant workspace.

– $x_i = u_i$ are added to the input buffer of AW_1, indicating that variables x_i in synthesized attributes y_i subscribe to the values of distant variables u_i. In like manner, $u_i = x_i$ are added to the output buffer of AW_2 indicated variable subscriptions it will have to fulfill. These subscriptions are fulfilled incrementally, that is, values are individually returned and sent to distant subscriptions as they are produced.

3.2.2. Task orchestration bus

Resolving a Case starts from an initialisation which consists in instantiating the root node of the main service with the axiom of the GAG. This creates an artifact with a single open node. The subsequent steps (micro steps) captured in the Active Workspaces model are sanctioned either by the application of business rules to open nodes or the consumption of a fulfilled subscription from its input buffer. Either way, executing a micro step adds data to the existing system and the only ordering on these steps is imposed by their data dependencies.

A business rule R is applicable at an open node X if its left hand side matches X and if any eventual logical expression on the variables in the inherited attributes evaluates to $TRUE$. This operation of pattern matching produces a substitution σ which is a redefinition of the variables in input positions in terms of variables in output positions of both the node X and the rule R. Several rules may match the open node and the choice of which to apply is made by the user. Once a rule is chosen, node X becomes closed and new open nodes $X_1, \ldots, X_n$ are created corresponding to subtasks on the right hand side of R. At the base, these open nodes are concurrently handled with an implicit ordering imposed by variable dependences. However, it is possible to add priorities, start- and due-time to tasks and hence to nodes and recommend a certain order in the execution of these tasks. These additions can be updated at any given moment to take into account new contextual realities. Open nodes for which no applicable rule is found correspond to services that have to be requested from a distant users.

Messages received at the input buffer also update the local configuration of the Active Workspace. These messages correspond either to the reception of a service call or to the fulfillment of a subscription. The former instantiates a root node for the corresponding service in the user's workspace while the latter recursively applies the effect of the subscription up the artifact tree.

3.2.2.1. Case Transfer, Delegation, and Synchronization

Case Delegation is naturally supported through service calls and is modeled in GAGs as terminal symbols and grammar axioms. A service is offered by a role and hence by users who play the role. A user cannot call a service he offers. In other words, users cannot call services attached to roles they play. Also, each service is designed to serve a particular role. That is, only users who play that particular role can call the service. Summarily, exchange of services only occur between roles and not within roles. However, users in the same role can communicate in two ways: *Case Transfer* and *Artifact Synchronization*.

In practice, *Case Transfer* is employed as a strategy to handle situations related to user unavailability and/or inability to complete work. To transfer a case, it suffices to transfer the initial service call to the new active-workspace and update the subscriptions accordingly. This creates a new artifact on which the distant user can start working.

Case Synchronization consists in weaving artifacts of the same service enacted in different workspaces. Practically, it can be used to share information between users working on the same case (for example after a case transfer). It can be either unidirectional (a

user shares his artifact with another user) or bidirectional two users synchronize artifacts in their workspaces. This feature considers artifacts as aspects and applies an operation reminiscent to the composition of aspects in aspect oriented programming.

3.2.2.2. Evolving the Active-Workspace

If we abstract the Active-Workspace model a level or two up, it becomes evident that this model has two major separate components: a dynamic underlying guarded attribute grammar specification, and an execution engine. New business rules, services, roles, and users added to the underlying grammar are automatically taken into consideration in subsequent executions of the system. This means that users can at any moment add, remove, or change the underlying grammar and these changes are directly visible (with no retrospective effect).

These two components form a single whole to provide users with the needed flexibility in designing, executing, and managing tasks in their active workspaces which by nature are perpetually evolving.

4. Discussion and Conclusion

Dynamic processes have been at the center of BPM research recently as per these reviews: [16] and [2]. Most of these research works have focused on flexible process design with users considered as part of the external environment[3][17][4][13][10]. A few other works show how exceptions and to some extent, uncertainty are managed in dynamic processes [12][9]. These works use a set of predefined exception handlers and again do not place users at a central position. The few researchers that have carried out work on user interactions have had to define an overlying user-workflow on a predefined process workflow[14][17]. These effectively enhance user interactions by adding flexibility to process enactment but lack flexibility in process design as the process has to be defined prior to its execution.

Active workspaces provide a holistic approach to dynamic process management with users, data, and processes being the essential building blocks. This model possesses to varying degrees the different forms of process flexibility presented in [16]. This explains why it naturally supports most forms of human collaboration in dynamic processes. We have used this model to show how such interactions can be supported. It is important to note that these operations might entail coupling the Active-Workspace model with external databases, knowledge bases, time servers, process performance monitors, etc. These certainly increase an overhead on the Active-Workspace model but have no negative effect on the specifications.

5. References

[1] Eric Badouel, Loic Helouet, Georges-edouard Kouamou, Christophe Morvan, and Robert Fondze Jr Nsaibirni. Active Workspaces : Distributed Collaborative Systems based on Guarded Attribute Grammars. *ACM SIGAPP Applied Computing Review*, 2015.

[2] Claudio Di Ciccio, Andrea Marrella, and Alessandro Russo. Knowledge-Intensive Processes: Characteristics, Requirements and Analysis of Contemporary Approaches. *Journal on Data Semantics*, pages 29–57, 2014.

[3] R. Hull, E. Damaggio, F. Fournier, M. Gupta, Fenno Terry Heath, S. Hobson, M. H Linehan, S. Maradugu, A. Nigam, P. Sukaviriya, and R. Vaculín. Introducing the Guard-Stage-Milestone Approach for Specifying Business Entity Lifecycles. In *Web Services and Formal Methods - 7th International Workshop, WS-FM 2010, Hoboken, NJ, USA*, volume 6551 of *Lecture Notes in Computer Science*, pages 1–24. Springer, 2011.

[4] Kunzle V, Reichert M PHILharmonicFlows: towards a framework for object-aware process management *Journal of Software Maintenance and Evolution: Research and Practice*,2011

[5] M.M. Wagner, L.S. Gresham, and V. Dato. Chapter 3 - case detection, outbreak detection, and outbreak characterization. In M.M. Wagner, A.W. Moore, and R.M. Aryel, editors, *Handbook of Biosurveillance*, pages 27 – 50. Academic Press, Burlington, 2006.

[6] International Society for Disease Surveillance. Final Recommendation: Core Processes and EHR Requirements for Public Health Syndromic Surveillance. Technical report, ISDS, 2011.

[7] R. Nsaibirni, G. Texier and GE. Kouamou. Modelling Disease Surveillance using Active Workspaces. *Conference de Recherche en Informatique (CRI), Yaounde*, 2015.

[8] Centers For Disease Control World Health Organization. Technical Guidelines for Intergrated Disease Surveillance and Response in the African Region. *Technical report, WHO/CDC, Georgia, USA* 2001.

[9] Andrea Marrella, Massimo Mecella, Sebastian Sardina SmartPM: An Adaptive Process Management System through Situation Calculus, IndiGolog, and Classical Planning *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourteenth International Conference, {KR}* 2014, Vienna, Austria, July 20-24, 2014

[10] Roger Atsa Etoundi, Marcel Fouda Ndjodo, and Ghislain Abessolo Aloo. A Formal Framework for Business Process Modeling. *International Journal of Computer Applications*, 13(6):27–32, 2011.

[11] Claudio Di Ciccio, Andrea Marrella, and Alessandro Russo. Knowledge-intensive Processes: An overview of contemporary approaches. *CEUR Workshop Proceedings*, 861:33–47, 2012.

[12] Reichert M, Rinderle S, Kreher U, Dadam P Adaptive Process Management with ADEPT2 *ICDE*, 2005

[13] ter Hofstede AHM, van der Aalst WMP, Adams M, Russell N Modern Business Process Automation: YAWL and its Support Environment. *Springer*, 2009

[14] Friess Michael Business spaces for human-centric BPM , Part 1: Introduction and concepts. *IBM DeveloperWorks* 2011.

[15] Roman Vaculín, Richard Hull, Terry Heath, Craig Cochran, Anil Nigam, and Piyawadee Sukaviriya. Declarative business artifact centric modeling of decision and knowledge intensive business processes. In *Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOC*, number Edoc, pages 151–160, 2011.

[16] Wil M. P. van der Aalst. Business Process Management: A Comprehensive Survey. *ISRN Software Engineering*, 2013:1–37, 2013.

[17] W. M. P. van der Aalst, M. Pesic, H. Schonenberg Declarative workflows: Balancing between flexibility and support *Computer Science - Research and Development*, 2009:99–113, 2009

A Distributed Pairwise Learning

On Distributing Bayesian Personalized Ranking from Implicit Feedback

Modou Gueye

LID
Université Cheikh Anta Diop
Dakar
Sénégal
modou2.gueye@ucad.edu.sn

ABSTRACT. Pairwise learning is a popular technique for collaborative ranking with implicit, positive only feedback. Bayesian Personalized Ranking (BPR) was recently proposed for this task and its ranking is among the bests. Because its learning is based on stochastic gradient descent (SGD) with uniformly drawn pairs, it converges slowly especially in the case of a very large pool of items.
We propose an approach to distribute its computation in order to face its scalability issue.

RÉSUMÉ. Le classement par pairs d'objets est une approche populaire d'apprentissage pour la recommandation d'objets à un individu. On se base sur l'hypothèse que ce dernier s'intéresse plus à un objet qu'il pris qu'un autre qu'il n'a pas considéré. De cette hypothèse, un classement des objets selon les intérêts qu'il porterait sur eux peut-être appris.
Nous proposons dans ce papier, une nouvelle approche permettant de paralléliser l'apprentissage du classement et donc de réduire considérablement le temps de calcul.

KEYWORDS : Distribution, Bayesian pairwise learning, Matrix factorization

MOTS-CLÉS : Distribution, Classement par pair, Factorisation de matrice

1. Introduction

Collaborative ranking with implicit, positive only feedback (so called one-class collaborative filtering) aims to make personalized ranking by providing a user with a ranked list of items [4]. In this kind of application, the collected data from user actions/behaviors are in an *one-class* form like what they purchased, clicked on or listened. Such data are referred as "implicit feedback" of users [2]. Contrary to the explicit ones in rating prediction where users rate items, and therefore we directly know the preference relationship between users and some items, we have here to infer user preferences from implicit feedbacks. That say to say, we have to only consider the presence or not of some users' actions (e.g., purchases, clicks, or even search events) in order to rank items for a given user when making recommendations to it.

For more formalization, let us consider an online shop and its users' history of purchases $S \subseteq U \times I$ with U the set of all its users and I the one of items to sell. The task of the recommender system is here to provide the user u with a personalized total ranking $\succ_u \subset I^2$ of all items, where $\succ_u$ has to meet the properties of a total order [9].

Collaborative ranking has been steadily receiving more attention, mostly due to the "one-class" characteristics of collected data in various services (e.g., "bought" in Amazon, "like" in Yahoo!Music, and "clic" in Google Advertisement). Bayesian Personalized Ranking (BPR) was recently proposed for this task. It is a matrix factorization technique which is able to learn individual ranking from implicit data. BPR is also admitted as one of the best current RS for item recommendation [9, 5, 8, 6, 3]. It takes pairs of items as basic units and maximize the likelihood of pairwise preferences over observed items and unobserved items. However BPR uses stochastic gradient descent and converges slowly especially if the pool of items is very large.

In this paper, we present a new approach to face the scalability of BPR by distributing its computation. Our proposal can be adapted to both shared-memory configuration or fully-distributed one. In the sequel, we first present the underlying ideas of BPR and its generic algorithm, then we detail our method to parallelize it. Finally we show that our proposal reduces almost proportionally the execution time according to the degree of distribution.

2. Bayesian Personalized Ranking

The key idea of BPR is to use partial order of items to train a recommendation model, contrary to previous works which just considered single user-item examples [2, 4]. BPR introduces the interpretation of positive-only data as partial ordering of items. When we observe that a user u has selected an item i (e.g., user u purchases item i in an online shop), we assume that this user prefers this item than the others without observed feedbacks. Thus from this assumption, one can infer partial order of items for the user.
Figure 1 shows an example of inference. On the left side, we have a matrix of observations collected from user actions from which user specific pairwise preferences $i \succ_u j$ between pairs of items can be inferred. On the right side, we present pairwise preferences deduced from user u_1. The symbol plus (+) indicates that he prefers an item than another, while minus (-) says the contrary. For items that he has both seen, we cannot infer any preference.

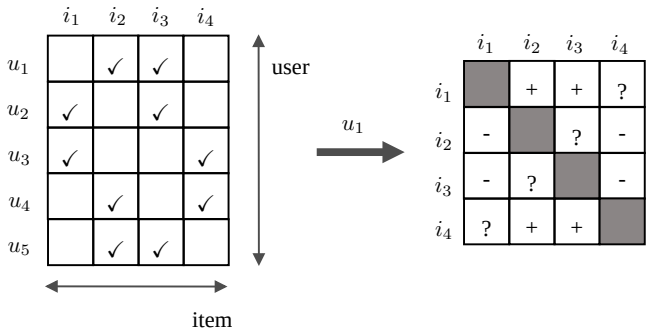


Figure 1 – Preferences retrieved from positive user-item occurrences

Let be $I_u^+ := \{i \in I : (u,i) \in S\}$ the set of implicitly-preferred items of user u. We can extract a pairwise preference dataset $\mathcal{P} : U \times I \times I$ by uniformly drawing for each user couples of an implicitly-preferred item and another one without observed feedback as follows

$$\mathcal{P} := \left\{(u,i,j)|i \in I_u^+ \wedge j \in I \backslash I_u^+\right\}$$

Each triplet $(u,i,j) \in \mathcal{P}$ implies that user u prefers item i than j. Due to the very large number of possible triplets, $\mathcal{P}$ is usually extracted by sampling techniques.

As BPR uses matrix factorization, it represents each user u (resp. each item i) by a vector p_u (resp. q_i) of latent factors. Thus, for each triplet $(u,i,j) \in \mathcal{P}$ we have the following order relation between the interests of u in i and j:

$$p_u \cdot q_i^T > p_u \cdot q_j^T, \;\; (u,i,j) \in \mathcal{P} \tag{1}$$

Hence, the main goal of BPR's optimization criterion ($BPR\text{-}OPT$) is to find an arbitrary model class to maximize the following posterior probability over all triplets in $\mathcal{P}$:

$$BPR\text{-}OPT = -\sum_{(u,i,j)\in\mathcal{P}} \ln \sigma(f_{uij}) + \lambda_\Theta \|\Theta\|^2 \tag{2}$$

For simplification, we posed f_{uij} as $p_u \cdot q_i^T - p_u \cdot q_j^T$. Θ represents the parameter vector of the arbitrary model class and λ_Θ the model specific regularization parameters. σ is the logistic sigmoid. The latter is used to approximate to non-differentiable Heaviside loss function [9]. Stochastic gradient descent (SGD) is used to learn the optimization criterion. In each step the gradient over the training data in $\mathcal{P}$ is computed and then the model parameters are updated with a learning rate α:

$$\Theta \leftarrow \Theta + \alpha \frac{\partial\, BPR\text{-}OPT}{\partial \Theta} \tag{3}$$

Algorithm 1 presents the learning of the optimization criterion with SGD.

Although BPR is among the best ranking technique, it converges slowly due to its sequential appraoch and pairs sampling, especially if the number of items is large [8]. Because that BPR relies on sampling pairs of items, its computation time grows relatively to the size of the pool of items to carry out.

In many large applications, we have to handle matrices with millions of both users and

Algorithm 1: Learning BPR

Data: $\mathcal{P}$, Θ
Result: Θ

Initialize Θ;
repeat
  Draw (u, i, j) from $\mathcal{P}$;
  $\Theta \leftarrow \Theta + \alpha \left((1 - \sigma(f_{uij})) \cdot \frac{\partial f_{uij}}{\partial \Theta} + \lambda_\Theta \Theta \right)$;
until *convergence*;
return Θ

items, and so many entries [1]. At such scales, distributed algorithms for matrix factorization are essential to achieve reasonable performance as discussed in [1]. This make BPR not suited for web-scale applications. We propose below a way to do it by generalizing the Distributed SGD (DSGD) of Gemulla et al. [1]. In our knowledge, there is not currently any proposition on this topic in the literature. Of course, one may think to used DSGD-liked approaches as in [1, 7, 10]. Thus it can partition the matrix of observations as illustrated in the left side of Figure 1 into independent blocks as in Figure 2. Independent blocks constitute a stratum (in gray color). Therefore parallel learning may be done on each block, stratum-by-stratum. Although this idea seems fine and was well applied to rating prediction, that is not the case for preference ranking. Indeed here we do not consider couples of user-item (i.e., the user and an item that he rated) but triplets (u, i, j) where the first item is more preferred by the user than the second. Thus using DSGD limits each computation node to take the items i and j from only its current block. Therefore the user-specific rankings that one will make may have partial, and block-limited order. Indeed, in the pairwise preference dataset $\mathcal{P}$, any triplet u, i, j where i and j are in different blocks can not be considered.

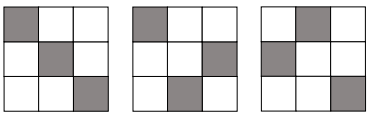


Figure 2 – Interchangeable blocks for a 3-by-3 gridded matrix

We propose a novel pair-blocks strategy which always keeps the notion of interchangeability of Gemulla et al. in [1]. They formulated it as two blocks which do not share any column nor row are interchangeable (i.e., independent). Thus two SGD instances can separately process them at the same time without any worrying. They define a stratum as a list of interchangeable blocks. The strata are processed in turn.

In the next section, we formalize and detail our proposal based on "pair-blocks" interchangeability. We show how we avoid partial, and block-limited order while ensuring distributed computing.

1. http://2016.recsyschallenge.com/

3. Distributed Bayesian Personalized Ranking

As we said above, block-based parallel gradient descent as introduced in [1, 7] is an original approach for distributing matrix factorization. Their well-minded concept of "interchangeability" underlies their contribution. We can define it as follows

Definition 1. *Blocks interchangeability*
Let be U_1 and U_2 to subsets of U, similarly I_1 and I_2 two subsets of I. Let be $\mathcal{B}_1 := U_1 \times I_1$ and $\mathcal{B}_2 := U_2 \times I_2$ two data blocks. They are interchangeable iff $U_1 \cap U_2 = \emptyset$ and $I_1 \cap I_2 = \emptyset$ (i.e., they do not share any row nor column).

From this definition, one can run operations completely in parallel on these blocks. Hence we introduce our notion of interchangeable pair-blocks, but for convenience, we define first our consideration of pair-block.

Definition 2. *Pair-block*
A pair-block φ is a couple of not interchangeable blocks $\mathcal{B}_1$ and $\mathcal{B}_2$ such as $U_1 = U_2$.

Definition 3. *Pair-blocks interchangeability*
Two pair-blocks are interchangeable if each block of the one is interchangeable with each block of the others.

Figure 3 shows two interchangeable pair-blocks represented with different colors. The matrix of observations can be expressed as unions of strata. Each stratum contains a group of interchangeable pair-blocks. In Figure 3, we list the sequence of strata that one have to set up when targeting two processors. As one can remark in this figure, in the two last

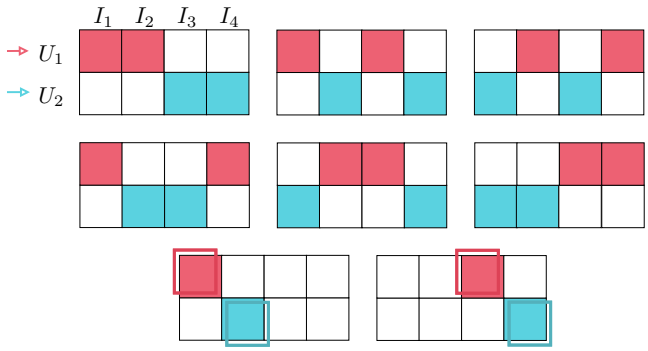


Figure 3 – Interchangeable pair-blocks-based strata

strata, the couple of blocks in pair-blocks has the same block. This allows us the possibility to consider any triplet $(u, i, j) \in \mathcal{P}$ wherever the position of i and j in the matrix of observation. As well as the number of block-columns is the double of the number of processors since each processor must have its own input. Therefore the sequence of strata must be carefully chosen in order to avoid re-using a pair-block in two different strata. Let be n the number of processors (e.g., two processors in Figure 3), each stratum must have its own n interchangeable pair-blocks to be processed in a distributed manner. From this point, it is easy to compute the number of strata to made since it becomes a combination problem. Indeed for n processors, the number N_s of strata to consider is the one of all 2-combinations of the $2n$ block-columns $C_{2n}^2 + 2n$ in order to cover all the triplets in $\mathcal{P}$

given by $N_s = C_{2n}^2 + 2n = n(2n+1)$.
One consequence of the use of pair-blocks is that we are able to join all any two blocks of the same block-row. Contrary to DSGD, we can infer preference ranking for each user over all the items. Therefore our ranking is not partial or block-limited while we are able to process each of our pair-blocks-based stratum in parallel. The processors of computation are synchronized when starting learning on a stratum.
We called our approach of distributing bayesian personnalized ranking by DBPR. Algorithm 2 details its functioning. Lines 7 to 10 are the distributed part. In Line 2 the strata

Algorithm 2: Learning DBPR

Data: $\mathcal{P}$, Θ, n
Result: Θ

1 Initialize Θ;
2 Generate strata $\mathcal{S}$;
3 // To balance workloads across the computing resources
4 Balance pair-blocks' data;
5 **repeat**
6 **foreach** $s \in \mathcal{S}$ **do** // We take the strata in turn
7 **for** $\varphi \in s$ **do in parallel** // Processing of pair-blocks φ
8 Draw (u, i, j) from $\mathcal{P}_\varphi$;
9 $\Theta \leftarrow \Theta + \alpha \left(\left(1 - \sigma(f_{uij}) \right) \cdot \frac{\partial f_{uij}}{\partial \Theta} + \lambda_\Theta \Theta \right)$;
10 **end**
11 **end**
12 **until** *convergence*;
13 **return** Θ

are generated and their pair-blocks' data balanced in Line 4.

4. Experimentation

We demonstrate in this section the efficiency of our proposal. Due to the limited paper size and the closeness of DBPR and BPR ranking qualities (see Section 7.2), we compare here their learning times. We led a set of experiments with two publicly available datasets.

4.1. Datasets

Due to the lack of implicit feedback datasets, researchers usually rely to transforming rating datasets [9, 5]. Thus we evaluate our algorithm using two different rating datasets: MovieLens@1M and MovieLens@10M [2]. As we want to solve an implicit feedback task, we first take only the ratings with a value ≥ 4 (the range of ratings is from 1 to 5), then we generated user-item pairs by removing the rating scores. Thus we obtain implicit, positive only feedback datasets. Table 1 shows the final characteristics of the datasets.

2. http://www.grouplens.org/node/73

Table 1 – Characteristics of the datasets

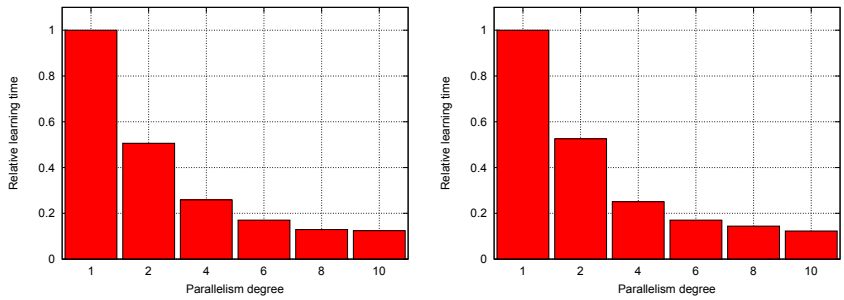
Dataset	$\|U\|$	$\|I\|$	$\|S\|$
MovieLens@1M	6,036	3,483	450,771
MovieLens@10M	56,071	10,119	4,010,795

4.2. Setup

We implemented DBPR and BPR in C/C++ and used shared-memory processing. We generated all strata by backtracking. Then to balance the amount of data in the pair-blocks, we used a round-robin-based approach which permutes both users and items. Two indexes allow us to find the final position of a user or an item.
Our evaluation consisted to run DBPR with increasing degree of parallelism, and compares its computation time to the one of BPR [3]. Of course, we included in the final processing time of DBPR the one spent to generate strata and balance data between the blocks. We ran our experiments on a linux computer (Intel/Xeon with 24 cores at 2.93 GHz, and 64 GB of memory).

4.3. Learning time vs Parallelism degree

On each dataset, we launched one instance of BPR, and successively instances of DBPR with increasing degrees of parallelization. To ensure considering the same number of triplets per iteration for both BPR and DBPR, with compute the number of triplets per both iteration and pair-blocks as follows $N_\varphi = \frac{N}{n \times N_s}$, where N represent the number of triplets per iteration for BPR. With this consideration, we ensure that all our executions do the same amount of calculation. For each dataset, we drew $10 \times |S|$ triplets at each learning iteration. The number of factors per user and item is fixed to 10 and the total number of iteration to 200.
Figure 4 points out the contribution of DBPR on learning time relatively to the one of BPR. The latter equals 494 and 6,053 seconds for respectively the learning time on MovieLens@1M and MovieLens@10M. We can observe that the learning time decreases almost



(a) MovieLens@1M (b) MovieLens@10M

Figure 4 – Relative learning time vs Parallelism degree

proportionally to the degree of parallelization thanks to the independence of pair-blocks in each stratum.

3. One can consider that an execution of BPR corresponds to the one of DBPR without parallelization

5. Conclusion

DBPR is a new proposal to improve the learning time of BPR-liked models. In our experimentation, we demonstrated its efficiency as it is able to nearly decrease the computation time proportionally to the degree of parallelization. Time reduction allows to learn BPR models from very large datasets by adapting our proposal to distributed framework like MapReduce.

Following the statement of the law of large numbers and the central limit theorem, one can expect a better ranking precision by increasing the size of the dataset $\mathcal{P}$ while ensuring moderated learning time with DBPR.

6. References

[1] Rainer Gemulla, Erik Nijkamp, Peter J. Haas, and Yannis Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 69–77, New York, NY, USA, 2011. ACM.

[2] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, pages 263–272, Washington, DC, USA, 2008. IEEE Computer Society.

[3] Lukas Lerche and Dietmar Jannach. Using graded implicit feedback for bayesian personalized ranking. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 353–356, New York, NY, USA, 2014. ACM.

[4] Rong Pan, Yunhong Zhou, Bin Cao, Nathan N. Liu, Rajan Lukose, Martin Scholz, and Qiang Yang. One-class collaborative filtering. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, pages 502–511, Washington, DC, USA, 2008. IEEE Computer Society.

[5] Weike Pan and Li Chen. Gbpr: Group preference based bayesian personalized ranking for one-class collaborative filtering. In Francesca Rossi, editor, *IJCAI*. IJCAI/AAAI, 2013.

[6] Shuang Qiu, Jian Cheng, Ting Yuan, Cong Leng, and Hanqing Lu. Item group based pairwise preference learning for personalized ranking. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 1219–1222, New York, NY, USA, 2014. ACM.

[7] Benjamin Recht and Christopher Recht. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

[8] Steffen Rendle and Christoph Freudenthaler. Improving pairwise learning for item recommendation from implicit feedback. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 273–282, New York, NY, USA, 2014. ACM.

[9] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 452–461, Arlington, Virginia, United States, 2009. AUAI Press.

[10] Yong Zhuang, Wei-Sheng Chin, Yu-Chin Juan, and Chih-Jen Lin. A fast parallel sgd for matrix factorization in shared memory systems. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 249–256, New York, NY, USA, 2013. ACM.

7. Annexes

7.1. Biographie

M. Gueye hold a PhD degree from Telecom ParisTech, a leading French engineering school specialized in computer science, under the supervision of Pr Talel Abdessalem (Telecom ParisTech) and Dr Hubert Naacke (University Pierre & Marie Curie, France). His thesis' subject was about designing scalable and accurate recommender systems.

M. Gueye is currently an Assistant Professor at University Cheikh Anta Diop (Sénégal). His research interests are in large scale data management and mining, recommender systems and web information extraction.

7.2. performance of DBPR in terms of ranking

Due to the limited size of the paper, we report here the performance of DBPR in terms of quality measures commonly employed in the recommendation field. For the performance evaluation, we used the Precision, Recall, F1 and NDCG measures which are references in this field.

Tables 2 and 3 show the ranking qualities of BPR and some instances of DBPR with increasing parallelism degree (2, 4 and 8 degree).

In almost all the measures, we see that the ranking quality of DBPR is close enough to the one of BPR. The slight lost of quality when the parallelism degree increases can be related to the pair-blocks-based learning of DBPR. Indeed, each processor unit is constrained to sample triplets into its current pair-blocks. Although this ensures independant processing, but we can not sample so much different triplets as BPR allows. We target to face this drawback in our future work. Indeed with the decreasing of computation time thanks to the distributed approach of DBPR, we can increase the number of triples to use in each iteration in order to expect better ranking in recommendations.

Table 2 – top@5 comparison of DBPR and BPR on MovieLens@1M

Algorithm	Recall	Precision	F1	NDCG
BPR	0.1057	0.3997	0.1671	0.0782
DBPR-2	0.1032	0.39	0.1632	0.0765
DBPR-4	0.0997	0.3886	0.1586	0.0732
DBPR-8	0.0972	0.3865	0.1553	0.0718

Table 3 – top@10 comparison of DBPR and BPR on MovieLens@1M

Algorithm	Recall	Precision	F1	NDCG
BPR	0.1766	0.3573	0.2363	0.1095
DBPR-2	0.1743	0.3502	0.2327	0.1061
DBPR-4	0.1726	0.3513	0.2314	0.1023
DBPR-8	0.1688	0.3481	0.2273	0.0994

Requêtes XPath avec préférences structurelles et évaluations à l'aide d'automates

Maurice TCHOUPÉ TCHENDJI *, Brice NGUEFACK *

* Département de Maths-Informatique
Faculté des Sciences, Université de Dschang
BP 67, Dschang-Cameroun
ttchoupe@yahoo.fr
brice_nguefack@yahoo.fr

RÉSUMÉ. Le concept de requêtes avec préférences a émergé dans la communauté des Bases de Données Relationnelles, pour permettre aux utilisateurs d'obtenir des réponses beaucoup plus pertinentes à leurs préoccupations, exprimées via des requêtes dites avec préférences. De telles requêtes ont généralement deux parties : la première permet d'exprimer les contraintes strictes et la seconde, des préférences ou souhaits. Toute réponse à une requête avec préférences doit nécessairement satisfaire la première partie et préférentiellement la seconde. Toutefois, s'il existe au moins une réponse satisfaisant la seconde partie, toutes les réponses ne satisfaisant que la première partie seront exclues du résultat final : elles sont dominées. Dans ce papier, nous explorons une approche d'importation de ce concept dans les Bases de Données XML via le langage XPath. Pour ce faire, nous proposons le langage PrefSXPath, une extension du langage XPath permettant d'exprimer les requêtes XPath avec préférences structurelles, puis, nous présentons un algorithme d'évaluation des requêtes PrefSXPath à l'aide des automates.

ABSTRACT. The concept of preferences queries emerged in the Relational Databases community, allowing users to get much more relevant responses to their concerns, expressed via requests say with preferences. Such requests usually have two parts: the first is used to express the strict constraints and the second, preferences or wishes. Any response to a query with preferences must necessarily satisfy the first part and preferably the latter. However, if there is at least a satisfactory answer of the second part, those satisfying only the first part will be excluded from the final result: they are dominated. In this paper, we explore an approach of importation of this concept in a XML Database via XPath language. To do this, we propose the PrefSXPath language, an extension of XPath for expressing XPath queries with structural preferences, then we present a query evaluation algorithm of PrefSXPath using automata.

MOTS-CLÉS : XML, XPath, Requêtes avec Préférences, Base de Données XML, Skyline, Automates.

KEYWORDS : XML, XPath, Preferences Queries, XML Database, Skyline, Automata.

1. Introduction

Les documents semi-structurés XML ont une structure flexible [1] et ne contiennent que du texte. Ils permettent alors d'échanger et de stocker les données plus aisément qu'avec les Bases de Données Relationnelles (BDR) classiques : on parle de Bases de Données XML (BDs XML).

L'exploitation des données stockées dans une BD se fait généralement via un langage dédié d'interrogation (SQL pour les BDR, XPath [15], XQuery [16] pour les BD XML, ...) permettant à l'utilisateur de rédiger des requêtes qui peuvent dans certains cas, contenir d'une part des exigences obligatoires appelées *contraintes*, et d'autre part, des exigences optionnelles appelées *préférences* ou *souhaits* : on parle alors de *requêtes bipolaires* ou de requêtes avec préférences[1][6, 9].

Des langages spécifiques (généralement des extensions du SQL ou du XPath) ont été proposés pour l'écriture des requêtes avec préférences : *SQLf* [3], *Preference SQL* [8], *Preference Queries* [5], ... pour les BDR, *XPref* [11], *Prefrence XPATH* [14], ... pour les BD XML. Les extensions de XPath prenant en compte les préférences ([11, 14]) procèdent généralement par réécriture de requêtes vers le format XPath pure. Bien plus, ils ne s'intéressent généralement qu'aux préférences portant sur les valeurs alors qu'un document XML intègre aussi une structure qu'il convient de prendre en compte lors de son interrogation.

Ce papier a pour objectif de proposer une autre approche d'importation du concept de requête avec préférences dans les BDs XML via le langage XPath en se focalisant sur la structure : nous traitons des *préférences structurelles*. A cet effet, nous proposons le langage *PrefSXPath*, une extension du langage XPath permettant d'exprimer les requêtes XPath avec préférences structurelles. Bien plus, nous présentons aussi une approche d'évaluation des requêtes *PrefSXPath* à l'aide des automates suivant la technique utilisée par Bin Sun et al. [2] pour l'évaluation des requêtes classiques (sans préférences). Notons que notre proposition se distingue de celles de [14, 11] sur bien des points : ils se sont intéressés principalement aux préférences portant sur les valeurs quand nous le faisons sur les structures, leur extension de XPath est localisée au niveau des prédicats alors que la notre porte également sur des nœuds sans prédicats ; là où ils font une réécriture de requête nous proposons un algorithme d'évaluation, ...

La figure 1 donne une vue synoptique de la démarche d'évaluation proposée ; elle se décline en deux étapes. En fait, étant donné une requête *PrefSXPath* Q à appliquer à un document XML D, après construction comme dans [2] d'un automate A relativement à la requête Q et extraction d'un index T_q de D, dans la première étape, A est utilisé pour générer toutes les réponses de Q dans D sans tenir compte des préférences. Parallèlement à la production des réponses de cette étape, des structures de données contenant des informations pour la sélection des meilleures réponses sont produites et utilisées dans la seconde étape pour construire une table relationnelle *preferenceTable*. Cette table indique de façon booléenne pour chaque réponse r et pour chaque préférence p contenue dans Q si une occurrence de p a été utilisée pour produire la réponse r. L'opérateur Skyline [12] est ensuite appliqué sur la table *preferenceTable* pour choisir les meilleures réponses : ce sont celles associées à des tuples non dominés.

Dans la suite de ce manuscrit, nous présentons dans la section 2 des notions relatives aux documents XML, et respectivement dans les sections 3 et 4 , une grammaire

1. c'est la seconde appellation que nous adopterons par la suite.

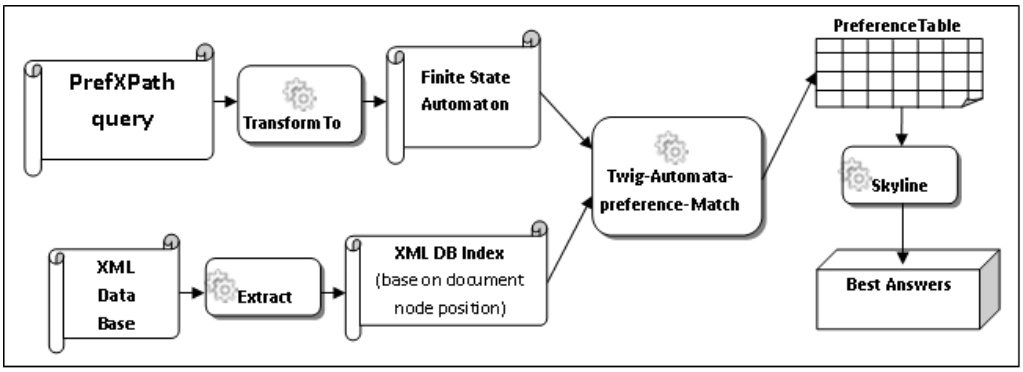


Figure 1. *une vue synoptique de notre approche d'évaluation de requêtes XPath avec préférences.*

pour le langage *PrefSXPath* et l'algorithme d'évaluation. La section 5 est consacrée à la conclusion ; nous y faisons un bilan de notre travail et présentons quelques pistes pour des travaux futurs.

2. Documents XML et interrogations

2.1. Documents XML : représentations et indexations

Un *document XML abstrait*[2] est représenté par un arbre étiqueté $D = (N_d, E_d)$ où N_d est un ensemble de nœuds étiquetés, E_d un ensemble d'arcs reliant chacun deux nœuds de N_d. Pour tout nœud $x \in N_d$ la fonction $label_d(x)$ retourne son étiquette.

Les documents XML sont généralement exploités à travers un index. Une présentation détaillée de ceux couramment utilisés est donnée dans [7, 10]. Dans ce papier nous utiliserons l'indexation basée sur la position des nœuds (aussi appelée la notation positionnelle) [10] dans laquelle chaque nœud du document est représenté par un triplet *(Start, End, Level)* : *Start* et *End* représentent respectivement les positions de début et de fin de l'élément dans le document ; *Level* est la profondeur du nœud dans la représentation arborescente du document. Avec cette convention, comme dans [17], l'index d'un document est constitué d'un ensemble de listes chainées T_a d'occurrences de nœuds de type a triées suivant la composante *Start* du triplet. L'un des avantages de cette représentation est qu'elle permet, étant donné deux nœuds a et b quelconques représentés respectivement par les triplets *(starta, enda, levela)* et *(startb, endb, levelb)*, de déterminer les relations *parent-enfant* ou *ancêtre-descendant* en temps constant. En effet, *a est un ancêtre de b* si et seulement si *starta<startb<enda*. Si de plus *levela +1=levelb*, alors, *a* est le *parent* de *b*. Par la suite, nous dirons d'un nœud *a* qu'il *recouvre* un nœud *b* si *b* est un descendant de *a*.

2.2. Documents XML : requêtes et évaluations

A l'instar d'une BD classique, un document XML contient des informations (des données). Il encapsule aussi une structure dont il faut impérativement tenir compte lors de son interrogation. Ainsi une requête XML concerne non seulement le contenu (les données)

2. Dans un document abstrait, on fait abstraction des nœuds de textes et des nœuds d'attributs : ceux-ci ne sont pas d'un intérêt avéré pour les traitements purement structurels qui nous intéressent dans ce papier.

mais aussi la structure (les relations structurelles que l'utilisateur souhaite avoir entre les différentes occurrences des éléments). Comme pour la représentation des documents, une requête XML Q peut être représentée par un arbre $Q = (N_q,\ E_q)$ dans lequel N_q est un ensemble des nœuds étiquetés, E_q un ensemble d'arcs reliant chacun deux nœuds de N_q. Dans E_q on distingue deux types d'arcs : ceux reliant un nœud père à un nœud fils noté x/y et ceux reliant un nœud à un de ses descendants noté $x//y$. Pour tout arc $x \in N_q$ la fonction $label_q(x)$ retourne son étiquette, et on note $\Sigma = \{label_q(x),\ x \in N_q\}$.

Soient, $Q = (N_q,\ E_q)$ une requête, $D = (N_d,\ E_d)$ un document XML, deux nœuds $n_d \in N_d$ et $n_q \in N_q$: On dira que n_d *est une occurrence de* n_q dans D si $label_q(n_q) = label_d(n_d)$. On dira aussi que la requête Q est *satisfaite* dans D si chacun des nœud de N_q possède au moins une occurrence dans N_d et telle que les relations de parenté existant entre les nœuds de N_q soient les mêmes qui existent entre leurs occurrences.

2.3. Requêtes XPath et évaluations

XPath (XML Path) [15] est un langage essentiel[3] de requête XML permettant de sélectionner des nœuds d'un document XML de sorte que le chemin allant de la racine du document à chacun des nœuds sélectionnés satisfait un motif (la requête) donné. En plus des relations parent-enfant et ancêtre-descendant, on trouve aussi dans les requêtes XPath des prédicats associés à des nœuds. Ce sont des expressions booléennes comprises entre les symboles '[' et ']' et accolées à un nœud requête pour signifier que les occurrences licites de ce nœud sont celles pour lesquels le prédicat est évalué à *True*. De nombreuses techniques d'évaluation de requêtes XPath existent [13, 4], parmi lesquelles celle développée par Bin Sun et al. [2] sur laquelle nous nous appuyons pour étayer notre approche d'extension du langage XPath à l'aide des préférences. Dans l'approche de [2] une requête XPath est transformée en un automate en appliquant un ensemble de schémas de construction associés aux différents types de chemin existant dans le langage XPath. Certaines transitions de l'automate produit sont décorées par des *actions* dont l'exécution dans l'algorithme de matching, lors du franchissement de telles transitions permet de construire progressivement la solution de la requête.

3. PrefSXPath : un langage pour l'expression des requêtes XPath avec préférences

Dans cette section, nous présentons la grammaire de *PrefSXPath* ainsi que des éléments du vocabulaire utilisé dans la suite de ce manuscrit.

3.1. Une notation pour les préférences dans la grammaire de XPath

Nous introduisons la notation ' !' comme opérateur unaire dans la grammaire de *PrefSXPath* (voir ligne 7 du listing ci-dessous) pour l'expression des préférences dans un (sous-)chemin XPath. Par exemple, dans la requête *PrefSXPath* $Q_1 = /a/(b[c])!/d$ les sous-chemins $/a$ et $/d$ représentent des *contraintes*, alors que $/(b[c])!$ représente une *préférence*. Q_1 est interprétée comme une requête retournant toutes les occurrences *di* du nœud résultat *d* satisfaisant : (1) Le chemin allant de la racine à *di* possède (obligatoirement) une occurrence de *a* et éventuellement une occurrence de *b* si celle-ci est parente d'une occurrence de *c*. Les occurrences *di* candidates à faire partie de la réponse sont

3. Il est utilisé par bien d'autres langages d'interrogation XML comme XQuery [16], XSLT, ...

donc celles pour lesquelles il existe dans le document un sous-arbre de l'une des formes */ai/bi[ci]/di* ou */ai/di*. (2) S'il existe même une seule réponse candidate, disons *dp*, telle que le chemin allant de la racine à *dp* contienne une occurrence de *b* (*/ap/bp/dp)*), alors, seules les réponses candidates intégrant une occurrence de *b* seront retournées comme réponses à la requête : on dit qu'elles *dominent* les autres solutions. Ainsi présenté, la notion de préférence permet de ne retenir comme résultat à la requête que les *meilleures réponses* c-à-d. les plus préférées[4].

En nous basant sur la grammaire donnée dans [2] pour décrire un sous ensemble significatif du langage XPath, nous considérons la grammaire suivante pour les expression *PrefSXPath*.

Syntaxe BNF des expressions *PrefSXPath*

```
AbstTwigExpr ::= '/' RltvTwigExpr
                 | '(' RltvTwigExpr ')'
RltvTwigExpr ::= RltvTwigExpr BinaryOp RltvTwigExpr
                 | RltvTwigExpr UnaryOp
                 | Step
BinaryOp ::= '/' | '//'
UnaryOp ::= '*' | '+' | '?'|!
Step ::= Name
         | Name '[' Predicate ']'
Predicate ::= RltvTwigExpr
```

3.2. Quelques définitions et notations

Considérons la requête avec préférence $Q_3 = /a/(b[(c[d/e!])!/i])!/a[d]/g!/h$ ayant *h* pour *nœud résultat*. En remarquant que plusieurs nœuds d'une expression *PrefSXPath* peuvent avoir le même label, pour pouvoir désigner sans ambigüité tout nœud par son label, on peut lui ajouter un indice relativement à sa position dans l'expression. Q_3 par exemple est reécrit en $Q_3 = /a_1/(b[(c[d_1/e!])!/i])!/a_2[d_2]/g!/h$.

Dans Q_3, aux nœuds *b*, *c* et a_2 sont associés des prédicats donnés respectivement par les *chemins prédicats* $[(c[d_1/e!])!/i]$, $[d_1/e!]$ et $[d_2]$; de tels nœuds sont appelés *nœuds clés*[5] . Le chemin obtenu en supprimant dans la requête tous les chemins prédicats est appelé *chemin principal* : Q3 a pour chemin principal $/a_1/b!/a_2/g!/h$. Le sous-chemin $(b[(c[d_1/e!])!/i])!$ permettant d'exprimer une préférence est appelé *chemin préférence* et le nœud clé *b* de ce chemin est appelé *nœud préférence* de ce chemin. En fait, on appelle *nœud(s) préférence(s)* d'un chemin préférence le(s) nœud(s) résultat(s) de ce chemin[6].

Quand un nœud préférence figure dans un chemin prédicat, il est lié au chemin principal par un unique nœud clé qui est son pivot. Le *pivot*[7] d'un nœud préférence est en fait,

4. Notons que le concept de *préférence* est différent de celui de *optionnelle* déjà présente dans le langage XPath et symbolisé par '*?*'. Par exemple, avec la requête $Q_2 = /a/(b[c])?/d$ dans laquelle le sous chemin $/(b[c])?$ est optionnel, les occurrences di de d qui seront retournées sont celles appartenant à un sous arbre du document de l'une des deux formes $/ai/bi[ci]/di$ ou $/ai/di$.

5. Lors de l'exécution de la requête, quand on arrive sur un nœud clé, on doit suivre deux chemins parallèlement : celui du prédicat et celui du *chemin principal*.

6. Par exemple, les requêtes suivantes *a!, (a[b/c])!, (a /b([c/d])!/e)!, (a/(b[c/d])!)!* ont pour nœuds préférences respectivement {a}, {a}, {e}, {a, b}.

7. En considérant la notation positionnelle, si $np(sn,en,ln)$ est un nœud préférence et $nc1(sc1,ec1,lc1),\ldots,nck(sck,eck,lck)$ sont k nœuds clés ancêtres de np (on a dans ce cas,

le nœud clé situé sur le chemin principal tel que le chemin qui les relie ne passe pas par un autre nœud clé situé sur le chemin principal. Notons que si un nœud préférence est situé sur le chemin principal, il est sont propre pivot. Dans Q_3, les nœuds préférences *b*, *c*, *e* et *g* ont respectivement pour pivots les nœuds *b*, *b*, *b* et *g*.

L'approche d'évaluation utilisée ici est descendante : on évalue une requête en commençant par l'évaluation du nœud le plus proche de la racine et en évoluant vers les feuilles. Nous appellerons donc par la suite *requête partielle* tout (chemin) préfixe d'une requête : $/a/(b[(c[d/e!])!/i])!$ est une requête partielle de Q_3.

4. Evaluation des requêtes PrefSXPath à l'aide des automates

4.1. De l'expression PrefSXPath à l'automate

Le tableau 1 donne à l'exemple de celui présenté dans [2] pour chaque motif de requête *PrefSXPath*, le schéma de l'automate à construire correspondant. Les transitions de ces automates sont étiquetées soit par des éléments de Σ, soit par λ quand la requête permet d'exprimer une relation *ancêtre-descendant* ($x//y$), soit par le nom d'une des actions décrites à la section 4.2.2.

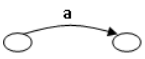
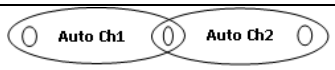
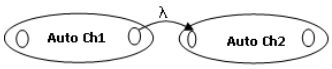
Chemin *PrefSXPath*	**Automate correspondant**
$a \in \Sigma,\ /a$	
$Ch1/Ch2$	
$Ch1//Ch2$	
$Ch?$	
$Ch!$, a étant un nœud préférence de Ch	
$c \in \Sigma,\ c[Ch]$	
Ch^*	

Tableau 1. *Motifs de requêtes PrefSXPath et schémas d'automates correspondant*

$sci < sn \leq en < eci,\ \forall i, 0 < i < k$) et situés sur le chemin principal, alors le nœud pivot de np est le nœud ncl (scl, ecl, lcl), $0 < l < k$ tel que ($ecl == minimum(ec1, \ldots, eck)$) : c'est le nœud y ayant la plus petite valeur de *endPos*. On peut donc statiquement calculer le nœud pivot de tout nœud préférence. On supposera donc dans la suite qu'on dispose d'une fonction *getPivotId* qui retourne le pivot d'un nœud préférence donné. Par exemple, *getPivotId (c) = b*.

4.2. Algorithme d'évaluation

4.2.1. Les structures de données

Rappel : l'évaluation d'une requête *PrefSXPath* est descendante. Au cours de l'évaluation, le résultat de l'évaluation du nœud requête courant est stocké dans une variable appelée *currentAnswer*. Ce qui suit est une description des autres variables utilisées par l'algorithme.

partialSolutionStack : pile utilisée lors du traitement des (sous-)chemins optionnels et préférences. En effet, lors du franchissement d'un arc étiqueté par l'une des actions *BeginOP* ou *BeginPref* ou encore *Push*[8], la solution partielle courante (contenue dans *currentAnswer*) y est stockée en vue de son utilisation à la fin du traitement du sous-chemin optionnel ou préférence (traversé d'un arc étiqueté EndOp ou EndPref) pour construire la nouvelle solution courante.

prefNodeStack : pile utilisée pour stocker et restituer les nœuds préférences contenus dans le chemin préférence courant : lors du franchissement d'un arc étiqueté *BeginPref(a)*, le nœud préférence *a* y est empilé et lors du franchissement d'un arc étiqueté *EndPref*, on y effectue un dépilement.

answerPivotTable : tableau associatif dont chaque entrée contient une paire *(clé, valeur)*. *clé* est un identifiant d'un nœud pivot, et la valeur associée est le résultat de l'évaluation de la sous-requête ayant ce nœud comme nœud résultat ; il est initialisé à *currentAnswer* au moment du franchissement d'un arc étiqueté *clé*. Lors du franchissement d'un arc étiqueté *EndPref* l'entrée associée à la clé du pivot du nœud préférence courant (c'est celui se trouvant actuellement au sommet de la pile *prefNodeStack*) est utilisée pour filtrer les occurrences du pivot à stocker dans l'entrée de la table *infoPrefNodeTable* correspondant au nœud préférence courant.

infoPrefNodeTable : tableau associatif dont chaque entrée contient une paire *(clé, valeur)*. clé est un identifiant d'un nœud préférence, et la valeur associée est constituée des occurrences du pivot de ce nœud recouvrant au moins une de ses occurrences.

preferenceTable : tableau de booléens renseignant pour chaque réponse *r* de la requête et pour chaque nœud préférence *p* si une occurrence de *p* a été utilisée pour la production de la solution *r*.

4.2.2. Les actions

Le tableau 2 associe à chaque action la description des traitements à effectuer lors du franchissement d'un arc étiqueté par cette action.

4.2.3. L'algorithme

Etape1 : évaluation sans tenir compte des préférences

L'algorithme de la première étape d'évaluation d'une requête *PrefSXPath Q* sur un document *D* (algo. 1) prend en entrée un index Tq de *D* et l'automate *A* associé à *Q*. Il retourne le triplet *(currentAnswer, infoPrefNodeTable, answerPivotTable)* utilisé dans la seconde étape pour construire la table *preferenceTable* de laquelle sont extraite les meilleures réponses. Dans l'algorithme 1, on parcourt l'automate *A* à partir de l'état initial en exécutant suivant le type de l'étiquette de la transition qui part de l'état courant le traitement correspondant : si c'est une valeur $a \in \Sigma$, (on est sur un chemin de type $ch1/a/ch2$), alors, on filtre dans la liste Ta les occurrences de *a* qui sont fils d'un élément de la solution courante *currentAnswer* (algo. 1, ligne 15) et si *a* est un pivot, on remplit l'entrée correspondante dans la table *answerPivotTable* (algo. 1, lignes 16-18) ;

8. Voir la description des actions à la section 4.2.2.

Nom action	**Description des traitements associés**
Push	Empile la solution partielle courante dans la pile *partialSolutionStack*
FilterUp	Filtre dans le résultat se trouvant au sommet de la pile *partialSolutionStack* ceux qui sont ancêtres d'une occurrence se trouvant dans *currentAnswer*.
PopRestore	Dépile le résultat au sommet de la pile *partialSolutionStack* et en fait la solution courante.
BeginPref (a)	Annonce le début de traitement d'un chemin préférence ayant *a* pour nœud préférence : on doit empiler la solution (partielle) courante dans *partialSolutionStack* ainsi que le nœud préférence *a* du chemin préférence qu'on s'en va traiter dans la pile *prefNodeStack*.
BeginOp	Annonce le début de traitement d'un chemin optionnel : on doit empiler la solution (partielle) courante dans *partialSolutionStack* :
EndOp	Annonce la fin de traitement d'un chemin optionnel. On doit mettre à jour la solution courante (*currentAnswer*) en y ajoutant les occurrences du résultat présentement au sommet de la pile *partialSolutionStack*.
EndPref	Annonce la fin du traitement d'un chemin préférence. On doit tout d'abord restaurer le nœud préférence courant (c'est celui au sommet de la pile *prefNodeStack*) puis, filtrer l'entrée correspondant à son pivot dans la table *answerPivotTable* avec la solution courante *currentAnswer* pour n'y retenir que les occurrences du pivot recouvrant au moins une occurrence du nœud préférence courant contenue dans *currentAnswer*. Le résultat du filtrage est stocké dans la table *infoPrefNodeTable* à l'entrée correspondant au nœud préférence courant. Pour finir, on doit effectuer tous les traitements énumérés ci-dessus pour l'action *EndOp*.
BeginStar	Annonce le début de traitement d'un chemin étoilé (chemin avec star (*)) du type $ch1/(ch)*/$. On doit ajouter à la solution partielle résultant du traitement de la sous requête *ch*1, les solutions provenant du traitement de zéro ou plusieurs occurrences de *ch*. La solution partielle V de $ch1/(ch)*/$ sera la somme $V = V_0 + V_1 + V_2 \ldots + V_n$, où V_i est la solution partielle résultant de la consommation de *ch* pour la i^{ime} fois. Les V_i ne sont pas indépendants : $V_{(i+1)}$ dépend de V_i dans la mesure où, les éléments de la solution partielle $V_{(i+1)}$ sont des descendants des éléments de V_i qui correspondent à la chaine ch : $V_{(i+1)} = \{v_i | \exists v_k \in V_i, v_i Desc_{ch}(v_k)\}$. $v_i Desc_{ch}(v_k)$ signifie que v_i est un descendant de v_k suivant la chaine *ch*.
EndStar	Annonce la fin de la i^{ieme} consommation d'un chemin étoilé. Si lors de ce i^{ieme} passage on a récolté une solution, c.-à-d. si (V != ϕ), on doit l'ajouter à la solution partielle courante et se reconnecter au début du chemin étoilé. Sinon, on sort de ce chemin pour poursuivre le traitement de la requête résiduelle.

Tableau 2. *Actions et traitements associés*

sinon si c'est λ, (on est sur un chemin de type $a//b$) alors, on filtre dans la liste Tb les occurrences de *b* qui sont descendants d'un élément de la solution courante *currentAnswer* (algo 1, lignes 6-8). Enfin, si c'est une action *(BeginStar, BeginPref, EndPref, ...)*, les instructions prévues pour le traitement de celle-ci (tableau 2) sont exécutées[9].

entrée: - L'index T_q du document ;
- L'automate *A* associé à la requête;
sortie : Toutes les occurrences du nœud résultats de la requête satisfaisant la requête

```
1  currentState = q0 ; answerPivotTable = ϕ ; infoPrefNodeTable = ϕ;
2  currentAnswer ='/' ;                /* On initialise à la racine */;
3  tantque currentState != FA faire
4      a=labelTrans currentState        /* On récupère le label de la transition
       courante  */;
5      si a= = λ alors                                              /* ch//b */
6          currentState = δA(currentState, a);
7          b= labelTrans currentState;
8          (currentAnswer, Tb) = filter ANC-DESC currentAnswer Tb;
9          currentState = δA(currentState, b);
10     sinon
11         si a ∈ Σact alors
12             (currentAnswer, qc) = perform_Action (currentAnswer, q, a);
13             currentState = qc;
14         sinon                                                    /* a ∈ Σ */
15             (currentAnswer, Ta) = filter CHILD currentAnswer Ta;
16             si (isPivot(a)) alors
17                 indiceA= indice(a, answerPivotTable);
18                 answerPivotTable[indiceA] = currentAnswer ;
19                 currentState = δA(currentState, a);
20 return (currentAnswer, answerPivotTable, infoPrefNodeTable ) ;
```

Algorithm 1: Twig-Automata-Preference-Match pour l'étape 1

Etape 2 : extraction des meilleures réponses

La seconde étape est divisée en deux phases : dans la phase 1 on construit la table *preferenceTable* et dans la phase 2 on sélectionne les meilleures réponses en appliquant *l'opérateur skyline*[10][12] sur les tuples de la table *preferenceTable* pour ne retenir que les réponses contenues dans l'ensemble constitué des tuples non dominés.

Dans ce qui suit, pour des besoins d'illustrations, nous considérons la requête *PrefSX-Path* $Q = /a_1/\ldots/a_{(k-2)}/b![\ldots]/a_k/\ldots/a_{(k+l)}/c[d_1/\ldots/d_{k_{pred}}/g![\ldots]/\ldots]/a_{(k+l+2)}/\ldots/a_s/f$ ayant f comme nœud résultat, possédant deux nœuds préférences *b* et *g* situés respective-

9. La fonction *perform_Action(action, q, currentAnswer,)* ; (algo. 1 ligne 12) permet d'effectuer les traitements liés à l'action *action* étiquetant l'arc associé à l'état *q*, les réponses courantes étant dans *currentAnswer*.

10. L'opérateur skyline [12] permet de sélectionner les meilleurs n-uplets c.-à-d. ceux qui ne sont pas dominés au sens de la relation de préférence. De façon sommaire, il peut être présentée comme suit : soient deux tuples $p = (p_1, \ldots, p_k, p_{k+1}, \ldots p_n)$ et $q = (q_1, \ldots, q_k, q_{k+1}, \ldots q_n)$ d'une table relationnelle R de schéma $R(P_1, \ldots P_K, P_{K+1}, \ldots P_n)$. Pour les requêtes dans lesquelles les préférences portent sur les champs $P_{k+1}, \ldots, P_n$, on dira que ***p domine q*** et on note $p > q$, si les trois conditions suivantes sont satisfaites : (1) $p_i = q_i$, pour tout $i = 1, 2, \ldots k$. (2) $p_i \geq q_i$ pour tout $i = (k+1), \ldots, n$. (3) il existe i, $(k+1) \leq i \leq n$, et $p_i > q_i$.

ment sur le chemin principal et sur un chemin prédicat. Considérons aussi que :
• Q a pour chemin principal $/a_1/\ldots/a_{(k-2)}/b!/a_k/\ldots/a_{(k+l)}/c/a_{(k+l+2)}/\ldots/a_s/f$ de longueur s+1.
• $/a_k/\ldots/a_{(k+l)}/c/a_{(k+l+2)}/\ldots/a_s/f$ est le suffixe du chemin principal comprenant $K = s+1-(k-1) = s-k+2$ nœuds, parmi lesquels on suppose avoir $p,\ 0 \leq p \leq K-1$ nœuds préférences.
• Le (sous-)chemin $d_1/\ldots/d_{(k_{pred})}$ contenu dans le prédicat associé au nœud c est le préfixe d'un chemin principal se terminant juste avant le nœud g et contenant k_{pred} nœuds[11], parmi lesquels on suppose qu'on ait $p_{pred},\ 0 \leq p_{pred} \leq k_{pred}$ autres nœuds préférences.
• $a_{(k+l+2)}/\ldots/a_s/f$ sont les $k_{princi} = s+1-(k+l+1) = s-k-l$ nœuds du chemin principal de Q compris entre c et le nœud résultat f, parmi lesquels on suppose y avoir $p_{princi},\ 0 \leq p_{princi} \leq k_{princi}$ nœuds préférences.

A l'aide d'une telle requête, présentons comment s'effectue le remplissage de la table *preferenceTable*. Elle comporte pour le cas de cet exemple au moins deux colonnes étiquetées b et g (ce sont les deux seuls nœuds préférences effectivement mis en exergue dans la requête Q). De façon générale, nous examinerons le cas où le nœud préférence est situé sur le chemin principal et celui dans lequel il ne l'est pas.

• **Cas du nœud préférence « b » situé sur le chemin principal**

Les occurrences $f_j\left(f_{js}, f_{je}, f_{jl}\right)$[12] de f intégrant une occurrence $b_i(b_{is}, b_{ie}, b_{il})$ de b satisfont : $\begin{cases} b_{is} < f_{js} < f_{je} < b_{ie} & (1) \\ f_{jl} = b_{il} + m, \quad m \in \{K-p, K-p+1, \ldots, K\} & (2) \end{cases}$

L'équation (1) exprime le fait que $b_i(b_{is}, b_{ie}, b_{il})$ recouvre $f_j\left(f_{js}, f_{je}, f_{jl}\right)$. Tenant compte de ce que le (sous-)chemin allant de bi à fi peut éventuellement contenir une occurrence pour chacun des p nœuds préférences compris entre les nœuds b et f, l'équation (2) exprime le fait que f_j doit être situé à une profondeur comprise entre $(K-p)$ et (K) de bi.

• **Cas du nœud préférence « g » situé dans un chemin prédicat**

Les occurrences $f_j\left(f_{js}, f_{je}, f_{jl}\right)$ de f qui intègrent une occurrence $g_i(g_{is}, g_{ie}, g_{il})$ de g ayant pour pivot l'occurrence $c_v\,(c_{vs}, c_{ve}, c_{vl})$ de c[13] satisfont :

$$\begin{cases} c_{vs} < f_{js} < f_{je} < c_{ve} & /*\, c_v \textit{ recouvre } f_j\, */ & (3) \\ c_{vs} < g_{is} < g_{ie} < c_{ve} & /*\, c_v \textit{ recouvre } g_i\, */ & (4) \\ f_{jl} = c_{vl} + m, & m \in \{k_{princi} - p_{princi}, \ldots, k_{princi}\} & (5) \\ g_{il} = c_{vl} + n, & n \in \left\{k_{pred} - p_{pred}, \ldots, k_{pred} + 1\right\} & (6) \end{cases}$$

Les équations (3) et (4) expriment le fait que c_v doit recouvrir à la fois f_j et g_i. De même que pour le cas précédent, le (sous-)chemin allant de c_v à f_j (resp. de c_v à g_i) contient éventuellement des occurrences des k_{princi} (resp. p_{pred}) nœuds optionnels compris entre les nœuds c_v et f_j (resp. c_v et g_i). L'équation (5) (resp. 6) exprime le fait que f_j (resp. g_i) doit être situé à une profondeur comprise entre $(k_{princi} - p_{princi})$ et (k_{princi}) (resp. $(k_{pred} + 1 - p_{pred})$ et $(k_{pred} + 1)$ de c_v) de c_v.

Ainsi donc, pour tout résultat f_j de la requête, pour un nœud préférence b situé sur le chemin principal, l'entrée $preferenceTable[f_j, b] = 1$ s'il existe au moins une occurrence b_{is} de b dans la liste présente dans l'entrée $infoPrefNodeTable[b]$ recouvrant f_j (équation 1) et située à la "bonne profondeur"' dans le sous arbre ayant b_{is} pour racine (équations 2), sinon, elle est mise à zero. De même, pour un nœud préférence g situé dans

11. Tous appartenant au chemin principal.
12. Rappelons que f_{js}, f_{je}, f_{jl} sont respectivement les composants du triplet (start, end, level) représentant le nœud f_j dans le document.
13. Rappel : g a pour pivot c.

un prédicat, l'entrée $preferenceTable[f_j, g] = 1$ s'il existe au moins une occurrence c_{vs} du pivot c de g dans la liste présente dans $answerPivotTable[c]$ recouvrant à la fois f_j (équation 3) et au moins une occurrence g_{is} de g dans la solution partielle contenue dans $infoPrefNodeTable[g]$ (équation 4). f_j et g_{is} doivent être situées à la bonne profondeur dans le sous arbre ayant c_{vs} pour racine (équations 5 et 6) ; sinon, elle est mise à zero.

Enfin, l'*opérateur skyline* est appliqué aux tuples de la table $PreferenceTable$ ainsi construite pour déterminer les meilleures solutions.

5. Conclusion

Nous avons exploré dans ce papier une approche d'expression et d'évaluation de requêtes *XPath* avec préférences. Pour ce faire le langage *PrefSXPath* a été proposé et ainsi qu'un algorithme d'évaluation des mots de ce langage (des requêtes) sur un document XML.

Bien que l'algorithme proposé dans ce manuscrit ait été déroulé sur bien des exemples (faute de place, ils n'ont pas été déroulés ici) avec des résultats très satisfaisants, une étude analytique complète de ses performances est en cours de réalisation. Le travail présenté dans ce papier est le point de départ d'un travail plus ambitieux ayant pour but la généralisation de l'approche présentée dans ce manuscrit. On projette de considérer les préférences comme des *aspects* exprimables via un DSL (Domain Specific Language) qu'on construira, et de les injecter dans les algorithmes d'évaluations d'expressions *XPath* déjà existants par un *tisseur de préférences* qu'on définira.

6. Bibliographie

[1] Abiteboul, S. *« Querying Semi-Structured Data »*, In Proceedings of the International Conference on Database Theory (ICDT), Delphi, Greece, pp. 1-18, 1997.

[2] Bing Sun, Bo Zhou, Nan Tang, Guoren Wang, Ge Yu, and Fulin Jia. *« Answering XML Twig Queries with Automata »*, In Jeffrey Xu Yu, Xuemin Lin, Hongjun Lu, and Yanchun Zhang, editors, Advanced Web Technologies and Applications, 6th Asia-Pacific Web Conference, APWeb 2004, Hangzhou, China, April 14-17, 2004, Proceedings, volume 3007 of Lecture Notes in Computer Science, pp. 170-79. Springer, 2004.

[3] Bosc P., Pivert O. *« SQLf : a relational database language for fuzzy querying »*, IEEE Trans. On Fuzzy Systems, vol(3) pp.1-17, 1995.

[4] C.Y. Chan, P. Felber, M.N. Garofalakis, R. Rastogi, *« Efficient filtering of XML documents with XPath expressions »*, in : Proceedings of the 18th International Conference on Data Engineering (ICDE), IEEE Comput. Soc., pp. 235-244, 2002

[5] Chomicki J. *« Preference Formulas in Relational Queries »*, In ACM Trans. on Database Systems (TODS), vol. 28(4), 2003.

[6] Dubois D., Prade H. *Bipolarity in flexible quering.* Proc. of the 5th Int. Conf. on Flexible Query Answering Systems (FQAS), Copenhagen, Denmark, 2002.

[7] Gang Gou and Rada Chirkova. *« Efficiently Querying Large XML Data »*, Repositories : A Survey IEEE Transactions On Knowledge And Data Engineering, VOL. 19, NO. 10, pp. 1381-1402 OCTOBER 2007.

[8] Kiebling W. *« Foundations of Preferences in Database Systems »*, In Proc. of the 28th Int. Conf. on Very Large Databases (VLDB), pp. 311-322., Hong Kong, China, 2002.

[9] Lietard L., Rocacher D. and Tbahriti S.-E. *« Preferences and Bipolarity in Query Language »*, , International Conference of the North American Fuzzy Information Processing Society (NAFIPS 2008), New-York, USA. pp. 1-6, 2008.

[10] Q. Li and B. Moon. *« Indexing and querying XML data for regular path expressions »*, Proceedings of the 27th VLDB Conference, pp. 361-370, 2001.

[11] Rakesh Agrawal , Jerry Kiernan , Ramakrishnan Srikant , Yirong Xu, *« An XPath-based preference language for P3P »*, Proceedings of the 12th international conference on World Wide Web, May 20-24, Budapest, Hungary, 2003

[12] Stephan Börzsönyi, Donald Kossmann, and Konrad Stocker. *« The skyline operator »*, In Proceedings of the 17th International Conference on Data Engineering, April 2-6, 2001, Hiedelberg, Germany, pp. 421-430, 2001.

[13] T. Green, G. Miklau, M. Onizuka, D. Suciu, *« Processing XML streams with deterministic automata »*, in : Proceedings of the 9th International Conference on Database Theory (ICDT), Springer, pp. 173-189, 2003.

[14] W. Kießling, B. Hafenrichter, S. Fischer, S. Holland : *« Preference XPATH : A Query Language for E-Commerce »*, Proc. 5th Intern. Konf. für Wirtschaftsinformatik, Augsburg, pp. 425-440, Sept. 2001.

[15] W3C Consortium. *« XML Path Language (XPath) 2.0 »*, http ://www.w3.org/TR/XPath20/, 2006.

[16] W3C Consortium. *« XQuery 1.0 : An XML Query Language »*, http ://www.w3.org/TR/xquery/. 2006.

[17] Yao, J. T. and Zhang, M. *«A Fast Tree Pattern Matching Algorithm for XML Query », Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, (WI 2004), 20-24 September 2004, Beijing, China, pp. 235-241, 2004.

Empirical study of LDA for Arabic topic identification

Marwa Naili, Anja Habacha Chaibi and Henda Ben Ghézala

RIADI-ENSI
University of Manouba
Manouba 2010, Tunisia

maroua.naili@riadi.rnu.tn

RÉSUMÉ. Cet article met l'accent sur l'identification thématique pour la langue arabe. Nous étudions l'Allocation de Dirichlet Latente (LDA) comme une méthode non supervisée pour l'identification thématique. Ainsi, une étude approfondie de LDA a été effectuée à deux niveaux: le processus de lemmatisation et le choix des paramètres. Pour le premier niveau, nous étudions l'effet des différents lemmatiseurs sur LDA. Pour le deuxième niveau, nous nous focalisons sur les paramètres de LDA et leurs impacts sur l'identification. Cette étude montre que LDA est une méthode efficace pour l'identification thématique Arabe surtout avec le bon choix des paramètres. Un autre résultat important est l'impact élevé des lemmatiseurs sur l'identification thématique.

ABSTRACT. This paper focuses on the topic identification for the Arabic language. We study the Latent Dirichlet Allocation (LDA) as an unsupervised method for the Arabic topic identification. Thus, a deep study of LDA is carried out at two levels: Stemming process and the choice of LDA parameters. For the first one, we study the effect of different Arabic stemmers on LDA. For the second one, we focus on LDA parameters and their impact on the topic identification. This study shows that LDA is an efficient method for Arabic topic identification especially with the right choice of parameters. Another important result is the high impact of stemming algorithms on topic identification.

MOTS-CLÉS : Identification thématique, Allocation de Dirichlet Latente, paramètres de LDA, lemmatiseurs Arabes.

KEYWORDS: Topic identification, Latent Dirichlet Allocation, LDA parameters, Arabic stemmers.

1. Introduction

During the last few years, the number of textual documents has been vastly increasing. Thus, many techniques have been presented to deal with this big number of documents. However, the real challenge is to manage these documents based on their content, especially the thematic one. For this reason, topic Identification and classification draw a lot of intention in research fields dealing with different types of documents (text [7], XML [2], etc). Yet for Arabic textual documents, there is a flagrant lack of research. This can be explained by the high complexity of this language and the lack of Arabic resources. In this paper, we will focus on topic identification by studding LDA as an unsupervised method for Arabic topic identification.

This paper is organized as follows: Section 2 presents an overview of Arabic topic identification; Section 3 describes some Arabic stemmers; Section 4 deals with LDA; Section 5 is dedicated to the evaluation and the discussion; finally, the conclusion and future works are presented in section 6.

2. Overview of Arabic topic identification

Topic identification is the process of identifying the topic of a textual unity. According to most researchers, a topic is a cluster of words which are closely related to the topic. Clusters depend on the stemming process that specifies the type of words (root, stem, etc). For the Arabic topic identification, some methods have been used as:

- *TF-IDF [7]:* allows the construction of a vector space. Each vector represents a document by the combination between TF(w,d) and IDF(w). The topic with the highest similarity with the document will be considered as the document's topic.

- *SVM and MSVM* [13]: is a supervised method which classifies documents into two classes by constructing a hyperplane separator in the R^Nvector space. Yet, when the number of categories is superior than 2, the MSVM is used. In fact, the idea of this method is to find *n* hyperplane with *n* corresponds to the number of categories.

-*TR-Classsifier* [7]: is based on triggers which are identified by using the Average Mutual Information. In fact, topics and documents are presented by triggers which are a set of words that have the highest degree of correlation. Then, based on the TR-distance, the similarity is calculated between triggers to identify the topic of the document.

-*Named Entities approach* [10]: The idea of this approach is to reduce the dimension of vectors by using only the segments bounded by named entities pairs. Then, the mutual information is used to calculate similarity between topics and documents. Besides these methods, we can cite other methods used for topic identification such as

TULM and Neural networks in [7]. However, the major limit of these methods is that a training step is necessary to identify the topics and to construct a vocabulary for each topic. Thus, we opted to use the unsupervised method LDA. That means that there is no need to a training step because topics are identified in the process of topic identification.

3. Arabic stemmers

Arabic language is one of the most complex and ambiguous language because of its wide variety of grammatical forms and its complex morphology. Thus, the stemming process is more difficult for the Arabic language than other languages. The stemming process aims to find the lexical root or lemma of words by removing prefixes and suffixes with are attached to its root. As an example of Arabic stemmers we mention:

- *Khoja Stemmer [11]:* it extracts the root of a word by removing the longest suffix and prefix and then by matching the rest with verbal and nouns patterns.

- *ISRI Arabic Stemmer* [5]: it extracts the root of a word. But, unlike Khoja Stemmer, it doesn't use any root dictionary or lexicon.

- *The Buckwalter Arabic Morphological Analyzer* [12]: it returns the stems of words based on lexicons of stems, prefixes, suffixes and morphological compatibility tables.

- *Light Stemmer* [6]: Unlike Khoja Stemmer, it removes some defined prefixes and suffixes instead of extracting the original root words.

According to different studies [5,6] the most efficient stemmers are Khoja and Light Stemmers. These two stemmers are available freely on the web and might be the only available Open Source ones. Thus, we will study Khoja and Light Stemmers to evaluate the effect of the stemming process on the topic identification.

4. Latent dirichlet allocation (LDA)

LDA [3] is a generative model in which documents are represented as a mixture of topic. Each topic is a multinomial distribution over words that depends on the stemming process. Therefore, for each document *w* in the corpus *D*, the generative process is:

1. We choose N (a document is a sequence of N words) according to Poisson distribution (N ~ Poisson(ξ))
2. We choose θ (θ_d is the distribution over the topic of the document d) according to dirichlet allocation ($\theta \sim Dirichlet(\alpha)$)
3. For each of the N words w_n : Choose a latent topic z_n according to a multinomial distribution and choose a word $w_n$from $p\ (w_n|z_n, \beta)$

The θ variable takes values in the *(k-1)* simplex and its density is equal to:

$$\mathrm{p}(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_k - 1} \qquad (1)$$

Where $\alpha \in \mathbb{R}^k$, $\alpha_i > 0$ and $\Gamma(x)$ is the Gamma function.

Therefore, given α and β, the joint distribution of θ, *z* and *w* is equal to:

$$p(\theta, z, w|\alpha, \beta) = \mathrm{p}(\theta|\alpha) \prod_{n=1}^{N} \mathrm{p}(z_n|\theta)\ \mathrm{p}(w_n|z_n, \beta) \qquad (2)$$

Finally, by integrating over θ and summing over *z*, the marginal distribution of a document is as follow (equation 3):

$$p(w|\alpha, \beta) = \int \mathrm{p}(\theta|\alpha)\left(\prod_{n=1}^{N} \sum_{z_n} \mathrm{p}(z_n|\theta)\mathrm{p}(w_n|z_n, \beta)\right) \mathrm{d}\theta \qquad (3)$$

According to Steyvers and Griffiths [8], the choice of α and β has an effect on the performance of LDA. Besides, these parameters depend on the number of topics and the vocabulary size. Moreover, Steyvers and Griffiths [8] recommended to use $\alpha = 50/k$ and $\beta = 0.01$. However, Lu et al. [14] conduct an in-depth analysis of the choice of α with $\beta = 0.01$. According to this analysis, the performance of LDA is influenced by the initializing choiceα. This choice also depends on the field of application such as topic classification and information retrieval which are tested in this study. As result, they found that, for the topic classification, the optimal performance is obtained by α between 0.1 and 0.5. Yet, for information retrieval, the optimal performance is obtained by α between 0.5 and 2. However, according to Lu et al. [14], the best value of α is not stable and it depends on the collection of documents used for tests. On the other hand, Heinrich [4] estimated the values of α and β by using the information available from the Gibbs sampler. In fact, Heinrich [4] showed that hyper-parameters are best estimated as parameters of the Dirichlet-multinomial distribution.

Despite the high performance of LDA, few works dealing with LDA were presented in the field of Arabic topic identification [9,1]. According to these works, promising results have been obtained by LDA. However, we note that no one has studied LDA parameters in the field of topic identification. Therefore, in this paper, we will study in depth the LDA by studding the choice of α and more important the effect of different stemming algorithms to enhance the quality of topic identification.

5. Evaluation and discussion

In this section, we evaluated LDA with different stemmers. Thus, we presented three different versions: LDA-WS (**W**ithout **S**temmer), LDA-KS (**K**hoja **S**temmer) and LDA-LS (**L**ight **S**temmer). For this evaluation, we use the Arabic benchmark Al-Watan

which contains 20291 articles from Watan newspaper and it covers six topics: culture (2782 documents), economy (3468 documents), international news (2035 documents), local news (3596 documents), religion (3860 documents) and sport (4550 documents). To report the evaluation results, we use three metrics: **R**ecall, **P**recision and **F**-measure.

5.1. Identified topics based on different stemmers

	Culture	Economy	International News	Local News	Religion	Sport
LDA-WS	الله (god)	مليون (million)	قال (said)	السلطنة (sultanate)	الله (god)	المباراة (match)
	الإسلام (Islam)	ريال (real)	العراق (Iraq)	العمل (work)	قال (said)	المنتخب (team)
	الحياة (life)	عام (public)	المتحدة (united)	العام (the public)	صلى (pray)	الاول (first)
	الناس (people)	الدول (countries)	الامريكية (American)	العامة (the public)	وسلم (salaam)	المركز (position)
	الإسلامية(Islamic)	السلطنة(sultanate)	عام (public)	محمد (Mohammed)	رسول (prophet)	الثاني (second)
LDA-KS	علم(knowledge)	دول (countries)	عرق (vein)	جمع (collect)	سلم (salaam)	لعب (play)
	كون (universe)	شرك (share)	روس (Russian)	دور (role)	قول (saying)	فرق (teams)
	عمل (work)	عمل (work)	حكم (rule)	علم (knowledge)	صلى (pray)	تعهد (pledge)
	كتب (write)	صنع (production)	دول (countries)	عمل (work)	كون (universe)	دور (role)
	جمع (collect)	عوم (launch)	عمل (work)	قوم (nation)	رسل (Russell)	بطل (champion)
LDA-LS	اسلام (Islam)	شرك (share)	عراق (Iraq)	عام (public)	قال (said)	فريق (team)
	عرب (Arab)	عام (public)	اميرك (American)	عمل (work)	صل (pray)	منتخب (team)
	فن (art)	اقتصاد(economy)	دول (countries)	عمان (Amman)	رسول (prophet)	دور (role)
	كتاب (book)	دول (countries)	قال (said)	دور (role)	سلم (salaam)	مبارا (match)
	عالم (world)	قطاع (sector)	رئيس (president)	تعليم (education)	مسلم (Muslim)	بطول(championship)

Figure 1. ***Identified topics based on LDA-WS, LDA-KS and LDA-LS*.**

By conducting the three versions of LDA on AL-Watan corpus, we were able to identify all the six topics. As shown in Figure.1, the identified topics depend on the used stemmer. In fact, without using any stemming algorithms, the different topics were successfully identified by LDA-WS. However, the problem is that some words can figure more than once with different affix or suffix such as العام and العامة which mean public. This problem is resolved by using Khoja stemmer which extracts the root of words. Thus, by employing LDA-KS, the topics are present by roots. The limit of this method is that a root can have several meaning such as علم which has many meaning like: knowledge, flag, aware. Therefore, by using Khoja Stemmer, we might lose the meaning. Yet, Light Stemmer removes only the prefix to maintain the meaning such as the word المنتخب (the team) without stemming, نخب (pledge) with Khoja Stemmer and منتخب (team) with Light Stemmer. As conclusion, all the six topics have been successfully identified by LDA. Moreover, Light Stemmer is the most efficient stemmer because it solves the problem of repetition (which is caused by the absence of stemmer: LDA-WS) and the loss of meaning (which is caused by Khoja Stemmer LDA-KS).

5.2. Study of LDA parameter (α)

We study in depth the α parameter of LDA by using three values 0.1, 0.5 and 50/k (k is number of topics which is 6 in our study). These values are proposed by [8,14]. For

β, we used $\beta = 0.01$ which is recommended in most research. For each value of α, the results of LDA-WS, LDA-KS and LDA-LS are illustrates in table 1. First of all, we remark that LDA-LS is independent of α. Yet, LDA-WS and LDA-KS are strongly influenced by α and the best results are obtained by $\alpha = 0.5$. Furthermore, for α = 0.5, the results of LDA-LS and LDA-KS are very close. Based on this result and the results of the stemming process for the topic identification, Light Stemmer is the most efficient stemmer to use with LDA. In the other hand, regardless of the value of α and the stemming algorithm, the well identified topics are: sport (F = 91.86%), religion (F = 82.75%), economy (F = 75.13%). Yet, for the other topics, especially the culture topic, the performance of LDA is not stable. This can be explained by the fact that the vocabularies of these topics (culture, international and local news) are very close.

			Culture	Economy	Intern News	Local News	Religion	Sport	Average
LDA-WS	$\alpha = 0.1$	**R**	9.09%	70.10%	95.23%	84.73%	50.34%	85.25%	65.79%
		P	12.02%	**80.95%**	47.53%	**58.73%**	96.00%	**99.59%**	65.80%
		F	10.36%	**75.13%**	63.42%	**69.38%**	66.04%	**91.86%**	62.70%
	$\alpha = 0.5$	**R**	48.56%	70.30%	97.49%	81.01%	61.11%	84.13%	**73.77%**
		P	**46.73%**	79.72%	**67.21%**	56.98%	**97.16%**	99.43%	**74.54%**
		F	**47.63%**	74.72%	**79.57%**	66.90%	**75.03%**	91.14%	**72.50%**
	$\alpha = 50/k$	**R**	46.62%	69.49%	97.59%	80.70%	60.18%	84.28%	73.14%
		P	45.40%	79.04%	66.22%	56.47%	97.11%	99.48%	73.95%
		F	46.00%	73.96%	78.90%	66.44%	74.31%	91.25%	71.81%
LDA-KS	$\alpha = 0.1$	**R**	68.40%	64.27%	78.52%	50.08%	71.35%	75.82%	68.07%
		P	55.53%	57.72%	52.62%	50.75%	93.58%	99.34%	68.26%
		F	61.30%	60.82%	63.01%	50.41%	80.96%	86.00%	67.08%
	$\alpha = 0.5$	**R**	69.55%	54.67%	95.92%	78.28%	73.70%	79.98%	**75.35%**
		P	**55.76%**	**82.87%**	**76.28%**	**53.18%**	**94.33%**	99.29%	**76.95%**
		F	**61.90%**	**65.88%**	**84.98%**	**63.34%**	**82.75%**	**88.59%**	**74.59%**
	$\alpha = 50/k$	**R**	68.44%	63.98%	90.47%	50.78%	70.72%	75.54%	69.99%
		P	54.84%	57.79%	61.02%	50.79%	93.85%	**99.39%**	69.61%
		F	60.89%	60.73%	72.88%	50.78%	80.66%	85.84%	68.63%
LDA-LS	$\alpha = 0.1$	**R**	60.71%	63.32%	97.00%	77.11%	59.09%	83.49%	73.45%
		P	49.38%	**75.88%**	74.18%	54.20%	96.24%	**99.19%**	74.84%
		F	54.47%	**69.03%**	84.07%	63.66%	73.23%	90.67%	72.52%
	$\alpha = 0.5$	**R**	63.73%	62.51%	96.36%	77.14%	65.72%	83.54%	**74.83%**
		P	**54.19%**	75.54%	**75.60%**	**54.57%**	**96.10%**	**99.19%**	**75.86%**
		F	**58.57%**	68.41%	**84.73%**	**63.92%**	78.06%	**90.69%**	**74.06%**
	$\alpha = 50/k$	**R**	62.98%	62.92%	96.46%	76.42%	65.78%	83.36%	74.65%
		P	54.12%	75.47%	75.50%	53.97%	**96.10%**	99.06%	75.70%
		F	58.21%	68.63%	84.70%	63.26%	**78.10%**	90.53%	73.90%

Table 1. ***LDA-WS, LSA-KS and LDA-LS results with*** $\alpha = 0.1$, $\alpha = 0.5$ **and** $\alpha = 50/k$.

But the vocabularies of sport, religion and economy are more representative and unique for each topic which leads to an efficient topic identification.

5.3. Comparison with related works

To evaluate our work, we choose to compare our methods (LDA-KS and LDA-LS) with the works of Abbas et al. [7] and Koulali and Meziane [10]. The reason for this choice is that we used the same test corpus for the evaluation. Yet, we note that in these works [7,10], 90% of the corpus is used for the training step and only 10% for the test. This can explain the high performance of TF-IDF [7], MSVM [7], TR-Classifier [7] and the Named Entities approach (NE) [10]. However, as an unsupervised method which does not need any kind of training step, the results of LDA-KS and LDA-LS are promising. In fact, dispute culture and economy topics, the result for the rest of topics are comparable and even better some times. For example, for the international news topic, LDA-KS and LDA-LS are better than TF-IDF, MSVM and TR-classifier.

Works	**Culture**	**Economy**	**Intern News**	**Local News**	**Religion**	**Sport**	**Average**
TF-IDF	78.96%	90.03%	81.96%	78.43%	88.60%	96.91%	**86.04%**
MSVM	76.47%	95.50%	79.02%	68.64%	84.83%	89.75%	82.44%
TR-Classifier	81.60%	89.50%	83.77%	84.35%	91.97%	96.66%	88.02%
NE	75.66%	78.14%	90.15%	77.08%	88.26%	95.46%	84.15%
LDA-KS	61.90%	65.88%	84.98%	63.34%	82.75%	88.59%	74.59%
LDA-LS	58.57%	68.41%	84.73%	63.92%	78.06%	90.69%	74.06%

Table 2. *Comparison with related works*.

6. Conclusion

In this paper, we presented a deep study of LDA in the field of Arabic topic identification. In fact, we studied the effect of the stemming process on topic identification by using Arabic stemmers (Khoja and Light Stemmers). Besides, we studied in depth the parameters of LDA. As result, we showed that the choice of parameters influence the performance of LDA and the best result are obtained by $\alpha = 0.5$. Moreover, LDA depends on the stemming algorithms. Based on our evaluation, Light Stemmer is the best stemmer for the topic identification. Thus, based on the best choice of parameters and the stemming algorithm, the result of LDA is very promising in the field of topic identification. For further studies, we will use LDA for topic segmentation to realize a complete topic analysis of Arabic documents.

6. References

[1] A. Kelaiaia and H.F. Merouani. “Clustering with Probabilistic Topic Models on Arabic Texts”. In *Modeling Approaches and Algorithms for Advanced Computer Applications, Springer*, 65-74, 2013.

[2] A.A.Y. Yassine, and K. Amrouche. "Réseaux bayésiens jumelés et noyau de Fisher pondéré pour la classification de documents XML.", ARIMA Journal, Special issue CARI’12, 17:141-154, 2014.

[3] D.M. Blei, A.Y. Ng and M.I. Jordan, "Latent dirichlet allocation". The Journal of machine Learning research, 3, 993-1022, 2003.

[4] G. Heinrich. "Parameter estimation for text analysis". *University of Leipzig, Tech. Rep*, 2008.

[5] K. Taghva, R. Elkhoury and J. Coombs, "Arabic stemming without a root dictionary". International conference on Information Technology, 1:52-57, 2005.

[6] L. Larkey, L. Ballesteros and M. Connell, *Light stemming for Arabic information retrieval*. Arabic Computational Morphology, book chapter, Springer, 2007.

[7] M. Abbas, K. Smaïli and D. Berkani. "Evaluation of Topic Identification Methods on Arabic Corpora". *JDIM*, *9*(5), 185-192, 2011.

[8] M. Steyvers and T. Griffiths. *Probabilistic topic models*. Handbook of latent semantic analysis, 427(7):424-440, 2007.

[9] M. Zrigui, R. Ayadi, M. Mars and M. Maraoui, "Arabic text classification framework based on latent dirichlet allocation". *CIT*. Journal of Computing and Information Technology, 20(2): 125-140, 2012.

[10] R. Koulali and A. Meziane, "Feature Selection for Arabic Topic Detection Using Named Entities". In Proceeding of CITALA, Oujda, Morocco, pp. 243-246, 2014.

[11] S. Khoja and R. Garside, "Stemming Arabic text". Computer science, UK, 1999.

[12] T. Buckwalter, "Buckwalter Arabic morphological analyser version 2.0". LDC2004L02, ISBN 1-58563-324-0, 2004.

[13] V. Vapnik, "The natural of statitical learning theory". Springer, New York, 1995.

[14] Y. Lu, M. Qiaozhu and Z. ChengXiang. "Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA."*Information Retrieval* 14(2):178-203, 2011.

Approche hybride pour le développement d'un lemmatiseur pour la langue arabe

Mohamed Boudchiche et Azzeddine Mazroui

Département de Mathématiques et Informatique
Faculté des Sciences, Université Mohammed Premier, Oujda, Maroc
B-P 717, 60000, OUJDA
MAROC

moha.boudchiche@gmail.com, azze.mazroui@gmail.com

RÉSUMÉ. Nous présentons dans cet article un système d'analyse morphologique arabe qui attribue, pour chaque mot d'une phrase arabe, un lemme unique en tenant compte du contexte des mots. Le système proposé est composé de deux modules. Le premier consiste en une analyse hors contexte basée sur l'analyseur morphosyntaxique Alkhalil Morpho Sys 2. Dans le deuxième module, nous utilisons le contexte pour identifier le bon lemme parmi tous les lemmes possibles du mot obtenus par le premier module. A cet effet, nous utilisons une approche basée sur les modèles de Markov cachés, où les observations sont les mots de la phrase et les lemmes représentent les états cachés. Nous validons l'approche en utilisant un corpus étiqueté composé d'environ 500000 mots. Le système donne le bon lemme dans plus de 99,5% pour l'ensemble d'apprentissage et environ 94,3% pour l'ensemble de test.

MOTS-CLES : Traitement automatique de la langue arabe, Lemmatisation, Analyseur morphologique, Model de Markov caché, Algorithme de Viterbi.

1. Introduction

L'Internet connaît depuis quelques années une croissance exponentielle dans le domaine de la recherche d'information. Ainsi, les chercheurs ont développé pour certaines langues plusieurs outils permettant d'analyser et d'extraire l'information utile dans les documents numériques. Cependant, les différences entre les structures linguistiques des différentes langues ne permettent pas toujours d'étendre l'utilisation des programmes développés pour une langue donnée à une autre langue.

Dans le domaine du traitement automatique des langues naturelles (TALN), la lemmatisation occupe une place importante étant donné son utilisation dans plusieurs application du TALN telles que la traduction automatique, l'indexation, les résumeurs automatiques, la classification des textes et les dictionnaires interactifs [2, 3, 8] . En particulier, des travaux récents dans les systèmes de recherche d'information en langue arabe ont montré l'utilité de travailler avec les lemmes au lieu des mots.

La lemmatisation consiste à identifier pour chaque mot du texte son lemme qui représente la forme minimale du mot portant son sens principal. Les lemmes représentent les entrées des dictionnaires. Pour la langue arabe, le lemme d'un verbe est sa forme sans clitiques conjugué à l'accompli à la 3ème personne du singulier (le lemme du verbe 'سيمارسون/*fsymArswn*/ est 'مارس' /*mArs*/). Pour un nom, le lemme est sa forme au singulier masculin sans clitiques (le lemme du nom 'كمعلماتهم /*kmElmAthm*/ est 'معلم/*mElm*/). Si le nom n'a pas de masculin, alors son lemme est sa forme au singulier féminin (le lemme de 'بمدارسهم/*bmdArshm*/ est 'مدرسة' /*mdrsp*/). Enfin, pour une particule, le lemme est la particule sans clitiques (le lemme de 'كالّذي /*kAl*y*/ est 'الّذي /*Al*y*/).

Afin de répondre à une demande de plus en plus forte de lemmatiseurs pour la langue arabe, nous avons développé un système qui fournit les lemmes des mots d'une phrase arabe. Notre système commence par réaliser une analyse morphologique en utilisant la deuxième version de l'analyseur morphologique Alkhalil morpho Sys [1]. Cette analyse permet l'obtention pour chaque mot pris hors contexte ses différents lemmes potentiels. Pour identifier le lemme correct parmi ces lemmes potentiels, nous avons utilisé dans une deuxième étape les modèles de Markov cachés et l'algorithme de Viterbi. Afin de réaliser les phases d'apprentissage et de test, nous avons utilisé le corpus Nemlar [9] pour lequel nous avons ajouté au préalable l'étiquette lemme à tous ses mots.

L'article est organisé de la manière suivante. Nous présentons dans la deuxième section le corpus Nemlar utilisé dans les deux étapes d'apprentissage et de test. Nous consacrons la section suivante pour un aperçu sur l'analyseur Alkhalil Morpho Sys utilisé dans la phase morphologique de notre système. Le paragraphe 4 est réservé à une description de la méthode adoptée dans le développement du lemmatiseur. Les résultats de l'évaluation du système sont détaillés au paragraphe 5 et nous terminons le papier par une conclusion.

2. L'Analyseur Alkhalil Morpho Sys[1]

AlKhalil Morpho Sys 2 [1] est un analyseur morphosyntaxique développé avec le langage de programmation orienté objet Java par le Laboratoire de Recherche en Informatique de l'Université Mohammed Premier, Oujda, Maroc. Il permet d'analyser aussi bien les mots arabes non voyellés que les mots partiellement ou totalement voyellés. L'analyse se fait hors contexte et les taches de l'analyseur pour un mot donné sont :

- retrouver les voyellations possibles du mot (lorsque le mot entré n'est pas voyellé),
- identifier pour chaque voyellation possible du mot son lemme accompagné du schème, les clitiques attachés au mot, sa catégorie grammaticale et son stem accompagné du schème.

Nous avons utilisé cet analyseur dans la première phase de notre système.

3. Description de la méthode

La lemmatisation des mots des textes arabes sera réalisée en deux étapes. Dans la première étape, le système utilise la deuxième version de l'analyseur morphologique Alkhalil Morpho Sys pour analyser les mots de la phrase. Ainsi, l'analyseur nous fournit les différents lemmes potentiels de chaque mot. Ensuite, un traitement statistique basé sur les chaines de Markov cachées et l'algorithme de Viterbi sera réalisé dans la deuxième phase. L'objectif de ce traitement est la désambiguïsation qui consiste à identifier le lemme correct dans le contexte parmi les lemmes potentiels d'un mot obtenus dans la phase morphologique.

3.1. Analyse morphologique

Après une phase de prétraitement du texte entré (tokénisation, normalisation des mots, découpage des textes en phrase puis en mots), ces derniers subissent une analyse morphologique en utilisant la 2ème version de l'analyseur morphologique Alkhalil Morpho Sys. Nous obtenons ainsi tous les lemmes potentiels de chaque mot du texte pris hors contexte accompagnées de leurs informations morphosyntaxiques. En effet, pour chaque voyellation du mot, le système fournit les clitiques attachés aux stems, les POS tags, le stem et le lemme. Dans le cas d'un nom ou d'un verbe, le système fourni également la racine, les schèmes du stem et du lemme et l'état syntaxique.

3.2. Analyse statistique

Après avoir identifié les lemmes potentiels pour chaque mot de la phrase, nous appliquons un traitement statistique dont l'objectif est la sélection du lemme le plus probable parmi ces lemmes potentiels. Ce traitement est basé sur les modèles de Markov cachés, les techniques de lissage et l'algorithme de Viterbi.

Nous donnons dans la suite un bref aperçu de ces trois concepts mathématiques.

[1] http://oujda-nlp-team.net/?p=1299&lang=en

3.3. Modèles de Markov Cachés

Les modèles de Markov cachés (HMM) sont utilisés pour modéliser deux processus aléatoires dépendants dont les états du premier sont non observables (états cachés), et ceux du second sont observables (états observés). Les HMM servent à prédire les états cachés à partir des états observés.

En effet, si $O = \{o_1, o_2, \dots, o_r\}$ est un ensemble fini d'observations et $E = \{h_1, h_2, \dots, h_m\}$ est un ensemble fini d'états cachés, alors un double processus $(X_t, Y_t)_{t\geq 1}$ est un modèle Markov caché du premier ordre si :

- $(X_t)_t$ est une chaîne de Markov homogène à valeurs dans l'ensemble d'états cachés E vérifiant : $Pr(X_{t+1} = h_j / X_t = h_i, \dots, X_1 = h_k) = Pr(X_{t+1} = h_j / X_t = h_i) = a_{ij}$. a_{ij} est la probabilité de transitions de l'état caché h_i vers l'état caché h_j.
- $(Y_t)_t$ est un processus observable qui prend ses valeurs dans l'ensemble d'observations O vérifiant : $Pr\left(Y_t = o_k / X_t = h_i, Y_{t-1} = o_{k_{t-1}}, X_{t-1} = h_{i_{t-1}}, \dots, Y_1 = o_{k_1}, X_1 = h_{i_1}\right) = Pr(Y_t = o_k / X_t = h_i) = b_i(k)$.

  $b_i(k)$ est la probabilité d'observer l'état o_k étant donné l'état caché h_j.

Ainsi, les informations sur les états cachés peuvent être déduites à partir des données observées.

Soit S une phrase observée composée des mots $w_1, w_2, \dots, w_n$ et $E = \{l_1, l_2, \dots, l_m\}$ l'ensemble de tous les lemmes de la langue arabe.

Afin de rechercher les lemmes les plus probables dans le contexte des mots w_i de la phrase S, nous allons utiliser une modélisation par les HMM où les mots de la phrase représenteront les observations et leurs lemmes les états cachés.

Notre objectif est donc de trouver pour la phrase $S = (w_1, w_2, \dots, w_n)$ la séquence de lemmes la plus probable $(l_1^*, \dots, l_n^*)$ satisfaisant la relation suivante :

$$(l_1^*, \dots, l_n^*) = \underset{l_i \in L_i}{\operatorname{argmax}} Pr(l_1, \dots, l_n / w_1, \dots, w_n)$$

où L_i est l'ensemble des lemmes possibles du mot w_i obtenus suite à l'analyse morphologique de la première étape.

3.3.1. Algorithme de Viterbi

Pour trouver la séquence la plus probable des lemmes, nous allons utiliser l'algorithme de Viterbi [5], qui est bien adapté pour la recherche du chemin optimal. Ainsi, si nous notons $\emptyset(t, l_t^k)$ le maximum sur l'ensemble des chemins de longueur *(t-1)* de la probabilité que les *(t-1)* premiers mots aient les lemmes du chemin et le $t^{ème}$ mot w_t ait le lemme l_t^k , c.à.d. : $\emptyset(t, l_t^k) = \underset{\substack{l_i^{j_i} \in L_i \\ 1\leq i\leq t-1}}{\max} [\Pr(w_1, \dots, w_t / l_1^{k_1}, \dots, l_t^k)] \Pr\left(l_1^{k_1}, \dots, l_t^k\right)$,

alors, en utilisant les hypothèses markoviennes, nous pouvons facilement vérifier que :

$$\emptyset(t, l_t^k) = \left(\max_{l_{t-1}^j \in L_{t-1}} \emptyset(t-1, l_{t-1}^j) \Pr(l_t^k / l_{t-1}^j)\right) \Pr(w_t / l_t^k).$$

Cette équation permettra de calculer de manière récursive les valeurs de la fonction $\emptyset$.

Pour obtenir le chemin optimal, nous utilisons la fonction Ψ qui mémorise à l'instant t l'étiquette cachée qui réalise le maximum dans la définition de $\emptyset$. Elle est définie par :

$$\Psi(t, l_t^k) = \underset{l_{t-1}^j \in l_{t-1}}{argmax}\ \emptyset(t-1, l_{t-1}^j)\, Pr(l_t^k / l_{t-1}^j).$$

3.3.2. Méthodes de lissage

Afin de pouvoir programmer l'algorithme de Viterbi, il faut au préalable estimer les paramètres du modèle statistique, c'est à dire, les coefficients des matrices de transition et d'émission $A = (a_{ij})$ et $B = (b_i(t))$ où $a_{ij} = \Pr(l_j / l_i)$ et $b_i(t) = Pr(w_t / l_i)$.

Pour cela, nous avons appliqué sur un corpus d'apprentissage étiqueté de taille N la méthode d'estimation basée sur le maximum de vraisemblance [6].

Si w_t est un mot de la phrase S et (l_i, l_j) sont deux lemmes, alors nous notons :

- n_i : le nombre d'occurrences de l'état caché l_i dans le corpus C,
- n_{ij}: le nombre d'occurrences dans C de la transition de l'état caché l_i vers l'état l_j,
- m_{it} : le nombre de fois que le mot w_t correspond à l'état caché l_i dans le corpus C,

alors, les coefficients a_{ij} et $b_i(t)$ sont estimés en utilisant les équations suivantes :

$$a_{ij} = \frac{n_{ij}}{n_i}, 1 \le i \le N, 1 \le j \le N \quad \text{et} \quad b_i(t) = \frac{m_{it}}{n_i}, 1 \le t \le n, 1 \le i \le N$$

Etant donné qu'il n'existe pas de corpus d'apprentissage pouvant contenir toutes les transitions entre les mots de la langue arabe, les coefficients de transition peuvent pour certains exemples être estimés par la valeur zéro. Cela affectera négativement la recherche du chemin optimal par l'algorithme de Viterbi. Pour remédier à ce phénomène, des techniques de lissage sont alors utilisées. Ces techniques seront appliquées avant de faire tourner l'algorithme de Viterbi, et consistent à attribuer une probabilité non nulle à toutes les transitions du corpus de test. Pour cela, nous avons utilisé la méthode Absolute Discounting [4].

Ainsi, si $C = \{Ph_1, \dots, Ph_M\}$ est le corpus d'apprentissage de la langue arabe formé par *M* phrases Ph_k, et si nous posons :

- $N_{1+}(l_i \bullet)$: le nombre de tous les mots dont les lemmes correspondants sont répétés une fois et plus après le lemme li dans le corpus C,
- N_i: le nombre de mots annoté dans le corpus C avec le lemme l_i,
- z_i : le nombre de mots non annoté dans le corpus C avec le lemme l_i et pour lesquels l'analyseur Alkhalil génère ce lemme,

alors, les coefficients a_{ij} et $b_i(t)$ sont estimés par :

$$a_{ij} = \frac{max(n_{ij} - D, 0)}{n_i} + \frac{D}{n_i} P_{abs}(l_j) N_{1+}(l_i \bullet) \ et\ b_i(t) = \begin{cases} \frac{m_{it} - D}{n_i} & si\ m_{it} \neq 0 \\ \frac{N_i \times D}{n_i \times z_i} & sinon \end{cases}$$

avec la constante *D=0.5* et $P_{abs}(l_j) = \frac{n_j}{N}$

4. Corpus d'apprentissage et de test

Le projet NEMLAR (Network for Euro-Mediterranean Language Resources) lancé en 2003 visait le développement des ressources de la langue arabe dans le cadre d'une collaboration dans la région méditerranéenne. Le projet a réuni 14 partenaires de divers pays dans le cadre du programme MED-Unco soutenu par l'Union européenne [9].

Le corpus Nemlar est un ensemble de textes en langue arabe annotés par la société RDI Egypte pour le compte du Consortium NEMLAR qui détient les droits. Il contient environ 500,000 mots issus de 13 domaines différents répartis sur 489 fichiers.

Les étiquettes disponibles dans le corpus Nemlar pour un mot donné sont sa forme voyellée, son stem, les clitiques attachés au stem et sa catégorie grammaticale et son schème. Ce corpus est disponible sous deux formes : la forme voyellée et la forme non voyellée.

Afin de pouvoir utiliser ce corpus dans les phases d'apprentissage et de test de notre modèle, nous avons procédé à son enrichissement avec l'étiquette lemme en réalisant les trois étapes suivantes :

4.1. Analyse morphologique

Durant cette étape, nous commençons par analyser les mots du corpus voyellé en utilisant l'analyseur AlKhalil Morpho Sys 2. Ensuite, nous ne gardons que les lemmes dont les étiquettes lexicales associées (clitiques+stem+racine), et qui sont fournies par l'analyseur AlKhalil, coïncident avec les étiquettes lexicales du mot dans le corpus Nemlar.

4.2. Identification du lemme correct parmi les lemmes potentiels

Après avoir identifié les lemmes potentiels pour les mots du corpus Nemlar, nous avons demandé à un linguiste spécialisé d'identifier le lemme correct parmi ces lemmes. Dans le cas où le lemme correct ne figure pas parmi les sorties de la première étape, le linguiste attribue au mot son lemme.

4.3. Insertion de l'étiquette lemme

Après que le linguiste ait achevé son travail, nous sommes passés à la dernière étape qui consiste à insérer les lemmes dans le corpus Nemlar.

5. Evaluation

La phase d'apprentissage qui a servi à l'estimation des matrices de transition et d'émission a été réalisée sur 90% du corpus NEMLAR choisi aléatoirement. Des tests ont été ensuite réalisés sur deux sous-ensembles non voyellés du corpus NEMLAR :

- Le premier ensemble, appelé *Te*, constitue les 10% restants du corpus Nemlar qui n'ont pas été utilisés dans la phase d'apprentissage.

- Le deuxième ensemble, appelé *Tr*, constitue environ 25% du corpus d'apprentissage. Il a été tiré aléatoirement du corpus d'apprentissage.

La méthode d'évaluation consiste à comparer le lemme fourni par notre lemmatiseur avec celui attribué par les annotateurs de corpus. La précision est calculée par la formule suivante :

$$Précision = \frac{\text{le nombre de mots correctement lemmatisés}}{\text{la taille de l'ensemble de test}}.$$

Les résultats de test sont présentés dans la table 1.

	Précision
Ensemble *Tr*	99,21%
Ensemble *Te*	94,45%

Table 1. *Précision du lemmatiseur*

Les résultats obtenus montrent la robustesse de notre lemmatiseur. En effet, le système fourni un lemme correct dans 94.45% des mots du corpus de test *Te*, alors que ce taux augmente pour atteindre 99,21% dans l'ensemble d'apprentissage *Tr*.

Afin de situer les performances de notre lemmatiseur, nous avons réalisé une comparaison entre les taux d'erreurs de notre système et le système MADAMIRA2.

MADAMIRA (v1.0) est un système d'analyse morphologique de levée de l'ambigüité dans le contexte [7]. Il fournit plusieurs sorties morphosyntaxiques dont le lemme du mot.

Pour réaliser cette comparaison, nous avons exécuté le système MADAMIRA sur le corpus de test Te, et les résultats obtenus sont présentés dans le tableau 2.

	Précision
MADAMIRA	90,53%
Notre lemmatiseur	94,45%

Table 2. *Comparaison des précisions des deux lemmatiseurs*

Nous constatons que les performances de notre lemmatiseur sont largement meilleurs que celles de l'analyseur MADAMIRA. En effet, notre système a atteint une précision de l'ordre de 94.45% alors que celle de l'analyseur MADMIRA est en dessous de 91%.

6. Conclusion

Nous avons présenté dans cet article un lemmatiseur des phrases arabes. L'analyse morphologique opérée dans la première phase propose souvent plusieurs lemmes potentiels pour un mot donné. Pour choisir le lemme correct dans le contexte de la phrase

[2] http://innovation.columbia.edu/technologies/cu14012_arabiclanguage-disambiguation-for-naturallanguage-processing-applications

parmi ces lemmes, nous avons adopté une approche statistique basée sur des modèles de Markov cachés. Les résultats obtenus sont très encourageants. Afin d'améliorer davantage les performances du système, nous prévoyons agir sur deux niveaux :

- Niveau analyse morphologique : exploiter la richesse des informations fournies par l'analyseur Alkhalil pour mieux filtrer les transitions entre lemmes. En effet, l'absence d'une transition entre deux lemmes dans le corpus d'apprentissage n'est pas nécessairement due aux limites du corpus, mais peut être causée par le non compatibilité entre ces deux lemmes (par exemple, un lemme verbe ne peut succéder à un حرف جر).
- Niveau corpus : utiliser dans la phase d'apprentissage un corpus de taille plus importante. Cela permettra de mieux ajuster les estimations des matrices de transition et d'émission, et par suite améliorer la précision du lemmatiseur.

7. Bibliographie

[1] Boudchiche, M., Mazroui, A., Ould Abdallahi Ould Bebah, M., Lakhouaja, A., Boudlal, A., 2016. AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer. *J. King Saud Univ. - Comput. Inf. Sci*. doi:10.1016/j.jksuci.2016.05.002

[2] Hammouda, F.K., Almarimi, A.A., 2010. Heuristic Lemmatization for Arabic Texts Indexation and Classification. *J. Comput. Sci*. 6 6, 660–665.

[3] Koulali, R., Meziane, A., 2013. Experiments with arabic topic detection. *J. Theor. Appl. Inf. Technol*. 50.

[4] Manning, C.D., Schütze, H., 1999. Foundations of Statistical Natural Language Processing. *MIT Press, Cambridge, MA, USA*.

[5] Neuhoff, D., 1975. The Viterbi algorithm as an aid in text recognition. *IEEE Trans. Inf. Theory* 21, 222–226. doi:10.1109/TIT.1975.1055355

[6] Ney, H., Essen, U., 1991. On smoothing techniques for bigram-based natural language modelling, *in: 1991 International Conference on Acoustics, Speech, and Signal Processing. IEEE*, pp. 825–828 vol.2. doi:10.1109/ICASSP.1991.150464

[7] Pasha, A., Al-badrashiny, M., Diab, M., Kholy, A. El, Eskander, R., Habash, N., Pooleery, M., Rambow, O., Roth, R.M., 2014. MADAMIRA : A Fast , Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. *Proc. 9th Lang. Resour. Eval. Conf*. 1094–1101.

[8] Reqqass, M., Lakhouaja, A., Mazroui, A., Atih, I., 2015. Amelioration of the interactive dictionary of arabic language. *Int. J. Comput. Sci. Appl*. 12, 94–107.

[9] Yaseen, M., Attia, M., Maegaard, B., Choukri, K., Paulsson, N., Haamid, S., Krauwer, S., Bendahman, C., Fersøe, H., Rashwan, M., Haddad, B., Mukbel, C., Mouradi, A., Shahin, M., Chenfour, N., Ragheb, A., 2006. Building Annotated Written and Spoken Arabic LR's in NEMLAR Project, in: *LREC*. pp. 533–538.

Overview of the social information's usage in Information Retrieval and Recommendation systems

Abir Gorrab, Ferihane Kboubi and Henda Ben Ghezala

RIADI Laboratory-ENSI
University of Manouba
Manouba 2010, Tunisia
Abir.Gorrab@riadi.rnu.tn

ABSTRACT. Web 2.0 and social networks represent a huge and rewarding source of information. Our work revolves around the issue of access and identification of social information and their use in building a user profile enriched with a social dimension, and operating in a process of personalization and recommendation. We present several approaches of Social IR (Information Retrieval), distinguished by the type of social information integrated; as well as social recommendation approaches. We also expose a study of the modeling techniques of the user profile's social dimension, followed by a discussion and proposed directions.

RÉSUMÉ. Le web 2.0 et les réseaux sociaux représentent une source d'information énorme et enrichissante. Notre travail s'articule autour de la problématique d'accès et d'identification des informations sociales et leur exploitation dans la construction d'un profil utilisateur enrichi d'une dimension sociale, et son exploitation dans un processus de personnalisation et de recommandation. Nous présentons différentes approches de RI (Recherche d'Information) Sociale, distinguées par le type d'informations sociales intégrées; ainsi que des approches de recommandation sociale. Nous exposons également une étude des techniques de modélisation de la dimension sociale du profil utilisateur, suivie par une discussion et des directions envisagées.

KEYWORDS: social information retrieval, social recommendation, social networks, user profile

MOTS-CLÉS : recherche d'information sociale, recommandation sociale, réseaux sociaux, profil utilisateur.

1. Introduction

With the apparition of the social web and the explosion of social networks, users become able not only to consume, but also to product informational content. As a matter of fact, the huge number of web users and time spent daily on internet motivated researchers in IR and encouraged them to benefit from this content as an enlightening source of information. Besides, social networks and collaborative sites (such as Facebook, LinkedIn, Google+, Twitter, YouTube, delicious, CiteUlike, etc) are the most common and popular source of interactive content. In this paper, we focus on the impact of social information integration in an IR process and a recommendation system.

This paper is organized as follows: in section 2, we discuss the main approaches used in Social IR. While section 3 is devoted to the social recommendation, and section 4 is reserved to social dimension in user profile modeling, section 5 is dedicated to discussion and future directions. Finally, we conclude in section 6.

2. Social IR approaches

Social IR approaches are various. They are mostly based on social information identification and integration in a search process. In fact, there are several types of social information specific to each social network, such as folksonomies, tags, social relations (friends, co-authors, followers), comments, tweets, conversations, hashtags, like mentions, shares and many others. Proposed approaches widely use many social information, which can be integrated in different levels in IR process: user profile construction, query expansion and result ponderation. In this section, we present different social IR approaches that can be classified in three categories, according to social information used.

2.1. Approach based on annotations

Social annotations are a valuable informational source that enhances social IR by including user's area of interest**.** Bouhini et al. [2] propose a user profile generation approach from folksonomies. As a matter of fact, this work combines queries with user profile based on terms frequency. Actually, it presents two Social IR models inspired from BM25 model: BM25S Score Comb and BM25S Freq Comb, which combines query and user profile using respectively scores and terms frequency. Bao et al. [17] calculate similarity between web query and social annotations. They propose two algorithms that enhance web IR: SSR (Social Sim Rank) which computes similarity degree and SPR (Social Page Rank) that estimates web pages popularity. PengLi et al

[15] propose a TR-LDA model of annotations categorization. They introduce representation and ponderation methods of annotation categories. In this respect, they study the effect of annotations' incorporation in IR process.

2.2. Approach based on social relations

At first sight, users may be linked by different relationships that are specific to each social network, e.g., friend relationships, followers and co-authors. Works based on this approach usually use this informational content generated by relations, in a way that combines a social and a thematic score. In this context, Ben Jabeur et al. [9] investigate on a social model based on Bayesian network, incorporating two social relevance factors: User social importance, evaluated by a PageRank score; and the number of temporal neighbors. Moreover, Amer et al. [14] propose a probabilistic model of conversation indexation in twitter. This model incorporates social relations to measure users' influence, activity and expertise.

2.3. Approach based on social signals

Social signals like comments, shares and like mentions are being more explored in social IR works, due to the significant information they bring. Chelaru et al [16] study the impact of these social signals in video search on YouTube, by combining social information such as comments, like and dislike mentions, with basic search criteria (similarity between the query and video title). Hence, this unification enhances the performance of videos extraction process. Furthermore, Badache et al. [5] describe a language model exploiting temporal characteristics of social signals (number of like mentions, shares and comments) to estimate resources relevance and sort search results. Moreover, Ramesh et al. [13] examine the personalized social IR process and suggest an algorithm of user profile construction using pages liked on Facebook, through different user's accounts. This social content personalizes search results.

2.4. Comparative study

For more information about Social IR approaches, we led a comparative study of different categories. For each work we considered the following six points as a comparative criteria: (1) the social network used for the experimentation, (2) the techniques used in the presented models, (3) the metrics used for the evaluation, (4) if there is a combination of information, (5) if there is a combination of social networks and (6) if the work considered the temporal aspect. Table 1 summarizes the results of our study.

Social information		Social network	Techniques	Evaluation	Combining Information	Combining social networks	Temporal aspect
Annotati-ons	[2]		BM25and derivatives	MAP, P[0,1]	-	-	-
	[17]	Del.ic.ious	SSR, SPR	MAP, nDCG	-	-	-
	[15]	Del.ic.ious	TF_IDF, Inference algorithm	-	-	-	-
Social relations	[9]	Twitter, Citulike	PageRank, language model, ImpG:social score, TF-IDF	MAP, recall	√	√	√
	[14]	Twitter	BM25, language model, PageRank	Leave One Out approach, MAP	√	-	-
Social signals	[16]	Youtube	TF_IDF, Lucene, SentiWord-Net	nDCG	√	-	-
	[5]	Facebook, Twitter, LinkedIn, Del.ic.ious, Google+	Language model	MAP, nDCG, Recall, Precision	√	√	√
	[7]	Facebook	Clustering data TF-IDF	Performance measure	√	-	-

Table1. Comparative table of Social IR approaches categorized by social information types

In table1, we present some works related to the three approaches described in this section, based on annotations, social relations and social signals. Characteristics studied are the combination of many social information or social networks and the consideration of temporal aspect. These features enhance IR processes and improve their performances. In fact, many networks are used and many techniques are conducted, but

temporal aspect and the combination of different networks represent the greatest motivation for researchers.

3. Social recommendation

Social recommendation is a set of methods that try to suggest items or entities that seem to be interesting to the user, using his social information [12]. In fact, there are two main recommendation techniques. The first one is a content based approach which is based on recommending items similar to those the user has chosen in the past. The second one is a collaborative filtering approach; this approach recommends items to the user based on the choice of other people, who seem to have similar preferences. Moreover, Hybrid recommendation is a technique that comprises both content-based and collaborative filtering approaches, so as to provide the user with better recommendations. Additionally, many researchers have explored social information to improve recommender systems. Notably, Hafsi et al. [11] exploit user-generated content (rating and review) in books recommendation system. Their work measures books reputation and popularity concepts and tests three approaches: book tags and reviews indexation, themes interrogation and users similarity calculation. Unlike in [6], authors have proposed a content-based approach that compares users profiles' information in order to determine similarities between them and recommend friendship relations. On the other hand, Wang et al. [18] investigate on tag based social recommendation by calculating tags similarities and connecting users that are likely to have similar tastes and preferences. In the same context, Hannon et al. [7] propose an hybrid recommendation system using content and collaborative-based approaches that recommends users to follow in the social network Twitter, by analyzing their profiles.

4. Social dimension in user profile modeling

User profile modeling is an essential task in Personalized IR. This entity brings and organizes the information necessary to define the user and describe his interests.

Following the emergence of social networks, Social IR has widely evolved. Thus, the social dimension of the user profile has become an essential component in social personalization systems. A lot of works were directed towards the construction of a social profile based on annotations [8] [4], given the importance of the data they generate. Others have focused on the analysis of egocentric social network, they are interested in friendship relationships in social networks [3] [10]. This information produces relevant content for collaboration within social IR systems. It solves the cold start problem, or lack of user's activity on social networks. The temporal aspect is also

reflected in some works [10], which differentiates between recent and old social activities, to estimate their importance. Other social signals have also been integrated into the social dimension of user profile such as comments and shares. Once the profiles are built, some authors have thought of building virtual communities of users, based on similarity degree between the profiles. These communities are considered as a dimension in the profile. They are very rewarding and provide additional relevant information. In [1], Dridi et al model a user profile based on annotations and exploit it to detect communities based on annotations similarities. For community detection, Katz index is used. It calculates the similarity taking into account the direct and indirect links in a graph.

5. Discussion and future directions

In this section, we discuss different aspects related to research in Social IR topic. In fact, IR classical approaches do not take into account the user's social content provided by his interactions and social relations. Moreover, most of the existing approaches in Social IR use either social signals, tags or relational information. Some works started leveraging different types of information. Also, combining social content from many social networks and matching different user's social profiles improve the collection of relevant information that better describe the user and enhance his affluence.

The construction of a data collection relative to SIR systems is basically a major challenge. For this issue, we led a technical study of a set of social networks API that are likely to be the most known. networks don't present yet API for developers, like ResearchGate. Some social In the extraction process, the majority of social networks use the OAuth 2.0 for connection and authentication authority, like Twitter, Youtube, Google+, LinkedIn, and Foursquare. Delicious and CiteUlike require basic http authentication, while some other networks need API keys for authentication (Last.fm, Flickr). REST is the common API used to have access to resources, and the result is always a JSON or XML file. Actually, this study is our way to construct a data collection suitable for Social IR.

Temporality is a fundamental issue and the most central aspect in social content. This factor is being investigated in several works [10][6][11]but still presents new contribution areas. Temporal aspect supports the eventual and permanent evolution of users' tastes, preferences and behaviors. Indeed, information appreciated by users now may not remain the same after a moment. Besides, trend events attract users' attention at a specific moment and are no more important after a while. Thus, Social IR systems should be adapted to this evolution. The same as for Social IR systems, the freshness of the information is essential in recommendation systems. So, to enhance recommendation quality, temporal factor should be considered.

Social approaches present certainly some limitations. A big challenge is to map user's accounts across social networks [19], and to predict missing social information, by combining for example social information and the social graph [20], in order to have an enriched social user profile. We will make a deep study in works dealing with these challenges.

6. Conclusion

In this paper, a review of different aspects of Social IR is proposed. We presented a classification of Social IR approaches into three main categories, based on social information used. We also posed a study of Social recommendation systems. Then, we referred to user profile models proposed in Social IR studies, and specially the social dimension. In this respect, works included in this review reflect how deep the impact of social content in IR and recommendation process is. Furthermore, we discuss different aspects of Social IR. As for coming studies, we start the process of user profile construction, based on temporal social signals.

7. References

[1] A. Dridi and M. Kacimi. "Information Retrieval Framework based on Social Document Profile". CAiSE(Forum/DoctoralConsortium), 2014.

[2] C. Bouhini, M. Géry and C. Largeron. "Integrating user's profile in the query model for Social Information Retrieval". Eighth International Conference on Research Challenges in Information Science (RCIS) ,1-2. IEEE, 2014.

[3] D. Tchuente, M.F. Canut, N. Jessel, A. Péninou and F. Sèdes. "A community-based algorithm for deriving users' profiles from egocentrics networks: experiment on Facebook and DBLP". Social Network Analysis and Mining, 3(3), 667-683, 2013.

[4] H. Xie, X. Li, T. Wang, L. Chen, K. Li, F. L. Wang, Y. Cai, Q. Li and H. Min. "Personalized Search for Social Media via Dominating Verbal Context." Neurocomputing, 172, 27-37, 2016.

[5] I. Badache and M. Boughanem. "Document Priors Based On Time-Sensitive Social Signals." ECIR, 617-622. Springer,2015.

[6] J. Chen, W. Geyer, C. Dugan, M. Muller and I. Guy. "Make new friends, but keep the old: recommending people on social networking sites". InProceedings of the SIGCHI Conference on Human Factors in Computing Systems, 201-210, ACM, 2009.

[7] J. Hannon, M. Bennett and B. Smyth. "Recommending twitter users to follow using content and collaborative filtering approaches". In Proceedings of the fourth ACM conference on Recommender systems, 199-206, ACM, 2010.

[8] K. Han, J. Park and M. Y. Yi."Adaptive and multiple interest-aware user profiles for personalized search in folksonomy: A simple but effective graph-based profiling model". In 2015 International Conference on Big Data and Smart Computing (BIGCOMP), 225-231, IEEE, 2015.
[9] L. Ben Jabeur, L.T. Lechani and M.Boughanem. "Intrégration des facteurs temps et autorité sociale dans un modèle bayésien de recherche de tweets". CORIA, 301-316, 2012.
[10] M.F Canut, S. On-at, A. Péninou and F. Sèdes. "Time-aware Egocentric network-based User Profiling". In 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 569-572, IEEE, 2015.
[11] M. Hafsi, M. Géry and M. Beigbeder. "LaHC at INEX 2014: Social Book Search Track". Working Notes for CLEF 2014 Conference, 2014.
[12] M. R. Bouadjenek, H. Hacid and M. Bouzeghoub. "Social networks and information retrieval, how are they converging? A survey, a taxonomy and an analysis of social information retrieval approaches and platforms". Information Systems, 56, 1-18, 2016.
[13] N. Ramesh and J. Andrews. "Personalized Search Engine using Social Networking Activity". Indian Journal of Science and Technology, 8(4), 301, 2015.
[14] N. Ould Amer, P. Mulhem and M. Géry. "Recherche de conversations dans les réseaux sociaux: modélisation et expérimentations sur Twitter". Conférence en Recherche d'Infomations et Applications-12th French Information Retrieval Conference, 2015.
[15] P. Li, B. Wang, W. Jin, J. Nie, S; Zhiwei, Z. and B; He. "Exploring categorization property of social annotations for information retrieval". In Proceedings of the 20th ACM international conference on Information and knowledge management, 557-562, ACM, 2011.
[16] S. Chelaru, C. Orellana-Rodriguez and I. Sengor Altingovde. "Can social features help learning to rank youtube videos?". International Conference on Web Information Systems Engineering, 552-566, Springer Berlin Heidelberg, 2012.
[17] S. Bao, X. Wu, B. Fei , G. Xue , Z. Su and Y. Yu . "Optimizing web search using social annotations". In Proceedings of the 16th international conference on World Wide Web, 501-510, ACM, 2007.
[18] X. Wang, H. Liu and W. Fan. "Connecting users with similar interests via tag network inference". In Proceedings of the 20th ACM international conference on Information and knowledge management, 1019-1024, ACM, 2011.
[19] E. Raad, R. Chbeir and A. Dipanda."User profile matching in social networks". In 13th International Conference on, 297-304, IEEE, 2010.
[20] A. Mislove, B. Viswanath, P. Krishna Gummadi and P. Druschel "You are who you know: inferring user profiles in online social networks". In Proceedings of the Third ACM International Conference on Web Search and Web Data
Mining, 251-260, 2010.

Vers un système iconique d'aide à la décision pour les praticiens de la médecine traditionnelle

KOUAME Appoh[12], BROU Konan Marcellin[1], LO Moussa[2], LAMY Jean Baptiste[3].

[1](Institut National Polytechnique Félix Houphouët Boigny de Yamoussoukro) (Côte d'Ivoire) (kgerappoh@gmail.com, konanmarcellin@yahoo.fr)

[2](Université Gaston Berger) (Sénégal) (moussa.Lo@ugb.edu.sn)

[3](Université Paris 13, Bobigny (France)(jean-Baptiste.Lamy@Univ-Paris13.Fr)

RÉSUMÉ

«Mise en place d'une plateforme web social et sémantique pour le partage des connaissances des communautés ouest africaines » est un projet dans lequel un volet important concerne les praticiens de la médecine traditionnelle(PMT). Disposer d'un outil collaboratif de travail qui transcende leur état prégnant de non lettré constitue un défi majeur. C'est ce à quoi s'attelle ici, notre réflexion en amorçant l'aspect visuel via une composition iconique à approche ontologique d'ontoMEDTRAD. L'ontoMEDTRAD, avec ses composantes ontoConcept_term et ontoIcone, fait partie du système de gestion de la médecine traditionnelle, sysMEDTRAD. ontoConcept_term dénote les termes des concepts du domaine. ontoIcone devra inclure les icônes en alignement avec ces termes.

ABSTRACT

«Setting up a social and semantic web framework for knowledges sharing in West African communauties» is a project in which we note an important section relating to traditional medicine practitioner (TMP). People TMP, for the most part, are illiterate. Make available, a tool which transcends this characterization, constitutes a scientific issue in this work. Thus, the visual aspect via an iconic composition based on ontological approach of ontoMEDTRAD, is our goal. ontoMEDTRAD includes two modules which are ontoConcept_term and ontoIcone. It is also a part of the management system for traditional medicine (sysMEDTRAD). ontoConcept_term denotes the terms of concepts in this domain. ontoIcone will include the set of icons corresponding or in alignment with those terms.

MOTS-CLÉS : ontologie, web sémantique, Médecine traditionnelle, composition iconique, Afrique de l'Ouest.

KEYWORDS: ontology, semantic web, traditional medicine, iconic composition, West Africa.

1. Introduction

En médecine africaine, chaque habitant en zone rurale connaît et utilise les vertus d'un certain nombre de plantes. Ceci relève de la pharmacopée populaire. Ces usages en plus d'être culturels, lient les habitants à leurs terres [2]. Parmi les ressources médicinales traditionnelles, les plantes sont les plus utilisées. Les étalages des tradipraticiens en Afrique abondent de plantes. Ici, notre contribution vise, à augmenter le potentiel des soins en santé primaires pour les populations notamment de l'Afrique de l'ouest par la médecine traditionnelle(MT), persistante avant et depuis l'antiquité jusqu'à nos jours. L'ambition portée sur la MT d'en faire un système complémentaire de santé ne la rend pas substitut des offres de soins issues de la médecine moderne (MM). Pour les populations, l'accès aux soins de santé de la MM n'est pas équitable. Parallèlement à cette médecine conventionnelle, 80% des populations de l'Afrique de l'ouest, font appel à la MT locale [21]. A ce titre, sauver tout un réutilisable patrimoine de connaissances et d'expériences menacé de disparition, est d'un grand intérêt. Bien entendu, la transmission des savoirs en MT est amplement orale, puis la gestion des faunes et flores, empirique et à tâtons. Il importe alors de disposer d'un cadre idéal pour les praticiens de la MT (PMT), afin de partager, mutualiser, co-construire, sauvegarder et pérenniser les connaissances, les acquis et expériences dans ce domaine. Ici, notre SysMEDTRAD [17] conçu et s'adossant à un wiki sémantique, répond à cette exigence. Sa composante principale est ontoMedtrad, elle-même structurée en ontoCONCEPT_term et ontoIcone. La majorité des PMT étant illettrée, l'aspect visuel leur permet de s'affranchir des barrières linguistiques, de la lecture et de l'écriture textuelles. En général, une fois la maladie est déterminée chez un patient, le PMT pense à la plante d'où il tirera la recette des soins. Nous entamons l'aspect iconique des plantes antipaludiques, suite à la modélisation conceptuelle (annexe 2) dont le modèle de connaissances obtenu a permis de frmaliser l'ontoCONCEPT_term. Quatre points structurent ce document : (a) spécificités de la MT ; (b) approche iconique; (c) amorçage d'un système iconique ; (d) travaux liés.

2. Traits caractéristiques de la MT

Lors du recueil de données, on s'aperçoit que des PMT de grande renommée voilent à peine leur détermination de mourir avec leur secret [2]. D'autres PMT veulent une transmission des savoirs par lignée. Entre PMT, le manque d'acceptation et de partage mutuels des connaissances est avéré. Vu le caractère fortement implicite et caché de la MT, ont été sensibilisés au moins cinquante (50) PMT pour leur adhésion au projet. Le directeur du PNPMT[1] a eu un rôle de facilitation. En Côte d'Ivoire, depuis 2014 jusqu'à ce jour, à travers des ateliers et séminaires, ces PMT sont formés, afin de les amener au respect d'un certain nombre de normes de pratique et d'éthique dans leur art. Leur

1 PNPMT : Programme National de la Promotion de la Médecine traditionnelle, un des démembrements du Ministère ivoirien de la santé et de la lutte contre le sida(MSLS)

capacité en anatomie humaine a été renforcée par un médecin (MM), au premier trimestre de 2015. Un PMT, pour être enregistré au PNPMT, doit appartenir à une association de PMT. Une seule fédération des tradipraticiens de santé et naturothérapeutes (FTSN-CI) regroupe toutes les associations. Le désordre orchestré en leur sein, susceptible de perte de vies de patients, ne doit plus se perpétuer. La vision globale des gouvernants de la sous-région d'améliorer la MT et de la valoriser, est bien nette. En partie, la MT constitue un levier pour des découvertes en MM [4], où la majorité des médicaments obtenus sont des produits de synthèse suite à une chaîne de production (biologie, principes actifs des plantes médicinales, adjonction d'adjuvants et d'excipients). Entre MM et MT, les modes opératoires sont différents. Une autre spécificité de la MT est l'exercice inclusif par le PMT de deux fonctions, à savoir «médecin prescripteur» et « pharmacien ». En MT, le mode de prise en charge d'un patient n'est pas celui de la MM ; Dans le processus de traitement d'un malade, le PMT commence par détecter les signes symptomatiques sur le patient afin de déterminer la maladie. A ce type de diagnostic, se rajoute un autre d'ordre métaphysique, en plus de certains déterminants socio-culturels et environnementaux [11]. L'administration de soins par le PMT est en partie sous forme de remèdes appropriés, de fois, de manière extemporanée au regard de la première des quatre catégories de médicaments traditionnels selon l'OMS [5]. L'heure, la période et la saison de collecte de certaines plantes peuvent influer leurs vertus thérapeutiques. L'échange entre patient et PMT est bidirectionnel, se démarquant nettement de la position très dominante du rôle de médecin moderne [3] dans le même contexte. En somme, la MT vise un traitement exhaustif du patient (corps, âme et l'esprit ; environnement social et culturel), autrement dit du bien-être intégral du patient [11]. Ce qui précède rend impossible, ici, la réutilisation diligente et automatique des ressources terminologiques et ontologiques (RTO) existantes en MM.

3. Approche d'ontologie visuelle

Une ontologie est une spécification formelle et explicite d'une conceptualisation [25]. Ce travail d'«iconisation» est mené sur la base des modèles de données conçus en UML (annexe 2) puis formalisés (axiomatisation) sous protégé. Vu la complexité visuelle, pour la construction iconique, nous avons procédé par étape (en spirale et par incrément). Nous nous focalisons sur les plantes, ressources médicinales les plus prescrites par les PMT. Nous avons alors utilisé les éléments sémiologiques relatifs à un triptyque de stratégies de représentation [7]. Aussi avons-nous tenu compte du contexte d'utilisation de l'outil final dont disposeront les PMT. L'usage d'outils TIC (mobile, ordinateur...) comme moyen d'accès à SysMETRAD, milite fortement en faveur de l'utilisation de telles approches. Il est démontré la supériorité de la communication pictographique sur la communication verbale dans le dialogue homme-ordinateur [12][27].

3.1. Justification, collecte de données et principe de base

En MT africaine, les ressources médicinales sont de différentes origines : végétale, animale, minérale et métaphysique. La proportion des remèdes à base de plantes est plus

forte. Plus de cinq cent (500) plantes médicinales existent en Afrique de l'ouest selon l'UNESCO [26]. Notre approche d'ontologie visuelle nécessite qu'en plus de modéliser les concepts de ce domaine (ontoCONCEPT_term), il faut les représenter graphiquement par des icônes (ontoIcone). Idéalement, une correspondance systématique entre ces deux modèles devra être établie. Pour amorcer le volet iconique, nous partons du point de vue d'intérêt majeur « donner des soins de santé primaires à un patient ». Cette exigence est satisfaite par la définition de trois cas d'utilisations notés UC_i avec i=1 à 3 ; UC_1 : déterminer la maladie à partir des symptômes (faire un diagnostic) ; UC_2 : déterminer le remède (définir la recette via les parties des plantes et les plantes, le mode de fabrication et la forme de fin de préparation); UC_3 : déterminer le mode d'administration de la recette. En termes de scénario, notre choix s'est porté sur le paludisme, maladie classée parmi les plus mortelles et morbides en Afrique occidentale. Aussi, avons-nous axé notre réflexion prioritairement sur l'UC_2 consistant à spécifier la recette à base de plantes antipaludiques dont un inventaire est nécessaire. Pour définir une recette, le PMT commence par choisir la plante. Ensuite, il détermine les parties de plante qui seront utilisées dans cette recette. Suite à ce constat, il apparaît nécessaire de mettre au point une approche graphique pour présenter visuellement les plantes médicinales et permettre leur reconnaissance. Dans le florilège de plantes médicinales antipaludiques collectées en pharmacopée traditionnelle, nous avons ciblé vingt-deux (22) [22] [15] [14] [1] [23] pour trente (30) recettes (annexe 8). Autant pour ce recueil que pour la modélisation conceptuelle, nous avons parcouru moult sources et résultats d'études ethnobotaniques en biosciences comme celles précédemment citées. Notre démarche inclut les retombées en termes de connaissances reçues suite à notre collaboration directe avec des PMT (visites de terrain). Egalement avions-nous été instruis d'importants travaux de référence sur les pharmacopées traditionnelles [14] [19], des séminaires auxquels nous avons participé, et des documents du PNPMT et des ONG comme la PROMETRA. Une plante médicinale soigne une ou plusieurs maladies. Une recette peut être monospécifique (à base d'une plante), ou multispécifique (à base de plus d'une plante). L'association de plantes, mal assortie, est dangereuse [16] et donc déconseillée sauf en cas de maîtrise suffisante des effets secondaires. En médecine populaire et rurale, des associations de plantes sont assez connues pour purge (ou lavement) contre des maux de ventre. Les composantes d'une recette sont les parties de plantes à savoir : feuille, fruit, écorce, tige, racine, fleur,…, ou plante entière (herbe). Parmi elles, sont plus fréquentes, les feuilles, les fruits, les écorces. Majoritairement, les feuilles et les fruits sont sollicités aux taux respectifs de 60% et 15% [16]. Des critères sont définis pour caractériser les 22 espèces et permettre de les distinguer individuellement. À partir de la représentation visuelle et iconique d'une plante, il est aisé pour le PMT de faire sa prescription de recette. La description botanique au plan morphologique, devient alors intéressante et importante pour la retenue des critères. La quarantaine de critères extraits, sont réduits progressivement, par le truchement d'une méthode inclusive d'un logiciel d'apprentissage Weka, d'éléments sémiologiques, du contexte de travail projeté du PMT. Chaque valeur ou instance prise par un critère fait l'objet d'un iconème, unité signifiante d'icône. Ces iconèmes sont ensuite assemblés pour former les icônes. Il apparaît alors un problème d'optimisation

consistant à minimiser à la fois le nombre d'iconèmes et celui des icônes pour une meilleure représentation visuelle de toutes les espèces de plantes, tout en restant capable de les distinguer entre elles. En botanique, la systématique des végétaux suivant la hiérarchie «monde du vivant », domaine, règne, embranchement, classe, ordre, famille, genre, espèce, qui est aussi celle des plantes médicinales, est plus stable. Il est difficile pour la majorité des PMT de s'en prévaloir au regard de l'objectif ici visé. Au mieux, la priméité visuelle doit demeurer du fait du caractère prégnant d'illettré des PMT.

3.2. Variabilité descriptive de la botanique physique : les plantes (végétaux)

Le parcours descriptif par les botanistes des plantes est souvent variant. Cela est à même d'engendrer des divergences et contradictions. Les deux espèces *Adansonia digitata L et Lannea microcarpa Engl. et K. Krause* font partie de notre sélection de plantes médicinales respectivement notées (*ad)* et (*lm)*. (*lm)* est à feuille composée alterne et imparipennée avec 15 m de haut dans [24], tandis qu'elle est décrite comme étant à feuille opposée, de haut 10 m selon [9]. Notons que la morphologie d'une espèce végétale est diversement appréciable selon la région d'habitat, le pays, le climat et les saisons. La morphologie de (*ad),* plus élancée en Côte d'Ivoire, est différente de celle plus robuste et imposant au Sénégal, avec plus ou moins le même goût de fruit. Ceci montre bien les limites et insuffisances d'utiliser des photographies réalistes comme moyen sublime de reconnaissance visuelle des plantes. Les dimensions morphologiques d'une plante dans une photographie peuvent s'avérer trompeuses et laisser ainsi apparaître un fossé de compréhension dans le réel. Le fait de juxtaposer un objet physique connu aux côtés d'un autre méconnu dans une même image photo, permet d'avoir une vue plus rapprochée des dimensions réelles de l'objet méconnu. Par ailleurs, selon la FAO [10], un arbre a 5 m de hauteur minimale, là où l'Institut Forestier National (IFN) de France, donne 7 [13]. L'annexe 3 montre deux classifications par strate des végétaux faites par deux botanistes et laissant éclore une différence. Toute cette variabilité est à restreindre en utilisant des éléments visuels, semi-formels ou formels pour atteindre l'«iconisation » plus intelligible par la machine.

3.3. Technique de réduction et de choix des critères

La technique de réduction et de choix des critères englobe des règles sémiologiques et contextuelles. A l'aide des algorithmes Ranker et Jrip de Weka, nous avons procédé à supprimer progressivement les critères en prenant en compte leur poids de distinguabilité, les aspects visuels liés au PMT et les règles sémiotiques tirées surtout de la pensée peircienne, quand bien même celle de F. Saussure (diadique) ne nous a pas fait défaut pour comprendre les signes. Le signe, selon Peirce S, a trois dimensions : son objet, son icône et son interprétant. F. Saussure en a une perception double : signifiant et signifié. Ainsi, similarité visuelle, association sémantique et convention arbitraire [7] fondent le triptyque de stratégies pour notre représentation iconique des individus de plantes. Les critères retenus en termes de propriétés sur ces individus permettent de mieux les discriminer. Ranker classe par ordre décroissant les critères selon le poids de distinguabilité (voir tableau 1). JRip, sur la base des critères retenus parmi ceux précédemment classés, distingue de la meilleure manière (100%), les individus de plantes

6

par un raisonnement logique. Nous donnons un aperçu d'une partie des associations de classes (annexe 1). Relativement au paludisme, il y a des symptômes constants, et d'autres spécifiques à l'état du patient, soit enfant, femme en grossesse ou adulte. On admet que le paludisme est manifeste sur un patient. Les vingt-deux (22) plantes sont représentées à travers sept (7) critères selon le tableau ci-après (annexe 4).

Tableau 1 : critères visuels retenus

Poids du critère	N'ordre	Critère
2.845	15	formeDeFruit
2.799	16	couleurDefruit
2.215	18	couleurFleur
2.197	6	formeFoliaireFeuille
2.117	1	typeSilhouette
1.842	4	DispositionFeuille
1.529	8	pennation
		nom botanique (dont les instances ou individus sont distinctement discriminés)

4. Amorçage d'un système iconique

Au regard ce qui précède, pour le besoin iconique, on a admis cinq (5) silhouettes que sont **arbre**, **arbuste**, **palme**, **herbe** et **liane** pour représenter l'ensemble des instances de la classe Plante via *aPourSilhouette* (Tableau 2).

Tableau 2 : des associations entre classes (objects properties)

Domaine	Object Properties	Range	*fonctionnel* : *oui/non*
plante	*aPourCouleurDeFleur*	Couleur	non
plante	*aPourCouleurDeFruit*	Couleur	oui
plante	*aPourDispositionDeFeuille*	DispositionDeFeuille	oui
plante	*aPourFormeDeFeuille*	FormeDeFeuille	oui
plante	*aPourFormeDeFruit*	FormeDeFruit	oui
plante	*aPourPennation*	Pennation	non
plante	*aPourSilhouette*	Silhouette	oui

Tableau 3 : quelques attributs de classe (data properties)

Domaine	**Data Properties**	Range
Plante	*aPourNomBotanique*	string
Couleur	*aPourCodeHtml*	string
Plante	*aPourSynonyme*	string
Plante	*aPourNomVernaculaire*	string

Tous les iconèmes sont construits via le logiciel vectoriel **inkscape.** En conséquence, un iconème a été conçu pour chacune de ces 5 entités de Silhouette (Tableau 4).

Tableau 4 : iconèmes des cinq silhouettes retenues

arbre	arbuste	palme	liane	herbe
				

Fig .1 association d'iconèmes pour l'icône de Azadirachta indica

Les 22 espèces de plantes sont alors reparties entre ces silhouettes. Nous avons conçu des icônèmes (génériques et spécifiques) pour les parties de plantes concernées par les 7 critères retenus. L'Owlready-2.0 est une librairie qui a été intégrée dans python [18]. Ceci a l'effet de rendre accessible ontoMEDTRAD au programme python (annexe 6), dont l'exécution permet d'obtenir les icônes par composition d'iconèmes (fig.1 et tableau 5). Un exemple de résultat est l'icône de la plante « *Azadirachta indica »,* obtenu à partir d'une silhouette *arbre* et d'un fruit à forme *elliptique*. Les 33 iconèmes retenus sont décrits en appendice annexe 5. On admet la couleur verte pour les feuilles. La « pennation », la disposition de la feuille sont à iconèmes de couleur noire, y comprise la silhouette. Les fruits et les fleurs peuvent avoir des couleurs d'apparence « réelle ». La fleur pourrait changer de couleur une seule fois. La plante est reconnaissable à travers ses traits très caractéristiques au plan visuel.

7

Tableau 5 : de critères retenus et icône de Azadirachta indica A. Juss (neem)

Nom_botanique	*Silhouette*	*formeDeFruit*	*Fruit couleur*	*fleur couleur*	*FormeFoliaireFeuille*	*Penantion1*	*dispositionDe Feuille*
Azadirachta indica A. Juss.(Meliaceae)	arbre	elliptique	vert	blanc, jaune	entiereLanceolee	imparipennee, paripennee	opposee
pictogramme ou icône							

5. Travaux liés

Des travaux connexes à celui-ci sont de deux ordres (1) et (2) : ***-(1)*** *au titre de la MT* : Armel [6] a réalisé une ontologie sur la MT au Cameroun sans approche visuelle. [11] An Ontology for African Traditional Medicine (ATM) de G. ATEMEZING a son objet focalisé sur la validation des connaissances de la MT gérée par un système d'agents. Il n'aborde pas le fait que les PMT sont non lettrés. N. C. KUICHEU aborde une ontologie IcOnto en définissant l'icône b d'un concept X via X (a, b) [20], a étant le terme de X. Le langage utilisé est limité en définition. Dans les ontologies utilisées la description des plantes, principales ressources des recettes médicales traditionnelles, est très réduite. ***-(2)*** *au titre de la MM* : On a des thésauri, des taxinomies ou des ontologies pour la plupart des travaux réalisés. Cependant, ils ne comprennent pas d'ontologie visuelle sauf le Projet « VCM » de Lamy, où on a un langage iconique pour l'accès aux connaissances sur le médicament et le guide de bonnes pratiques cliniques. Une ontologie visuelle est validée sur la MM focalisée sur les aspects de facilité d'apprentissage et de vitesse de lecture [18] du praticien.

6. Conclusion

L'icône est devenue une réalité pour la plante médicinale, ressource la plus importante dans les recettes prescrites en MT africaine. Les descriptions botaniques et visuelles, afférentes à la même espèce de plante sont des plus variées. Il n'en demeure pas moins du discours des PMT sur la même recette traitant la même maladie. La difficulté liée à cette varibilité descriptive est surmontable. La composition sémi-formalisée, formalisée et schématisée est une tâche ardue qui doit s'appuyer sur un minimum de règles, en vue de son intégration dans un wiki sémantique, puis permettre d'autres catégorisations et extractions de connaissances. Ce système iconique (ontoIcone), amorcé sur les plantes devra servir d'outil d'assistance décisionnel pour les PMT à même de transcender la lecture et l'écriture textuelles voire les barrières linguistiques. Le processus de validation des icônes est à deux étapes : la première par les PMT lettrés, et la seconde pour l'ensemble. Ne serait-il pas judicieux d'établir de simple liaison entre icônes et images photos réalistes afin de pallier à l'insuffisance d'instruction pour des PMT, le temps de la standardisation. En perspective au terme de cette composition de plante, nous devons asseoir un véritable langage iconique à but inférentiel pour les PMT.

8

7. Bibliographie

[1]ADOU Lydie et al, 2014, Nephrolepis biserrata, une Ptéridophyte utilisée comme plante médicinale en Côte d'Ivoire, pp1-9
[2]AKE Assi L., 2011. Abrégé de Médecine et pharmacopée africaines, CI, pp1-150.
[3].A. KONAN, 2012, Place de la médecine traditionnelle dans les soins de santé primaires, (Côte d'Ivoire), Université Toulouse III-Paul Sabatier, pp54-118
[4]Albert Chominot, 2000, valorisation des plantes médicinales par l'industrie pharmaceutique complémentarités et contradictions, Courrier de l'environnement de l'INRA n°39, pp7-8
[5]AMARI A. et al, 2006, Adaptabilité des conditions d'enregistrement des médicaments génériques en RCI aux médicaments traditionnels améliorés, J. sci., Lab. de galénique et législation pharma., UFR pharma. et bio, Univ de Cocody, pp1-4
[6]ARMEL A. et al, 2012, Using METHONTOLOGY to Build a Deep Ontology for African Traditional Medicine: First Steps. Départ. Génie Informatique, IUT de Douala, Cameroun pp1-8
[7].Carlos N. et al, 2013, A Taxonomy of Representation Strategies in Iconic Communication, Department of Biomedical Informatics, University of Utah, USA pp1-20
[8] DOZON J.-P., 1988, Ce que valoriser la médecine traditionnelle veut dire, Politique africaine, n° 28, pp1-12
[9] Espèces arbustives, commentaire botanique http://www.bamanan.org/ (04/16)
[10] FAO, Archives de documents, http://www.fao.org/docrep/008/ae578f/AE578F05.htm (04/16)
[11]. G. ATEMEZING et al, 2009, An Ontology for African Traditional Medicine, pp1-10
[12].Guastello, et al., 1989, "Verbal versus pictorial representation of objects human computer interface," International Journal of Man-Machine Studies, Vol. 31, (1989), pp 99-120.
[13] IFN, https://fr.wikipedia.org/wiki/Arbre (04/16)
[14] J. Kerharo et al, 1974, pharmacopée traditionnelle sénégalaise : plantes médicinales et toxique
[15].KOFFI N. et al, 2009, Screening phytochimique de quelques plantes médicinales ivoiriennes utilisées en pays Krobou, UFR Biosciences, Labo. de Botanique., université de Cocody pp5-15
[16] KOFFI N. et al, 2008, Étude ethnopharmacologique de plantes antipaludiques utilisées en médecine traditionnelle chez les Abbey et Krobou d'Agboville (Côte d'Ivoire) pp1-9
[17] KOUAME Appoh et al, 2014, Architecture d'un système de gestion des connaissances de la médecine traditionnelle : sysMEDTRAD, INP-HB LARIMA, UGB LANI, CARI pp1-12
[18] LAMY JB et al, 2014, Validation de la sémantique d'un langage iconique médical à l'aide d'une ontologie : méthodes et applications (OWLready) ic2014, pp1-12
[19] Michel Arbonnier, 2002, Arbres-arbustes-et-lianes-de-zones sèche de l'Afrique de l'Ouest pp1-579
[20] N. C. KUICHEU et al, 2012, Description Logic Based Icons Semantics: An Ontology for Icons, school computer of Beijing Jiaotong University, China, pp1-4
[21]OMS/WHO, 2013, Stratégie de l'OMS pour la Médecine traditionnelle pour 2014-2023, pp1-25
[22] P. Zerbo et al, 2011, Plantes médicinales et pratiques médicales au Burkina Faso : cas des Sanan pp1-13
[23] Piba S. C. et al, 2015, Inventaire et disponibilité des plantes médicinales dans la forêt classée de Yapo-abbé, en Côte d'Ivoire, pp1-21
[24] PROTA (Plant *Resources* of Tropical Africa)
http://uses.plantnet-project.org/fr/Lannea_microcarpa (PROTA) (04/16)
[25] Tom GRUBER, 1993, A Translation Approach to Portable Ontology Specifications pp1-27
[26]Unesco, http://www.unesco.org/africa/VF/pages/afrique/2b.html) (04/16)
[27] Whiteside J., Jones S., Levy P.S., Wixon D., 1985, User performance with command, menu, and iconic interfaces; CHFCS, 185-191.

9

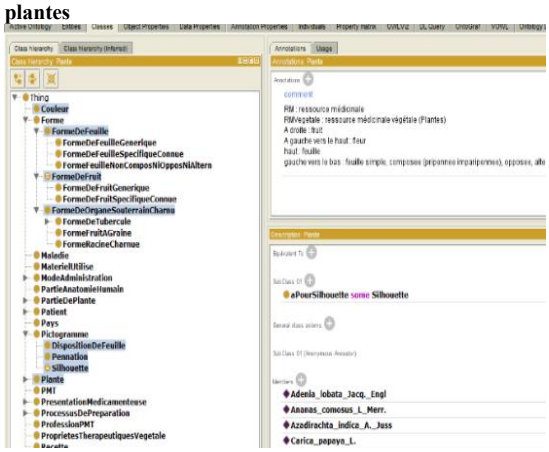
8. Appendices

Annexe 1

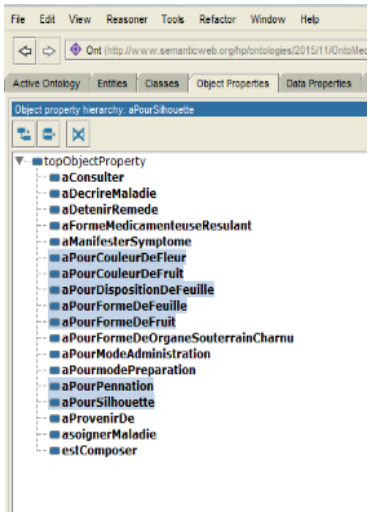
Annexe 1a : association de classes

Domaine	Object_Properties	Range
PMT	consulter	Patient
Patient	manifester	Symptôme
Symptôme	décrire	Maladie
Recette	soigner	Maladie
Recette	composer	Plante
PartieDePlante	provenirde	Plante
PMT	détenir	Recette

Annexe 1b : apercçu des classes selectionnées et utilisées pour la construction de pictogramme ou icone des plantes

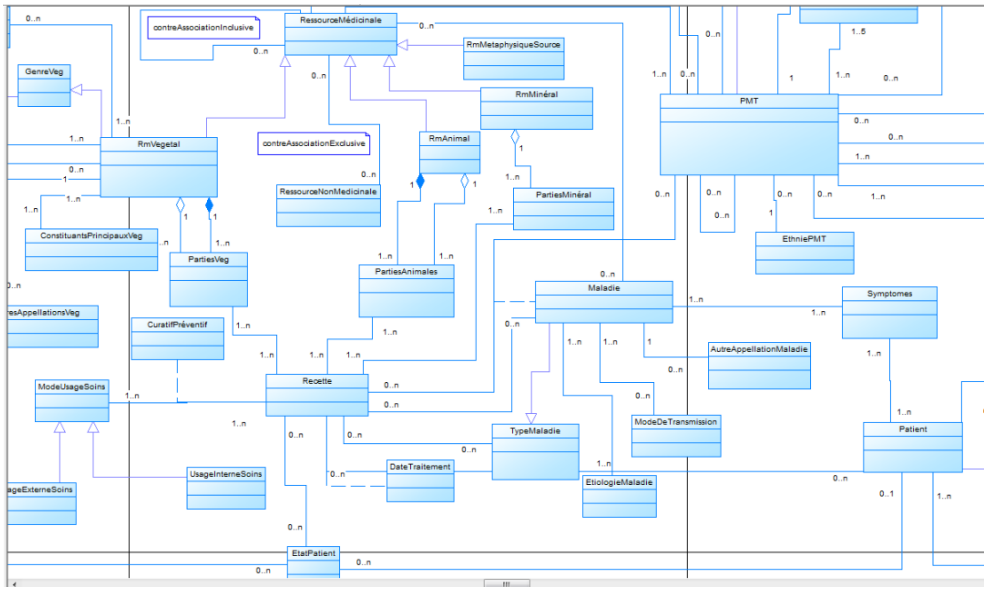


Annexe 1c : association de classes pour iconèmes



10

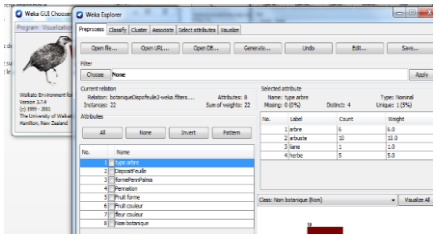
Annexe 2 : modèle de données pour ontoMEDTRAD

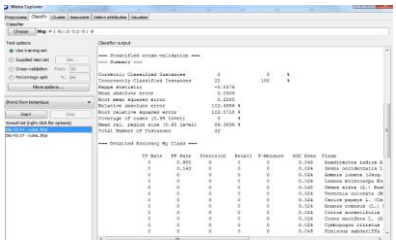


Annexe 3 : deux classifications par strate variées des plantes végétales, selon deux botanistes.

selon « Classe des plantes, Types biologiques chez les végétaux intertropicaux par J.-L. Trochain »	selon « FAMILLES DES PLANTES *par M. Adanson* » *Académie des Sciences*
1-Herbes mineures 2-Sous-arbrisseaux 3-Herbes majeures 4-Plantes herbacées 7-Plantes charnues 6-Arbrisseaux 8-Arbustes 9-Palmiers, bambous, 10-Arbres (des boisements clairs) 11-Arbres (des forêts) ou grand arbre	1 Arbres 2 Arbres fructifères 3 Arbres et Arbrisseaux. Baccifères 4 Siliqués 5 Arbrisseaux 6 Arbrisseaux grimpants 7 Herbes 8 Herbes pommifères et légumineuses

Annexe 4 : critères retenus pour la composition iconique d'une plante



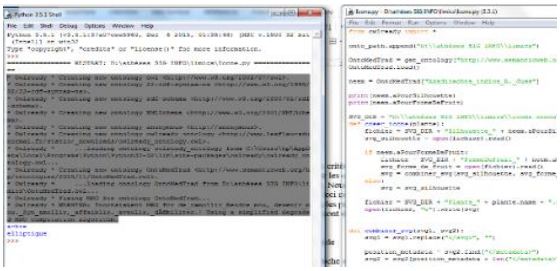


11

Annexe 5 : tableau des iconèmes

critères finaux	code de l'objet ou mnémonique	type d'objet	présence iconème (oui=O et non = N)	nombre d'instances	nombre d'icônes
type arbre ou silhouette	Silhouette	classe	O	5	5
fruit couleur	Couleur	classe	N	8	0
fleur couleur	Couleur	classe	O		
forme du fruit	FormeDufruit FormeDeFruitGenerique FormeDeFruitSpecifiqueConnue	classe sous classe sous classe	 O O	 7 3	 7 3
forme de la feuille (de couleur verte d'office)	FormeDeFeuille FormeDeFeuilleGenerique FormeDeFeuilleSpecifiqueConnue	classe sous classe sous classe	 O O	 8 2	 8 2
pennation de la feuille	Pennation	classe	O	2	2
disposition des feuilles	DispositionDeFeuille	classe	O	5	5
	Fleur	classe	O	1	1
NomBotanique	Plante	Classe	***résultat***	33	**33**

Annexe 6 : code exécuté sous python (idle)



Annexe 7 : tableau d'extrait de screening phytochimique de quelques plantes médicinales utilisées en pays Krobou (Agboville, Côte-d'Ivoire) ([15])

Tableau 1: Plantes et indications thérapeutiques

Signification des abréviations et sigles : Alc: Alcoolique ; Badig: Badigeonnage ; Citn: Citron; ER: Ecorce de racine ; ET: Ecorce de tig
Exp : Expression ; Fe : Feuille ; Fr : Fruit ; Ge : Graine ; Instil : Instillations ; KTK : Koutoukou ; Macré : Macéré ; Pge : Pétrissage ; Pu
Pulvérisation; Ram : ramollissement ; RF : Rameau feuillé ; TC : tubercule caulinaire; Tritn: Trituration.

Plantes	Affections, Symptômes, Effets thérapeutiques et phénomènes morbides	Parte utilisée	Modes de préparation	Formes médicamenteuses	Modes d'administration
Aframomum melegueta	Métrorragie	Fe Ge	Torréfaction, Pulv + eau	Poudre	Purge
Ageratum conyzoides	Accouchement facile	Fe	Expression	Extrait	Instil nasales
	Migraine	Fe	Pilage + eau	Pâte	Instil nasales
	Paludisme	Fe	Décoction	Décocté	Ablution
Boerhvia diffusa	Asthme	Fe	Infusion	Infusé	Boisson
	Choléra	Fe	Pétrissage +eau Décoction	Pâte Décocté	Purge Boisson
	Hémorragie après délivrance	Fe TC	Trituration Râpage	Extrait Extrait	Instillations vaginales
	Paludisme	Fe	Décoction	Décocté	Boisson
	Zona	Fe	Trituration +eau	Pâte	Badig, Purge
Capsicum frutescens	Rhumatisme	Fr	Pétrissage	Pâte	Friction
	Stimulant laxatif	Fr	Pétrissage + eau	Pâte	Purge
Chromolaena odorata	Diabète	Fe	Décoction	Décocté	Boisson
	Diarrhée	Fe	Pétrissage + eau	Pâte	Purge
	Paludisme	Fe	Décoction	Décocté	Boisson
Coffea canephora	Hypertension	Fe	Décoction	Décocté	Boisson
Cola nitida	Accouchement facile	ET	Ramollissement	Extrait	Instil buccales
	Migraine	ET	Pétrissage + eau	Pâte	Instil nasales
Cordia vignei	Paludisme	Fe	Décoction	Décocté	Boisson
	Accouchement facile	Fe	Ram, exp	Extrait	Instil buccales

12

Annexe 8 : **recettes avec modes de préparation et d'administration à base de 22 plantes médicinales antipaludiques**

	Plante	Partie utilisée	Mode de préparation	Mode d'administration
Mode de préparation et d'administration	*Azadirachta indica A. Juss.*(**Meliaceae)**[zerbo]	Feuille	décoction de l'organe avec les feuilles de *Senna occidentalis*	fumigation et boisson
	Senna occidentalis L. (famille **Fabaceae,** sfamille **Caesalpinaceae) [**zerbo] **Ordre** : Fabales.	Feuille	décoction de l'organe avec les feuilles de *Azadirachta indica*	boisson et bain
	Senna occidentalis (L.) Link **(Fabaceae-aesalpinioideae)** [nguessan]	Feuille	décoction	boisson
	Adenia lobata (Jacq.) Engl. (Passifloraceae) [nguessan]	Feuille	Décoction	boisson
	Adenia lobata (Jacq.) Engl. (Passifloraceae) [Piba]	tige	Décoction	boisson
	Lannea microcarpa Engl. et K. Krause **(Anacardaiceae)** [zerbo]	Ecorce-racine-feuilles	décoction du mélange	boisson et bain
	Senna alata (L.) Roxb. **(Caesalpinaceae)** [zerbo]	Feuilles	décoction de l'organe avec les feuilles de *Carica papaya* et de celles de *Vernonia colorata*	boisson et bain
	Senna alata (L.) Roxb. **(Fabaceae-Caesalpinioideae)**[nguessan]	Feuille	Infusion	Boisson
	Vernonia colorata (Willd.) Drake**(Asteraceae)** [zerbo]	Feuilles	décoction de l'organe avec les feuilles de *Senna alata*	boisson et bain
	Carica papaya L. **(Caricaceae)** [zerbo]	Feuille	Décoction	Boisson
	Carica papaya L. (Caricaceae) [nguessan]	Feuille	Décoction	Boisson
	Carica papaya L. (Caricaceae) [nguessan]	Fleur	Infusion	boisson
	Carica papaya L. (Caricaceae) [nguessan]	Fruit	Pétrissage + eau	Absorption
	Carica papaya L. **(Caricaceae)** *Ananas comosus (L.) Merr.* **(Bromeliaceae)** *Citrus aurantifolia (Christm.) Swingle* **(Rutaceae)** *Senna occidentalis (L.) Link* **(Fabaceae-aesalpinioideae)** *Cocos nucifera L.* **(Arecaceae)** [nguessan]	Fleur Fruit Fruit Feuille Racine	Décoction	Boisson
	Cymbopogon citratus (DC.) Stapf **(Poaceae)** [nguessan] *Hibiscus sabdariffa L.* **(Malvaceae)** *Manihot esculenta Crantz* **(Euphorbiaceae)**	Feuille Pétiole Feuille	Décoction+jus citron	Boisson
	Chrysophyllum africanum A. DC. **(Sapotaceae)** [nguessan]	Feuille Ecorce de tige	Décoction Pétrissage + eau	Bain de vapeur Boisson
	Senna alata (L.) Roxb. **(Fabaceae-Caesalpinioideae)** [nguessan] *Senna occidentalis (L.) Link* **(Fabaceae-Caesalpinioideae)** *Enantia polycarpa (DC.) Engl. et Diels* **(Annonaceae)**	Feuille Feuille Ecorce de racine	Décoction Décoction Pétrissage + eau	Bain de vapeur Ablution Purge, badigeonnage
	Enantia polycarpa (DC.) Engl. et Diels **(Annonaceae)** ou *(Annickia = Enantia)* [Piba]	Ecorce	décoction -Décocté, macération-macéré	Boisson, purge
	Ficus exasperata Vahl (Moraceae) [nguessan]	Ecorce de tige Ecorce de tige	Macération Pétrissage avec piment	Boisson Purge
	Combretum micranthum G. Don**(Combretaceae)** [zerbo]	Tige feuillée	décoction de l'organe avec la racine de Cochlospermum tinctorium	boisson et bain
	Musa x paradisiaca L. **(Musaceae)** [nguessan]	Feuille	Décoction +jus citron	boisson
	Scaphopetalum amoenum A. Chev. **(Malvaceae)** [nguessan]	Feuille Feuille	-Décoction -Macération alcoolique	Bain de vapeur, Boisson
	Cochlospermum tinctorium Perr. Ex A. Rich.**(Cochlospermaceae)** [zerbo]	Racines séchées	décoction	Boisson
	Nephrolepis biserrata **(Davalliaceae)** [adou lydie]	feuilles fraîches	décoction de l'organe avec des feuilles sèches de Carica papaya (mâle), de Musa Paradisiaca, et de et des jeunes feuilles de Citrus limon	Boisson
	Eucalyptus camaldulensis Dehnhardt **(Myrtaceae)** [zerbo]	Feuilles	décoction chaud de l'organe avec les feuilles de Carica papaya et celles Senna occidentalis	boisson, fumigation et bain

Nouvelle taxonomie des méthodes de classification basée sur l'Analyse de Concepts Formels

Marwa Trabelsi[1] — Nida Meddouri[1,2] — Mondher Maddouri[3]

[1] Laboratoire d'Informatique, Programmation, Algorithmique et Heuristique
Université d'El Manar, Tunisie
trabelsimarou@live.com

[2] Département Technologies Multimédia et Web
Institut Supérieur des Langues Appliquées et Informatique de Beja
Université de Jendouba, Tunisie
nida.meddouri@gmail.com

[3] Department of Computer Sciences, Community College of Hinakya
Taibah University, Medina
Kingdom of Saudi Arabia
mondher.maddouri@fst.rnu.tn

RÉSUMÉ. Des diverses approches ont été proposées dans la classification supervisée de données parmi lesquelles l'approche basée sur l'Analyse de Concepts Formels. Cet article présente une vue d'ensemble des méthodes de classification basée sur l'Analyse de Concepts Formels. Nous traitons ce sujet en proposant une nouvelle taxonomie de ces méthodes et en présentant une étude comparative basée sur la complexité théorique de ces dernières.

ABSTRACT. Various approaches have been proposed in supervised classification, among them the approach based on Formal Concept Analysis. This paper presents an overview of classification methods based on Formal Concepts Analysis. We address this issue by proposing a new taxonomy of these methods and presenting a comparative study based on the theoretical complexity.

MOTS-CLÉS : Fouille de Données, Classification Supervisée, Analyse de Concepts Formels

KEYWORDS : Data Mining, Supervised Classification, Formal Concept Analysis

1. Introduction

Le volume de données connait une évolution considérable et perpétuelle. Plusieurs travaux se sont focalisés sur l'extraction de connaissances à partir des données. Fayyad et al [21] ont définit l'extraction de connaissances à partir des données comme étant l'acquisition de nouvelles connaissances qui sont potentiellement utiles à partir des faits cachés au sein de grandes quantités de données. L'un des processus fondamentaux de l'extraction de connaissances repose sur la fouille de données, en particulier la classification supervisée. Ce processus peut être réalisé par les réseaux de neurones, les arbres de décisions, les réseaux bayésiens, les machines à vecteur de support ou encore l'Analyse de Concepts Formels [7]. Le choix de l'Analyse de Concepts Formels repose sur la capacité de cette dernière de traiter de grandes quantités de données et de simplifier la prédiction des classes [17, 7]. La classification basée sur l'Analyse de Concepts Formels consiste à construire des modèles appelés classifieurs à partir des données permettant de prédire des classes pour les futures données. Elle vise à découvrir tous les regroupements possibles de concepts et d'extraire ensuite les règles de classification selon les concepts générés à partir des données [17]. L'opération est réalisée en deux phases : **une phase d'apprentissage** dans laquelle un classifieur est construit pour décrire un ensemble prédéterminé de classes d'objets à partir d'un ensemble d'apprentissage et **une phase de classification** où le classifieur construit est utilisé pour associer une classe à chaque nouvel objet.

Cet article est organisé comme suit : la section 2 présente les notions de base de l'Analyse de Concepts Formels. Nous abordons, ensuite, la nouvelle taxonomie de méthodes de classification supervisée. Enfin, dans la section 3 nous proposons une étude comparative des méthodes présentées.

2. Classification supervisée basée sur l'Analyse de Concepts Formels

2.1. Analyse de Concepts Formels

L'Analyse de Concepts Formels est développée autour d'une communauté de mathématiciens. Ensuite, elle a attiré progressivement les chercheurs en informatique et a servi de fondement théorique à de nombreuses applications. Des multiples méthodes d'analyse de données et de représentation de connaissances ont été proposées. Ces méthodes traitent l'information sous la forme d'une hiérarchie de concepts formels [2].

Un contexte formel relie un ensemble fini d'objets *G* à un ensemble fini d'attributs *M* grâce à une relation binaire *I*. Il est représenté sous forme d'un triplet `K=(G,M,I)` où *I* vaut 1 quand un objet *g* vérifie l'attribut *m* avec *g* et *m* appartiennent respectivement à *G* et *M* notée `I(g,m)=1` [2]. Le contexte formel peut ainsi être illustré par un tableau de deux dimensions où on présente les objets sur les lignes et les attributs sur les colonnes. La case (i,j) indique la valeur de la relation entre l'objet g_i et l'attribut m_j avec i varie de 1 à n et j varie de 1 à m (n est le nombre des objets et m est le nombre des attributs). Á partir d'un contexte formel `K=(G,M,I)`, nous pouvons extraire tous les concepts formels possibles. L'ensemble ordonné[1] de tous les concepts peut être organisé sous forme d'un

1. Un treillis est un ensemble ordonné dans lequel toute paire d'éléments admet une borne supérieure et une borne inférieure

treillis appelé treillis complet de concepts formels, dit treillis de Galois [2]. Les méthodes de classification basées sur l'Analyse de Concepts Formels adoptent généralement une approche exhaustive ou une approche combinatoire. Dans ce qui suit, nous détaillons ces approches en donnant un survol des méthodes existantes de chaque approche.

2.2. Les méthodes de classification exhaustive

Les méthodes exhaustives se servent d'un seul classifieur et d'un regroupement de concepts formels durant les phases de la classification supervisée de données[2]. Cependant, elles varient entre elles selon la taille du treillis (complet ou demi) utilisé.

2.2.1. Méthodes avec treillis complet

GRAND [5, 18], RULEARNER [10, 18], GALOIS [4, 18], CBALATTICE [1], NAVIGALA [11], HMCS-FCA-SC [8] et SPFC [6] ont utilisé des treillis complets comme espace de recherche. Ces méthodes valident les caractéristiques associées aux concepts niveau par niveau dans les treillis. La navigation dans le treillis commence à partir du concept minimal où tous les concepts sont considérés comme des candidats.

GRAND[3] et GALOIS sont les premières méthodes qui utilisent des treillis complets. GRAND, lors de la phase d'apprentissage, organise l'information extraite d'un contexte formel sous forme d'un treillis complet à l'aide d'un algorithme incrémental qui considère le contexte ligne par ligne (colonne par colonne) et construit les treillis par ajout successif de concepts. Il réalise la mise à jour des treillis par l'ajout de nouveaux noeuds et par suppression des connexions redondantes [19]. Ensuite, elle applique les règles les plus spécifiques [5] à chaque objet. GALOIS construit, à l'image de GRAND, un treillis complet de façon incrémentale et ascendante. Dans la phase de classification, le système calcule la similarité entre le nouvel objet et chaque concept qui correspond au nombre de propriétés du concept vérifiées par l'objet [4].

D'autres méthodes comme NAVIGALA et RULEARNER sont par la suite issues de GRAND. NAVIGALA[4] a comme par ticularité lors de la construction du treillis, l'utilisation d'un contexte d'objets décrits par des vecteurs numériques de taille fixe. Ces vecteurs sont stockés dans une table discrète qui devient par la suite binaire [11].

De même, RULEARNER utilise un treillis complet pour la recherche des règles de classification. Elle construit le treillis de la même façon que GRAND. Durant la classification, elle se sert du vote majoritaire pour la détermination des classes des objets [10].

CBALATTICE construit un treillis complet de concepts et applique des règles d'association dans le but d'extraire des règles de classification. La méthode est incrémentale et progressive, toute augmentation du nombre d'objets, d'attributs et des classes peut être manipulée de manière très efficace [1].

HMCS-FCA-SC[5] a également eu recours à la construction d'un treillis complet afin de créer un modèle de classification hiérarchique. Durant la classification, elle emploie une mesure de similarité cosinus [6] entre le nouvel exemple et les concepts sélectionnés pour la classification des données [8].

2. La plupart des méthodes citées ont utilisé des échantillons de la base d'UCI (http ://archive.ics.uci.edu/ml/)
3. Graph-based induction
4. Navigation into Galois Lattice
5. Hierarchical Multi-label Classifier System - Formal Concept Analysis with Similarity Cosine
6. Consiste à calculer la similarité entre deux vecteurs à n dimensions en déterminant le cosinus de l'angle entre eux.

Après la construction du treillis, SPFC[7] affecte à chaque concept un score qui indique si les concepts sont convenables pour la génération des règles de classification. SPFC cherche, ensuite, les voisins des concepts pertinents (ayant les scores les plus élevés). Les objets inconnus seront classés dans les classes auxquelles appartiennent leurs voisins [6].

Les limites des méthodes exhaustives résident dans la construction d'un classifieur ayant une capacité de prédiction insuffisante et une complexité exponentielle de leurs algorithmes d'apprentissage en terme de temps et de ressources mémoires utilisés.

2.2.2. Méthodes avec demi treillis

Pour remédier à ce problème, d'autres travaux comme LEGAL [9, 18], CIBLE [18], CLNN & CLNB [23, 18], IPR [12], CLANN [14] et CITREC [3] ont eu recours aux demi treillis. Un demi treillis de concepts est une structure mathématique permettant de représenter une partie du treillis de concepts de façon sélective [3].

LEGAL[8] construit un sup-demi treillis[9] de concepts en s'appuyant sur certains paramètres d'apprentissage. Durant la phase d'apprentissage, elle construit un ensemble ordonné de concepts en se basant sur la classe de chaque instance. Les instances positives et négatives sont les instances étiquetés par une classe positive ou une classe négative dans le contexte formel. Au cours de la classification LEGAL applique le vote majoritaire [9].

CIBLE[10] est réalisée en deux étapes successives : elle commence par la construction d'un sup-demi treillis à partir d'un contexte binaire puis il se sert d'une mesure de similarité pour la classification des nouvelles instances [18].

CLNN & CLNB[11] construisent un sup-demi treillis de façon descendante. Ensuite, elles incorporent respectivement un classifieur Bayésien naïf et un classifieur plus proche voisin dans chaque noeud du demi treillis construit. CLNN & CLNB utilisent la même technique de vote qui est le vote majoritaire lors de la phase de classification [23].

CITREC[12] construit le treillis à partir d'un contexte réduit contenant seulement une instance représentative de chaque classe [3]. Dans la phase de classification, CITREC utilise aussi le vote majoritaire comme les méthodes CLNN & CLNB.

CLANN[13] construit un sup-demi treillis durant l'apprentissage en traitant les données qui possèdent seulement deux classes. Puis, elle utilise ce demi treillis pour construire un réseau de neurone qui réalise la classification [14].

IPR[14] introduit la notion de couverture. Elle fait recours à un algorithme glouton pour construire la couverture de concepts. L'algorithme choisit les concepts pertinents et chaque concept est déterminé grâce à une optimisation locale de la fonction d'entropie [12]. Pour chaque nouvel objet, IPR cherche les règles dont leur prémisse coïncide avec les attributs et applique la règle la plus pondérée pour cet objet.

La classification basée sur un demi treillis se déroule de la même manière que celle basée sur un treillis complet. Les méthodes basées sur les demi treillis permettent, par contre, de minimiser l'ensemble de règles de classification générées en gardant les plus pertinentes ce qui engendre un gain considérable aux niveaux du temps et d'apprentissage mais une perte d'information en même temps.

---

7. Classification by Selecting Plausible Formal Concepts in a Concept Lattice
8. Learning with Galois Lattice
9. Un sup demi-treillis est un ensemble ordonnée dans lequel deux éléments quelconques x et y admettent toujours une borne supérieure
10. Concept Induction Based Learning
11. Concept Lattices Nearest Neignbors and Concept Lattices Naive Bayes
12. Classification Indexée par le treillis de concepts
13. Concept Lattice-based Artificial Neural Network
14. Induction of Production Rules

Concernant les méthodes exhaustives citées auparavant, nous observons plusieurs inconvénients. D'une part, leurs complexités sont élevées. D'autre part, elles utilisent un classifieur faible et unique. Ainsi, on observe l'absence de l'aspect adaptatif dans la classification. Par conséquent, les chercheurs s'orientent vers les méthodes combinatoires.

2.3. Les méthodes de classification combinatoire

Dans le but d'améliorer la performance d'un classifieur unique (estimé faible) qui est adopté par les méthodes exhaustives, les méthodes combinatoires génèrent un ensemble de classifieurs et les combinent par les techniques de votes.

Plusieurs méthodes ont été proposées dans ce cadre : des méthodes qui réalisent l'apprentissage séquentiel telles que BFC [15], BNC [16] et bien d'autres basées sur l'apprentissage parallèle comme DNC [17], FCA-BRG [13] et RMCS [22].

L'apprentissage séquentiel consiste à générer des classifieurs d'une manière séquentielle c'est-à-dire, un classifieur, n'est généré qu'après la génération de son prédécesseur. À titre d'exemple, BFC[15] construit à partir d'un contexte formel une couverture formée seulement des concepts pertinents. Cette dernière se base sur le principe du boosting qui est une approche adaptative basée sur l'utilisation de plusieurs classifieurs du même modèle [20]. L'idée de BFC consiste à affecter des poids égaux aux exemples d'apprentissage parmi lesquels un sous-ensemble est sélectionné à l'aide d'un tirage probabiliste. À ce stade, un concept pertinent est extrait à partir du sous ensemble en sélectionnant l'attribut qui minimise la fonction d'entropie de Shannon[16]. BFC permet ensuite de générer une règle de classification déduite à partir du concept pertinent (extraite du sous ensemble) et de faire une mise à jour des poids aux exemples d'apprentissage. Cette procédure est appliquée récursivement pour construire finalement le classifieur final [15].

BNC[17] procède de la même manière que BFC dans la génération des classifieurs et le traitement de données d'apprentissage. En revanche, contrairement à BFC qui effectue le traitement de données binaires, BNC manipule des données nominales dans le but d'éviter la perte d'information issue de la représentation binaire [16].

L'apprentissage parallèle basé sur le principe de Dagging [20], consiste à diviser l'ensemble de données en plusieurs groupes à partir desquels les classifieurs sont produits. DNC[18] traite des données nominales et se déroule comme suit : un tirage de données est effectué afin de créer des groupes disjoints contenant des données stratifiés. Un classifieur de concept nominal [16] est ensuite construit dans chaque groupe. Enfin, la méthode utilise la technique de vote pour définir une combinaison de sortie des classifieurs [16].

Dans l'apprentissage parallèle, FCA-BRG[19] commence également par la division de la base initiale en des sous ensembles de données. Ces derniers sont, ensuite, utilisés pour la génération des sous contextes formels afin d'extraire les règles de classification. Un algorithme génétique sera enfin appliqué pour séléctionner les meilleures règles [13].

RMCS[20] construit des classifieurs basés sur les voisins. Ils ne réalisent la classification correcte d'un objet s'ils ont classé correctement ses voisins. RMCS commence par la construction d'une table de classification à partir d'un contexte formel (les objets du contexte utilisé sont privés d'un ensemble d'objets test). Dans cette table, RMCS affecte

15. Boosting Formal Concepts
16. L'entropie de Shannon, est une fonction mathématique qui correspond à la quantité d'information contenue ou délivrée par une source d'information.
17. Boosting Nominal Concepts
18. Dagging Nominal Concepts
19. FCA Based Rule Generator
20. Recommender based Multiple Classifier System

les classifieurs aux objets existants dans le contexte. Ensuite, RMCS cherche les voisins des objets de test à l'aide d'une métrique de similarité, puis elle sélectionne des classifieurs qui ont le nombre maximal de voisins trouvés. Les classifieurs sélectionnés sont ainsi recommandés pour la classification [22].

3. Discussion

Comme nous l'avons indiqué, les méthodes de classification basées sur l'Analyse de Concepts Formels sont regroupées en deux catégories principales : exhaustives et d'autres combinatoires. Les méthodes de chaque catégorie se distinguent entre elles sur certain aspects mais en partageant d'autres. Les méthodes exhaustives ont comme point commun la génération d'un seul classifieur ordinaire pour la classification des objets.

Le tableau 1 présente les méthodes exhaustives évoquées précédemment. Afin de dégager les particularités de chaque méthode, nous avons eu recours à cinq critères qui nous semblent les plus déterminants. Le critère le plus important de cette comparaison consiste à calculer un ordre de compléxité en fonction des paramètres de classification utilisés. D'autres critères ont été utilisés tels que la structure de concepts, le type de données, la méthode de sélection de concepts utilisée lors de la classification et la méthode de classification qui désigne la manière d'affectation des nouveaux objets aux classes.

système	**structure**	**données**	**sélection**	**classification**	**compléxité**
GRAND	treillis complet	binaires	cohérence maximale	vote majoritaire	$O(2^{k} * k^{4})$ k=min(m,n)
CIBLE	demi treillis	numériques	fonction de sélection	K-PPV	$O(\|L\| * m^{3})$ |L|=demi treillis
IPR	couverture	binaires	entropie	règles pondérées	$O(n^{2} * m^{2} * nm)$
CITREC	demi treillis	binaires	support	vote	$O(2^{m} * n)$
CLANN	demi treillis	binaires	algorithmes heuristiques	réseau de neurone	$O(2^{\min(m,n)})$
HMCS-FCA-SC	treillis complet	nominales	meilleurs concepts	mesure de similarité	$O(2^{m} + n2^{m})$

Tableau 1. *Comparaison théorique des méthodes exhaustives*

Comme le montre le tableau 1, les méthodes exhaustives possèdent une complexité exponentielle. Cela est dû principalement à la navigation dans la totalité de l'espace de recherche contrairement aux méthodes combinatoires qui distribuent le processus de classification sur des multiples classifieurs. Le problème est ainsi décomposé en plusieurs sous-problèmes. Le tableau 2 distingue les méthodes combinatoires. Pour des raisons comparatives nous avons utilisé les cinq critères du tableau 1 en y ajoutant la méthode de combinaison de classifieurs employés. Les tableaux 1 et 2 montrent que GRAND, IPR, CITREC, CLANN, BFC et RMCS opèrent à partir des données binaires, tandis que BNC et DNC traitent des données nominales. En revanche, CIBLE se distingue par rapport aux précédentes de sa capacité de traiter des données numériques. BNC et DNC utilisent le gain informationnel dans la sélection des concepts, tandis que IPR et BFC se servent de l'entropie de Shannon. Quant à CLANN, elle utilise des algorithmes heuristiques pour la sélection.

système	structure	données	sélection	combinaison	classification	compléxité
BFC	couverture	binaires	entropie	boosting	vote pondéré	O(nlog(n)+nm)
BNC	couverture	nominales	gain informationnel	boosting	vote pondéré	O(nlog(n)+nm) m=attribut nominal
DNC	couverture de concepts pertinents	nominales	gain informationnel	dagging	Vote majoritaire	O(n') n'= sous échantillon stratifié [17]
RMCS	treillis complet	binaires	distance euclidienne	dagging	maximum de voisin	O(nmlog(n))

Tableau 2. *Comparaison théorique des méthodes combinatoires*

Concernant la phase de classification, GRAND, CITREC et DNC utilisent le vote majoritaire. Le vote pondéré a été appliqué par IPR, BFC et BNC. En revanche, CLANN diffère des autres méthodes par l'utilisation du réseau de neurone.

La technique de combinaison (cf. section 2.3) a joué un rôle important dans l'optimisation de la complexité[21]. En effet, les méthodes combinatoires qui génèrent des classifieurs de manière séquentielle ont une complexité polynomiale logarithmique. De même, les méthodes qui génèrent des classifieurs parallèles arrivent a une complexité comparable de l'ordre de $nmlog(n)$ pour la méthode RMCS et de n pour DNC.

4. Conclusion

Dans cet article, nous nous sommes intéressés par la classification supervisée de données basée sur l'Analyse de Concepts Formels. Nous avons présenté dans un premier temps les méthodes de classification exhaustive qui se divisent en des méthodes basées sur des treillis complets et des méthodes basées sur des demi treillis. Dans un deuxième temps, nous avons décrit les méthodes de classification combinatoire qui elles-mêmes se décomposent en des méthodes basées sur l'apprentissage séquentiel et des méthodes basées sur l'apprentissage parallèle. Nos perspectives reposent sur la complexité et s'orientent vers les méthodes combinatoires qui offrent une complexité plus raisonnable, en particulier les méthodes qui génèrent des classifieurs parallèles.

5. Bibliographie

[1] A. GUPTA, N. KUMAR, V. BHATNAGAR, « Incremental classification rules based on association rules using formal concept analysis », *Machine Learning and Data Mining in Pattern Recognition*, vol. 10, n° 11-20, 2005.

[2] B. GANTER, G. STUMME, R. WILLE, « Formal concept analysis : foundations and applications », *Springer Science Business Media*, vol. 3626, 2005.

[3] B. DOUAR, C. LATIRI, Y. SLIMANI, « Approche hybride de classification supervisée à base de treillis de Galois : application à la reconnaissance de visages », *Extraction et Gestion des Connaissances*, n° 309-320, 2008.

[4] C. CARPINETO, G. ROMANO, « Concept data analysis : Theory and applications », *Livre, John Wiley Sons*, vol. 23, 2004.

21. Notez que 'n' est le nombre dobjets et 'm' le nombre dattributs

[5] G. OOSTHUIZEN, « The use of a lattice in knowledge processing », *Thesis, University of Strathclyde*, 1988.

[6] I. MADORI, Y. AKIHITO, « Classification by Selecting Plausible Formal Concepts in a Concept Lattice », *Formal Concepts Analysis meets information Retrieval* , vol. 14, n° 22-35, 2013.

[7] J. POELMANS, D. IGNATOV, G. DEDENE, « Formal concept analysis in knowledge processing : A survey on applications », *Expert systems with applications*, vol. 40(16), n° 6538-6560, 2013.

[8] M. FERRANDIN, J. NIEVOLA, F. ENEMBRECK, E. SCALABRIN, K. KREDENS, B. AVILA, « Hierarchical Classification Using FCA and the Cosine Similarity Function », *International Conference on Artificial Intelligence*, vol. 6, n° 281-287, 2013.

[9] M. LIQUIERE, E. MEPHU NGUIFO, « Legal : learning with galois lattice », *Journées Françaises sur lApprentissage*, n° 93-113, 1990.

[10] M. SAHAMI, « Learning classification rules using lattices » *European Conference on Machine Learning*, n° 343-346, 1995.

[11] M. VISANI, K. BERTET, J. OGIER, « Navigala : An original symbol classifier based on navigation through a galois lattice », *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, n° 449-473, 2011.

[12] M. MADDOURI, « Towards a machine learning approach based on incremental concept formation », *Intelligent Data Analysis*, vol. 8, n° 267-280, 2004.

[13] M. CINTRA, M. MONARD, H. CAMARGO, « FCA-BASED RULE GENERATOR, a framework for the genetic generation of fuzzy classification systems using formal concept analysis. », *In Fuzzy Systems (FUZZ-IEEE)*, n° 1-8, 2015.

[14] N. TSOPZÉ, E. MEPHU NGUIFO, G. TINDO, « CLANN : Concept Lattice-based Artificial Neural Network for Supervised Classification » *Concept Lattice and their applications*, vol. 331, 2007.

[15] N. MEDDOURI, M. MADDOURI, « Boosting formal concepts to discover classification rules », *Next-Generation Applied Intelligence*, n° 501-510, 2009.

[16] N. MEDDOURI, M. MADDOURI, « Adaptive learning of nominal concepts for supervised classification », *Knowledge-Based and Intelligent Information and Engineering Systems*, n° 121-130, 2010.

[17] N. MEDDOURI, H. KHOUFI, M. MADDOURI, « Parallel Learning and Classification for Rules based on Formal Concepts », *Knowledge-Based and Intelligent Information and Engineering Systems*, vol. 35, n° 358-367, 2014.

[18] P. NJIWOUA, E. MEPHU NGUIFO, « Treillis de Concepts et Classification Super-visée. », *Technique et Science Informatiques*,vol. 24(4), n° 449-488, , 2005.

[19] R. GODIN, R. MISSAOUI, H. ALAOUI, « Incremental concept formation algorithms based on Galois (concept) lattices », *Appeared in Computational Intelligence*, vol. 11(2), n° 246-267, 1995.

[20] S. KOTSIANTI, D. KANELLOPOULOS, « Combining bagging, boosting and dagging for classification problems », *Knowledge-Based Intelligent Information and Engineering Systems*, vol. 7, n° 493-500, 2007.

[21] U. FAYYAD, G. PIATETSKY-SHAPIRO, P. SMITH, « Advances in knowledge discovery and data mining », *National Conference on Artificial Intelligence*, vol. 2, n° 2, 1996.

[22] Y. KASHNITSKY, D. IGNATOV, « Can FCA-based Recommender System Suggest a Proper Classifier ? », *What can FCA do for Artificial Intelligence*, vol. 2, n° 2, 2015.

[23] Z. XIE, W. HSU, Z. LIU, M. LEE, « Concept lattice based composite classifiers for high predictability », *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 14(2-3), n° 143-156, 2002.

Kernel-based performance evaluation of coded QAM systems

Poda Pasteur[1,2] - Saoudi Samir[2] - Chonavel Thierry[2] - Guilloud Frédéric[2] - Tapsoba Théodore[1]

1 Ecole supérieure d'Informatique, Université polytechnique de Bobo-Dioulasso, Bobo-Dioulasso, Burkina Faso. pasteur.poda@univ-bobo.bf, theo_tapsoba@univ-ouaga.bf.

2 Département SC, Telecom Bretagne, Institut Mines-Telecom, Lab-Sticc UMR 6285, Brest, France. samir.saoudi@telecom-bretagne.eu, thierry.chonavel@telecom-bretagne.eu, frederic.guilloud@telecom-bretagne.eu.

RÉSUMÉ. Les estimateurs de taux d'erreur binaire par méthode à noyau sont d'un intérêt récent pour la réduction du coût des méthodes de Monte Carlo. Pour le moment, ils sont surtout appliqués à des modulations binaires. Dans ce papier, un estimateur à noyau est conçu pour des systèmes M-aires codés de Modulation d'Amplitude en Quadrature (MAQ). Les observations utilisées pour l'estimation sont définies sous forme de bits à valeurs souples bornées. Un noyau d'Epanechnikov est choisi et son paramètre de lissage obtenu sur la base du concept de bande passante canonique. Des simulations sont réalisées pour des systèmes MAQ-4 et MAQ-16 impliquant des canaux à bruit additif blanc Gaussien ainsi qu'à évanouïssements de Rayleigh. Les résultats obtenus montrent que l'estimateur proposé produit des gains en coût significatifs qui croissent avec E_b/N_0.

ABSTRACT. Kernel Bit Error Rate (BER) estimators are of recent interest for Monte Carlo sample size reduction. Until now, they mainly addressed binary modulation systems. In this paper, a kernel-based BER estimator is designed for coded M-ary Quadrature Amplitude Modulation (QAM) systems. The observations from which estimations are made are defined in the form of bounded soft bits. An Epanechnikov kernel function is selected and its smoothing parameter is derived based on the concept of canonical bandwidth. Simulations are run for 4-QAM and 16-QAM systems, involving additive white Gaussian noise and frequency-selective Rayleigh fading channels respectively. Simulation results show that the proposed estimator yields significative sample savings that grow with E_b/N_0.

MOTS-CLÉS : Taux d'erreur binaire, Estimateur à noyau, Méthode Monte Carlo, Fonction de densité de probabilité.

KEYWORDS : Bit Error Rate, Kernel estimator, Monte Carlo method, Probability density function.

1. Introduction

The Bit Error Rate (BER) is a measure of performance largely used in digital communications domain. Analytical BER estimation techniques have been studied [1], [2]. However, closed-form solutions are generally unavailable when considering complex digital communication systems. More successful have been simulation-based techniques at the core of which is the Monte Carlo (MC) method. The MC method is a universal technique that supplies an empirical determination of the BER estimate and that is commonly used as a reference for other methods. Its weak point is its high computational cost.

Since the 1970s, simulation-based techniques [3] were developed in order to reduce the sample size that the MC method requires to achieve accurate estimation. Recently, new BER estimation methods based on non-parametric *probability density function* (pdf) estimation have shown to achieve good performance for the uncoded binary-input Gaussian channel : namely Gaussian mixture models [4] and kernel estimators [5]. In [6], a kernel-based soft BER estimator is applied to Code Division Multiple Access (CDMA) schemes, for which efficient and reliable BER estimates have been reported. In [7], it is shown that kernel-based BER estimations can perform well in a blind way. Using Maximum Likelihood for the smoothing parameter optimisation, kernel method for BER estimation was applied to binary coded transmission schemes involving Turbo and Low Density Parity Check (LDPC) codes over CDMA systems [8].

To the best of our knowledge, BER estimation using kernel methods has been so far only applied to CDMA systems over Additive White Gaussian Noise (AWGN) channels. In this paper, we first address the issue of general M-ary modulations. Shifting from 2-ary real constellations to M-ary complex modulations involves the estimation of complex pdfs. As QAM systems are largely included in standards, we focus on this family of M-ary modulations. Secondly, we address the issue of estimating the BER when transmitting over frequency-selective fading channels. Hence, the distribution of the soft observations loses its Gaussian nature and finding an ad-hoc smoothing parameter for the kernel is not straightforward. In the remainder, we give a theoretical formulation of the Bit Error Probability (BEP) in Section 2 and present the principle of kernel-based estimation technique in Section 3. We describe the proposed kernel-based BER estimator in Section 4 while reporting simulation results in Section 5. In Section 6, we conclude the paper.

2. Theoretical formulation of the BEP

Let us consider a coded digital communication system that operates with Quadrature Amplitude Modulation (QAM) schemes. A signal containing coded QAM waveforms of alphabet $\{S_1, S_2, \ldots, S_M\}$ is transmitted over a noisy channel. M is the constellation size. At the receiver-end, we assume that the channel decoder delivers N independent and identically distributed soft bits $(X_j)_{1\leq j\leq N}$. Let X denote the univariate real random variable that describes the soft bits $(X_j)_{1\leq j\leq N}$ and let $f_X^{(0)}$ (resp. $f_X^{(1)}$) be the conditional pdf of X such that the transmitted bit $b_i = 0$ (resp. $b_i = 1$). The BEP can be stated as :

$$
\begin{aligned}
p_e &= \Pr[X > 0, b_i = 0] + \Pr[X < 0, b_i = 1] && (1)\\
&= \Pr[X > 0 \,|\, b_i = 0]\Pr[b_i = 0] + \Pr[X < 0 \,|\, b_i = 1]\Pr[b_i = 1] && (2)\\
&= \pi_0 \int_0^{+\infty} f_X^{(0)}(x)\,\mathrm{d}x + \pi_1 \int_{-\infty}^{0} f_X^{(1)}(x)\,\mathrm{d}x, && (3)
\end{aligned}
$$

where π_0 and π_1 are the *a priori* probabilities of bits values "0" and "1" respectively.

The BER is an estimate of the BEP. Based on the MC approach, it is estimated by counting the errors that occured on the transmitted data. Based on the kernel technique, the principle of its estimation is described in the following.

3. Kernel-based soft BER estimation

In kernel-based BER estimation, the marginal conditional pdfs $f_X^{(0)}(x)$ and $f_X^{(1)}(x)$ are estimated as follows :

$$\hat{f}_X^{(b_i)}(x) = \frac{1}{n_{b_i}} \sum_{j=1}^{n_{b_i}} \frac{1}{h_{b_i}} K\left(\frac{x - X_j}{h_{b_i}}\right), \tag{4}$$

where K is any even regular pdf with zero mean and unit variance called the kernel, n_{b_i} is the cardinality of the subset of the soft observations $(X_j)_{1\leq j\leq N}$ which are likely to be decoded into a binary "0" bit value (resp. "1") and h_{b_i} is a parameter called smoothing parameter (or bandwidth) that depends on the soft observations $(X_j)_{1\leq j\leq n_{b_i}}$. Then, p_e in Eq. (3) can be estimated as

$$\hat{p}_e = \pi_0 \int_0^{+\infty} \hat{f}_X^{(0)}(x)\,\mathrm{d}x + \pi_1 \int_{-\infty}^{0} \hat{f}_X^{(1)}(x)\,\mathrm{d}x. \tag{5}$$

The choice of the kernel K is related to the density function under estimation. Whenever the observed samples are distributed over a large scale, distributions with an infinite support (e.g., Gaussian distribution) are well suited. However, finite support distributions such as Epanechnikov or Quartic distributions should be selected to model K when the observed samples are bounded.

The design of the smoothing parameter h is a major issue since it significantly governs the accuracy of the estimation. To the end, optimisation of h with respect to some given constraints has been proposed. One of the most popular is the Asymptotic Mean Integrated Squared Error (AMISE) criterion. When the AMISE criterion is used, the optimal smoothing parameter is derived [9] as,

$$h^*_{\text{AMISE}} = \left[\frac{\int K^2(x)\,\mathrm{d}x}{\int f_X''(x)^2\,\mathrm{d}x \left(\int x^2 K(x)\,\mathrm{d}x\right)^2}\right]^{1/5} N^{-1/5}, \tag{6}$$

where $f_X''(x)$ is the second derivative of the pdf $f_X(x)$. Clearly, the constraint in Eq. (6) is the prior knowledge of the target distribution f_X, which is of course unknown and searched for. In practice, some reference distribution can be used to replace f_X, with mean and variance matching those of data. In the literature, the Gaussian distribution is a popular choice for f_X. Many designs of h^*_{AMISE} can be found including this recent one given as follows [10] :

$$h^*_{\text{Gau}} = (4/3)^{1/5} \min\left(\hat{\sigma}, IQR/1.34\right) N^{-1/5}. \tag{7}$$

where $\hat{\sigma}$ is the standard deviation of the data and IQR is their interquartile range.

4. Proposed kernel-based BER estimator scheme

Let us consider a digital communication system that includes a channel codec (encoder/decoder). The coded BER is the BER that is determined at the output of the channel decoder. A kernel-based soft coded BER estimator is proposed in this paper. Suited soft bits have to be given at the entry of the estimator. We define the soft bits as follows :

$$X_j = \Pr[b_j = 1|r] - \Pr[b_j = 0|r], \tag{8}$$

where r is the received signal. Let us assume that the channel decoder requires soft inputs in the form of Log-Likelihood Ratio (LLR). Each M-ary QAM soft symbol at the output of the channel carries $k = log_2(M)$ LLR bits $(L_j)_{1\leq j\leq k}$ that can be retrieved by a symbol-to-bit soft demapping [11]. We also assume that the outputs of the channel decoder are soft LLR bits. The *j*th LLR, L_j, is defined as

$$L_j = \log\left(\frac{\Pr[b_j = 1|r]}{\Pr[b_j = 0|r]}\right). \tag{9}$$

From Eq. (8), Eq. (9) and constraint $\Pr[b_j = 1|r] + \Pr[b_j = 0|r] = 1$, the soft bit X_j is derived in terms of the channel decoder output L_j as follows :

$$X_j = \frac{1 - e^{-L_j}}{1 + e^{-L_j}}. \tag{10}$$

Using the soft bits $(X_j)_{1\leq j\leq N}$, the proposed kernel-based estimator can perform, provided a kernel function K and a suitable smoothing parameter h are selected.

As shown in Eq. (10), the soft bits $(X_j)_{1\leq j\leq N}$ are bounded between -1 and $+1$. So, among the kernel function with bounded support, the Epanechnikov kernel function $K(x) = \frac{3}{4}\left(1 - x^2\right) I(|x| \leq 1)$ is chosen. Then it can be checked that the kernel estimator with bandwith h will be restricted to interval $[-1-h, 1+h]$. Since optimaly chosen h remains much smaller than 1 for large samples, we can consider that numericaly the support constraint for the distribution of X is satisfied when using the Epanechnikov kernel. Therefore, we need to find the corresponding smoothing parameter h^*_{Epa} that approximates well h^*_{AMISE} of Eq. (6). As h^*_{Gau} is a good approximation of h^*_{AMISE} of Eq. (6) in the context of a Gaussian kernel, the idea is to derive h^*_{Epa} from h^*_{Gau} based on the concept of canonical bandwidth [12]. The parameter h^*_{Epa} is then expressed as

$$h^*_{\text{Epa}} = \frac{\delta_{Epa}}{\delta_{Gau}} h^*_{\text{Gau}}, \tag{11}$$

where, from [12] $\delta_{Gau} \approx (1/4)^{1/10} = 0.7764$ and $\delta_{Epa} \approx 15^{1/5} = 1.7188$ are the canonical bandwidths of the Gaussian and Epanechnikov kernels.

At this stage, the expressions of the two marginal conditional pdfs $\hat{f}^{(0)}_X(x)$ and $\hat{f}^{(1)}_X(x)$ can be derived from Eq. (4) and then, Eq. (5) can be rewritten as follows :

$$\hat{p}_e = \pi_0 \int_0^{+\infty} \frac{1}{n_0} \sum_{j=1}^{n_0} \frac{1}{h^*_0} K\left(\frac{x - X_j}{h^*_0}\right) \mathrm{d}x + \pi_1 \int_{-\infty}^{0} \frac{1}{n_1} \sum_{j=1}^{n_1} \frac{1}{h^*_1} K\left(\frac{x - X_j}{h^*_1}\right) \mathrm{d}x, \tag{12}$$

where h^*_0 (resp. h^*_1), computed according to Eq. (11), is the selected optimal bandwidth which will govern the estimation accuracy of $\hat{f}^{(0)}_X(x)$ (resp. $\hat{f}^{(1)}_X(x)$). After transforma-

tions that are detailed in Appendix, Eq. (12) leads to the expression of the coded BER estimate as follows :

$$\hat{p}_e = \frac{\pi_0 L_0}{n_0} + \frac{\pi_1 L_1}{n_1} + \sum_{\substack{|\alpha_j|\leq 1,\\ 1\leq j\leq n_0}} \frac{3\pi_0}{4n_0}\left(\frac{2}{3} - \alpha_j + \frac{\alpha_j^3}{3}\right) + \sum_{\substack{|\beta_j|\leq 1,\\ 1\leq j\leq n_1}} \frac{3\pi_1}{4n_1}\left(\frac{2}{3} + \beta_j - \frac{\beta_j^3}{3}\right), \tag{13}$$

where $\alpha_j = -X_j/h_0^*$, $\beta_j = -X_j/h_1^*$, L_0 (resp. L_1) is the cardinality of the subset of $(\alpha_j)_{1\leq j\leq n_0}$ (resp. $(\beta_j)_{1\leq j\leq n_1}$) which are less than -1 (resp. greater than 1). Based on Eq. (13), coded BER estimates can be evaluated using soft bits $(X_j)_{1\leq j\leq N}$.

5. Simulation results

The proposed estimator has been simulated on a single-carrier QAM transmission scheme over the AWGN channel and also on a multi-carrier QAM transmission scheme over a frequency-selective Rayleigh fading channel. A Gray-coded 4-QAM and 16-QAM constellations were considered. The Rayleigh channel was ten taps long with a sample period of $12.8\mu s$, an $8Hz$ maximum Doppler shift and average taps gains given in $watts$ by the vector $[0.0616\ 0.4813\ 0.1511\ 0.0320\ 0.1323\ 0.0205\ 0.0079\ 0.0778\ 0.0166\ 0.0188]$. To mitigate inter-symbol and inter-carrier interferences, a Cyclic Prefix (CP) Orthogonal Frequency Division Multiplexing (OFDM) technique was implemented. The length of the CP was set to 9 and the number of OFDM sub-carriers set to 128. A 128-point FFT (Fast Fourier Transform) was performed. The Channel codec was a $4/7$-rate LDPC code with a Gallager-based parity check matrix built to be of rank 15. The number of iterations was set to 10 (resp. 30) for the AWGN (resp. Rayleigh) channel. An Epanechnikov kernel function and the smoothing parameter of Eq. (11) were selected.

We evaluate the performance in terms of absolute bias and Confidence Interval (CI). The absolute bias is defined as $|\,\mathbb{E}[\hat{p}_e] - p_e\,|$ where $\hat{p}_e$ represents an estimate of the coded BER. The true BER p_e is computed in the form of a benchmark using MC simulations. The CI has been calculated for a 95% confidence level. To validate the proposed estimator over the AWGN channel, *Figure* 1 offers a visual way to evaluate the bias for 4-QAM and 16-QAM transmission schemes. We can see that the kernel-based coded BER estimates data points are very close to the true BER (benchmark) from values greater than 10^{-1} down to 10^{-5}. *Table* 1 illustrates the bias and the CI using numerical data related to 4-QAM system simulation. From the observed CIs and their corresponding kernel sample sizes N_K, we derived (see [3]) the required sample sizes for MC simulations to yield equal performance and noted sample savings up to a factor 16. As for the performance achieved over the Rayleigh channel, the green curves with diamond marks in *Figure* 2 illustrate that coded BER estimates are close to their corresponding benchmarks. Detailed information about the bias, the CIs and the sample sizes is provided in *Table* 2 as far as 16-QAM transmission schemes are concerned. A thorough analysis of the observed numerical data let us notice that all the data points on the green curves are associated to coded BER values that fall into their corresponding CIs. The observed smallest CI is $[0.89p_e,\ 1.11p_e]$ and the largest of all is $[0.52p_e,\ 1.48p_e]$. If we considered $[0.50p_e,\ 1.50p_e]$ as the largest CI over which the estimator is declared not reliable and combining with the fact that all the mean values of the BER estimates are inside their corresponding CIs, we can conclude, at the light of the observed CIs, that the proposed estimator is reliable for BER values down to the neighbourhood of 10^{-4}.

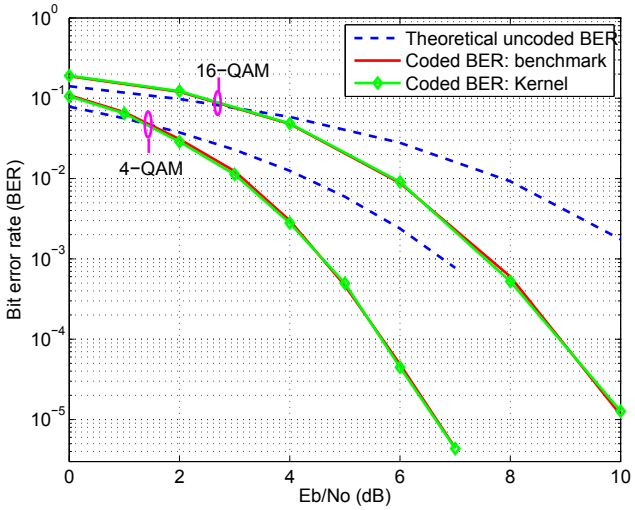


Figure 1. *Performance of the proposed estimator over the AWGN channel*

Regarding the efficiency, the two last Columns of *Table* 2 show that the proposed estimator requires less samples than the MC method. The given kernel (N_k) and MC (N_{mc}) sample sizes are those required for the two methods to achieve (almost) equal bias and CI. To illustrate this, let us consider the row of $E_b/N_0 = 12\,dB$ in *Table* 2. The proposed estimator achieved an efficiency described by a sample size of $50\,000$ against $127\,995$ for the MC estimator. In the same time, the proposed estimator achieved a CI of $[0.81p_e,\ 1.19p_e]$ versus $[0.80p_e,\ 1.20p_e]$ for the MC estimator. The two estimators performed the estimation with almost equal bias (0.0011 for the MC method against 0.0012 for the proposed kernel method). Moreover, for $E_b/N_0 = 20\,dB$ in *Table* 2, both the MC and the proposed estimators performed an estimate with equal bias and achieved CIs are $[0.62p_e,\ 1.38p_e]$ for the MC estimator against $[0.67p_e,\ 1.33p_e]$ for the proposed one. The corresponding sample saving achieved by the proposed estimator is at least of a factor 5.

Behind this efficiency of the proposed estimator is also hidden its performance in terms of the power consumption. The MC method and the proposed estimator yield almost equal CPU time for equal sample sizes ; e.g. : at $E_b/N_0 = 20\,dB$ and for a sample size of

Tableau 1. *Numerical results of coded* 4*-QAM BER estimation over AWGN channel*

E_b/N_0	Benchmark	Bias	CI	N_K
$00\,dB$	1.1×10^{-1}	0.03×10^{-1}	$[0.94p_e,\ 1.06p_e]$	10^3
$01\,dB$	6.7×10^{-2}	0.22×10^{-2}	$[0.90p_e,\ 1.10p_e]$	10^3
$02\,dB$	3.1×10^{-2}	0.22×10^{-2}	$[0.82p_e,\ 1.18p_e]$	10^3
$03\,dB$	1.2×10^{-2}	0.11×10^{-2}	$[0.93p_e,\ 1.07p_e]$	10^4
$04\,dB$	3.0×10^{-3}	0.18×10^{-3}	$[0.81p_e,\ 1.19p_e]$	10^4
$05\,dB$	4.7×10^{-4}	0.30×10^{-4}	$[0.89p_e,\ 1.11p_e]$	10^5
$06\,dB$	4.9×10^{-5}	0.38×10^{-5}	$[0.66p_e,\ 1.34p_e]$	10^5
$07\,dB$	4.4×10^{-6}	0.09×10^{-6}	$[0.54p_e,\ 1.46p_e]$	10^6

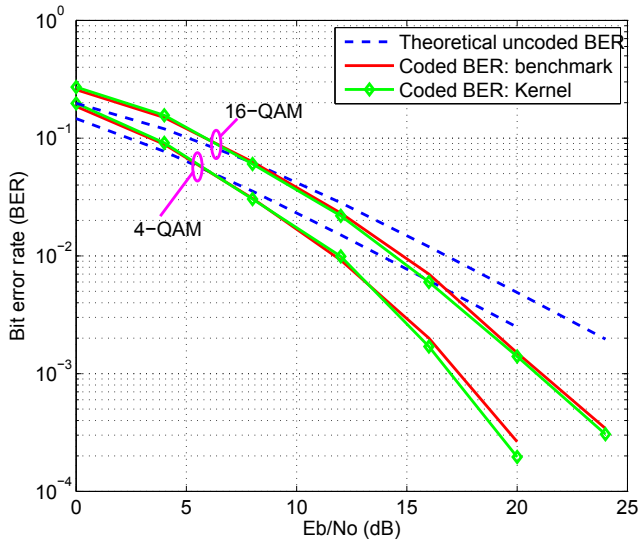


Figure 2. *Performance of the proposed estimator over Rayleigh channel*

$100\,000$, the CPU time engendered over the Rayleigh channel is 33.24 seconds for the MC method against 35.27 seconds for the proposed estimator. However, when the sample size increases it causes the CPU time to increase too. So, the sample saving due to the kernel method is beneficial in terms of power consumption. As an illustration, the performance achieved at $E_b/N_0 = 24\,dB$ (see Table 2) is at the cost of a CPU time of 7.27 minutes for the proposed estimator while being by far greater than 4.35 hours for the MC method.

6. Conclusion

In this paper, we proposed a kernel-based coded bit error rate estimator involving soft M-ary Quadrature Amplitude Modulation (QAM) symbols. An Epanechnikov kernel function was selected. The corresponding smoothing parameter was determined based on the concept of canonical bandwidth. Simulation results were reported for coded 4-QAM and 16-QAM single carrier transmissions over the additive white Gaussian noise channels and for coded multiple carrier modulations over a frequency-selective Rayleigh fading channel. Through curves and numerical data, the proposed kernel-based estimator showed to be, for equal reliability, more efficient than the Monte Carlo estimator. In future works, we will be interested in the possible efficiency improvement that might be achieved if different bandwidth selection strategies were implemented.

Aknowledgements

This work started thanks to the funding of *Projet RESEAU, SCAC-Ambassade de France, Burkina Faso*. Part of it has been performed in the framework of the *FP7 project ICT-317669 METIS*, which is partly funded by the European Union.

Tableau 2. *Numerical results of coded* 16*-QAM BER estimation over Rayleigh channel*

E_b/N_0	Benchmark	Bias	CI	N_k	N_{mc}
$00\,dB$	2.58×10^{-1}	0.13×10^{-1}	$[0.89p_e,\ 1.11p_e]$	1.0×10^3	3.0×10^3
$04\,dB$	1.50×10^{-1}	0.06×10^{-1}	$[0.86p_e,\ 1.14p_e]$	2.0×10^4	1.9×10^4
$08\,dB$	6.28×10^{-2}	0.26×10^{-2}	$[0.87p_e,\ 1.13p_e]$	5.0×10^4	5.1×10^4
$12\,dB$	2.31×10^{-2}	0.12×10^{-2}	$[0.81p_e,\ 1.19p_e]$	5.0×10^4	1.3×10^5
$16\,dB$	7.00×10^{-3}	1.00×10^{-3}	$[0.73p_e,\ 1.27p_e]$	5.0×10^4	1.0×10^5
$20\,dB$	1.50×10^{-3}	0.08×10^{-3}	$[0.67p_e,\ 1.33p_e]$	1.0×10^5	$> 5.1 \times 10^5$
$24\,dB$	3.42×10^{-4}	0.36×10^{-4}	$[0.54p_e,\ 1.46p_e]$	4.1×10^5	$> 2.6 \times 10^6$

7. Bibliographie

[1] S. BENEDETTO, E. BIGLIERI, R. DAFFARA, « Modeling and performance evaluation of non linear satellite links-A volterra series approach », *IEEE Journal on Selected Areas in Communications*, vol. AES-15, 1979, pp. 494-507.

[2] K. YAO, L. B. MILSTEIN, « The use of moment space bounds for evaluating the performance of a non linear digital communication system », *IEEE Transactions Communications*, vol. 31, 1983, pp. 677-683.

[3] M. C. JERUCHIM « Techniques for estimating the bit error rate in the simulation of digital communication systems », *IEEE Journal on Selected Areas in Communications*, vol. 2, n° 1, 1984, pp. 153-170.

[4] S. SAOUDI, T. DERHAM, T. AIT-IDIR, P. COUPE « A Fast Soft Bit Error Rate Estimation Method », *EURASIP Journal Wireless Communications and Networking*, doi :10.1155/2010/372370, 2010.

[5] M. ROSENBLATT, « Remarks on some non-parametric estimates of a density function », *The Annals of Mathematical Statistics*, vol. 27, n° 3, 1956, pp. 832-837.

[6] S. SAOUDI, M. TROUDI, F. GHORBEL, « An Iterative Soft Bit Error Rate Estimation of Any Digital Communication Systems Using a Nonparametric Probability Density Function », *EURASIP Journal Wireless Commun. and Networking*, doi :10.1155/2009/512192, 2009.

[7] S. SAOUDI, T. AIT-IDIR, Y. MOCHIDA, « A Novel Non-Parametric Iterative Soft Bit Error Rate Estimation Technique for Digital Communications Systems », *In : IEEE International Conference on Communications*, 2011, pp. 1-6.

[8] J. DONG, T. AIT-IDIR, S. SAOUDI, « Unsupervised bit error rate estimation using Maximum Likelihood Kernel methods », *In : IEEE Vehicular Technology Conference*, 2012, pp. 1-5.

[9] M. C. JONES, J. S. MARRON, S. J. SHEATER, « A brief survey of bandwidth selection for density estimation », *Journal of the American Statistical Association*, vol. 91, n° 433, 1996, pp. 401-407.

[10] A. Z. ZAMBOM, R. DIAS, « A review of Kernel density estimation with applications to econometrics », *International Econometric Review*, vol. 5, n° 1, 2013, pp. 20-42.

[11] Q. WANG, Q. XIE, Z. WANG, S. CHEN, L. HANZO, « A Universal Low-Complexity Symbol-to-Bit Soft Demapper », *IEEE Transactions on Vehicular Technology*, vol. 63, n° 1, 2014, pp. 119-130.

[12] J.S. MARRON, D. NOLAN, « Canonical kernels for density estimation », *Statistics & Probability Letters*, vol. 7, 1988, pp. 195-199.

8. Appendix

The BER estimate as given in Eq. (12) is

$$\hat{p}_e = \pi_0 \int_0^{+\infty} \frac{1}{n_0} \sum_{j=1}^{n_0} \frac{1}{h_0^*} K\left(\frac{x - X_j}{h_0^*}\right) \mathrm{d}x + \pi_1 \int_{-\infty}^{0} \frac{1}{n_1} \sum_{j=1}^{n_1} \frac{1}{h_1^*} K\left(\frac{x - X_j}{h_1^*}\right) \mathrm{d}x, \tag{A.1}$$

where n_0 (resp. n_1) is the cardinality of the subset of the soft observations among $(X_j)_{1 \leq j \leq N}$ which are likely to be decoded into a binary "0" bit value (resp. "1") and h_0^* (resp. h_1^*) is the selected optimal smoothing parameter which will govern the accuracy of the estimation of $\hat{f}_X^{(0)}(x)$ (resp. $\hat{f}_X^{(1)}(x)$). More explicitly, as $K(x) = \frac{3}{4}\left(1 - x^2\right) I(|x| \leq 1)$, we have

$$\begin{aligned} \hat{p}_e = & \tfrac{\pi_0}{n_0} \textstyle\int_0^{+\infty} \sum_{j=1}^{n_0} \tfrac{3}{4h_0^*} \left[1 - \left(\tfrac{x - X_j}{h_0^*}\right)^2\right] I\left(\left|\tfrac{x - X_j}{h_0^*}\right| \leq 1\right) \mathrm{d}x \\ & + \tfrac{\pi_1}{n_1} \textstyle\int_{-\infty}^{0} \sum_{j=1}^{n_1} \tfrac{3}{4h_1^*} \left[1 - \left(\tfrac{x - X_j}{h_1^*}\right)^2\right] I\left(\left|\tfrac{x - X_j}{h_1^*}\right| \leq 1\right) \mathrm{d}x. \end{aligned} \tag{A.2}$$

Then, using one of the properties of the integral, we get

$$\begin{aligned} \hat{p}_e = & \tfrac{\pi_0}{n_0} \textstyle\sum_{j=1}^{n_0} \int_0^{+\infty} \tfrac{3}{4h_0^*} \left[1 - \left(\tfrac{x - X_j}{h_0^*}\right)^2\right] I\left(\left|\tfrac{x - X_j}{h_0^*}\right| \leq 1\right) \mathrm{d}x \\ & + \tfrac{\pi_1}{n_1} \textstyle\sum_{j=1}^{n_1} \int_{-\infty}^{0} \tfrac{3}{4h_1^*} \left[1 - \left(\tfrac{x - X_j}{h_1^*}\right)^2\right] I\left(\left|\tfrac{x - X_j}{h_1^*}\right| \leq 1\right) \mathrm{d}x. \end{aligned} \tag{A.3}$$

Now, let us set the following changes of variables :

$$\begin{cases} u = \frac{x - X_j}{h_0^*} \\ \\ v = \frac{x - X_j}{h_1^*}. \end{cases}$$

We obtain

$$\begin{aligned} \hat{p}_e = & \tfrac{3\pi_0}{4n_0} \textstyle\sum_{j=1}^{n_0} \int_{-X_j/h_0^*}^{+\infty} \left(1 - u^2\right) I\left(|u| \leq 1\right) \mathrm{d}u \\ & + \tfrac{3\pi_1}{4n_1} \textstyle\sum_{j=1}^{n_1} \int_{-\infty}^{-X_j/h_1^*} \left(1 - v^2\right) I\left(|v| \leq 1\right) \mathrm{d}v, \end{aligned} \tag{A.4}$$

and then,

$$\hat{p}_e = \frac{3\pi_0}{4n_0} \sum_{j=1}^{n_0} \int_{[\alpha_j, +\infty[\cap [-1, 1]} \left(1 - u^2\right) \mathrm{d}u + \frac{3\pi_1}{4n_1} \sum_{j=1}^{n_1} \int_{]-\infty, \beta_j] \cap [-1, 1]} \left(1 - v^2\right) \mathrm{d}v, \tag{A.5}$$

where $\alpha_j = -X_j/h_0^*$ and $\beta_j = -X_j/h_1^*$. Depending on the values of α_j (resp. β_j), three cases are possible among which one leads to zero ; hence we get,

$$\hat{p}_e = \frac{3\pi_0}{4n_0}\left\{\sum_{\substack{\alpha_j<-1,\\ 1\leq j\leq n_0}}\left[t-\frac{t^3}{3}\right]_{-1}^{1}+\sum_{\substack{|\alpha_j|\leq 1,\\ 1\leq j\leq n_0}}\left[t-\frac{t^3}{3}\right]_{\alpha_j}^{1}\right\} + \frac{3\pi_1}{4n_1}\left\{\sum_{\substack{\beta_j>1,\\ 1\leq j\leq n_1}}\left[t-\frac{t^3}{3}\right]_{-1}^{1}+\sum_{\substack{|\beta_j|\leq 1,\\ 1\leq j\leq n_1}}\left[t-\frac{t^3}{3}\right]_{-1}^{\beta_j}\right\}. \tag{A.6}$$

Finally, the BER estimate expression is as follows :

$$\hat{p}_e = \frac{\pi_0 L_0}{n_0}+\frac{\pi_1 L_1}{n_1}+\frac{3\pi_0}{4n_0}\left\{\sum_{\substack{|\alpha_j|\leq 1,\\ 1\leq j\leq n_0}}\left(\frac{2}{3}-\alpha_j+\frac{\alpha_j^3}{3}\right)\right\} + \frac{3\pi_1}{4n_1}\left\{\sum_{\substack{|\beta_j|\leq 1,\\ 1\leq j\leq n_1}}\left(\frac{2}{3}+\beta_j-\frac{\beta_j^3}{3}\right)\right\}, \tag{A.7}$$

where L_0 (resp. L_1) is the cardinality of the subset of $(\alpha_j)_{1\leq j\leq n_0}$ (resp. $(\beta_j)_{1\leq j\leq n_1}$) which are less than -1 (resp. greater than 1).

Management of Low-density Sensor-Actuator Network in a Virtual Architecture

Vianney Kengne Tchendji*, Blaise Paho Nana*

*Department of Mathematics and Computer Science
Faculty of Science
University of Dschang
PO Box 67, Dschang-Cameroon
vianneykengne@yahoo.fr, blaisepaho@gmail.com

RÉSUMÉ. Les réseaux de capteurs sans fi (RCSF) font face à de nombreux problèmes dans leur mise en œuvre, notamment la connectivité des nœuds, la sécurité, l'économie d'énergie, la tolérence aux pannes, le routage [3]. Dans ce document, nous considérons un RCSF peu dense, caractérisé par une mauvaise couverture de la zone d'intérêt, et l'architecture virtuel introduite par Wadaa et al [1] qui permet de partitionner efficacemen ce type de réseau en clusters. Dans l'optique de router optimalement les informations collectés par chaque capteur jusqu'à une station de base (nœud sink, supposé au centre du réseau), nous proposons une stratégie de déplacement des capteurs mobiles (actuators) qui permet de: sauvegarder la connectivité du RCSF, optimiser le routage, économiser l'énergie des capteurs, améliorer la couverture de la zone d'intérêt, etc.

ABSTRACT. Wireless sensor networks (WSN) face many implementation's problems such as connectivity, security, energy saving, fault tolerance and routing problems [3]. In this paper, we consider a low-density WSN where the distribution of the sensors is poor, and the virtual architecture introduced by Wadaa et al [1] which provides a powerful and fast partitioning of the network into a set of clusters. In order to effectively route the information collected by each sensor node to the base station (sink node, located at the center of the network), we propose a strategy to allow mobile sensors (actuators) to move in order to: save connectivity of WSN, improve the routing of collected data, save energy of the sensors, improving the coverage of the interested area, etc.

MOTS-CLÉS : Réseau de capteurs sans fil architecture virtuelle, cluster vides, actuator, routage

KEYWORDS : Wireless sensor network, virtual architecture, empty cluster, actuator, routing

1. Introduction

For few years now, many improvements have been made in domains such as micro-electro-mechanical systems (MEMS) technology [9], wireless communications, and digital electronics. This enabled the development of micro components that easily combine data collection tools and wireless communication devices, and then opens a wide scope to wireless sensor networks (WSN) [3, 5, 8, 11].

Usually called microsensors or simply sensors, these devices with limited resources (bandwidth, computing power, available memory, embedded energy, etc.) have revolutionized traditional networks by bringing the idea to develop sensors networks based on the collaborative effort of a large number of sensors operating autonomously, and communicating with each other via short-range transmissions [6, 7]. These resource limitations added to the radio communication that have sensors, are factors that raise many problems (interference, intrusion, disconnection, data integrity, etc.).
In fact, it is common to see WSN composed of several thousand units [4]. In large networks, the sensors can be grouped into clusters based on their proximity in order to significantly increase the scalability, economy energy, routing, and consequently the lifetime of the network. The structure provided by this partitioning allows the use of various techniques to improve the quality of a WSN, such as data aggregation [10, 11].

In this paper, we consider a low-density WSN where the distribution of the sensors is poor, and the virtual architecture introduced by Wadaa et al [1] which provides a powerful and fast partitioning of the network into a set of clusters. In order to effectively route the information collected by each sensor node to the base station (sink node, located at the center of the network), we propose a strategy to allow mobile sensors (actuators) to move in order to : save connectivity of WSN, improve the routing of collected data, save energy of the sensors, improving the coverage of the interested area, etc.

The rest of this paper is organised as follows : we first present the virtual architecture in which we work, then we present a technique of detecting empty clusters, followed by our method of strengthening strategic points by the actuators, then the technique used to proper move the actuators is presented. A conclusion ends the paper.

2. Virtual architecture sensor network

2.1. Anatomy of a sensor

This is the basic equipment of any WSN. It has three main tasks : information collection from the deployment area, light treatment (optional) on the collected data and sharing these data with other sensors through multi-hop routing. Despite the great diversity (temperature sensors, humidity, pressure, etc.) existing on the market, they are all mounted on the same architectural diagram mainly made of a unit of : capture, processing, storage, communication, and energy. This material may be supplemented or reduced according to the developper [3]. One can for example add a locating system such as a GPS (Global Positioning System), a mobilizer (to get an actuator). The main and optional elements (represented by dashed lines) are shown in figure 1.

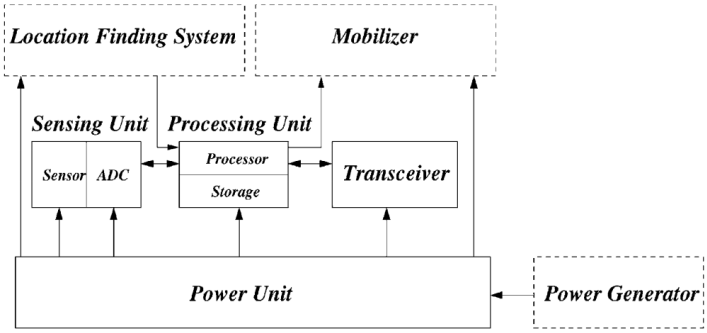
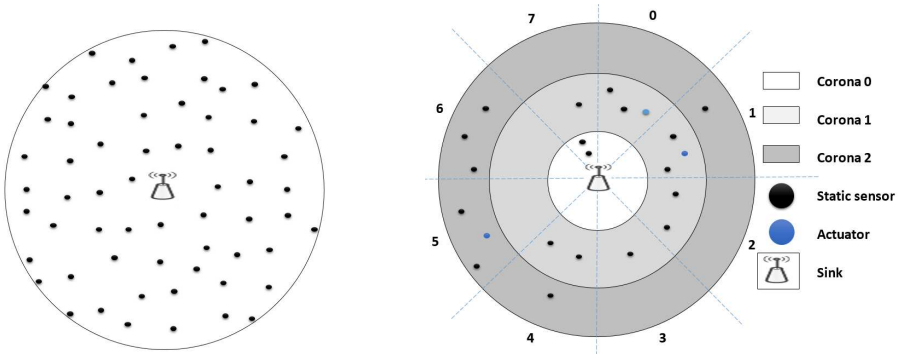


Figure 1 – Hardware architecture of a sensor.

2.2. Virtual network architecture

Let's consider a special sensor called the sink or base station (BS) unconstrained by common sensors's limits and capable of omnidirectional transmissions according to different radius and transmissions at various angles. Once deployed in the supervised area (figure 2a), the sensors can be grouped in clusters (as described in [1]) depending on the corona and the angular sector in which it is located (see figure 2b). Thus, the intersection of the corona i and the angular sector j forms the cluster (i, j). Since the network is sparse, it is important to identify the empty clusters. This allow to have an overview of the area covered by the sensors, and to achieve a better monitoring.



(a) Example of a basic WSN. (b) A sparse virtual architecture.

Figure 2 – An virtual architecture representation.

3. Detection of empty clusters and election of clusterheads

Knowing the distribution of sensors allows the sink to detect empty clusters and build the message propagation tree of figure 3b. For c coronas and s sectors, the sink counts $c \times s$ clusters in its virtual architecture ; therefore sink regularly updates to each message received two tables $h(c, s)$ and $relay(c, s)$. Each entry (i, j) of the table $h(c, s)$ contains 1 if cluster (i, j) is not empty and 0 otherwise, allowing the sink to get a global view of

the sensors's distribution ; and each entry (i, j) of the table $relay(c, s)$ contains the coordinates of the relay cluster of the cluster (i, j). For the rest we advance some hypotheses.

3.1. Hypotheses

The network is fully clustered as shown in figure 2b using the technique described in [1]. Each cluster has a gateway node or clusterhead (CH :node by which a message gets out of the cluster. This reduces redundancy.). Furthermore,

– Each sensor has a unique identifier ID in the network. The sensors are static and form a connected network. Adding or removing a sensor is a rare event ;

– A node is able to estimate its residual energy E_r and the sink has the ability to broadcast messages in the network at different radius, or at different angles ;

– The time is divided into slots of length r, parameters c and s are known to all sensors, and the local clock of each sensor is synchronized with the sink's ;

– A message sent by a sensor reaches all the sensors located in its transmission range after a slot ;

– Clusterisation is made such that all sensors of a given cluster can communicate with each other and some sensors of neighbor clusters.

Here is our protocol, its main lines are taken from S. Faye and J. F. Myoupo [2]. We are introducing the concept of clusterhead (CH).

3.2. Sink's algorithm

The sink periodically broadcasts the date on which the discovery algorithm will begin. All sensors are awake and sink initiates the detection by spreading in the first corona a discovery message $Detect(-1, -1)$. Due to network connectivity, it is certain that at least one sensor will receive this message. During the algorithm, each message transmitted by a sensor towards the sink contains the coordinates of its cluster and those of his relay cluster. Table $h(c, s)$ is initialized to 0. At each received messages from a sensor of the cluster (i, j), BS puts 1 in $h(i, j)$ and, in the entry $relay(i, j)$, it assigns the value contained in the variable *relay* of the received message. At the end of the algorithm, a cluster (i, j) is considered empty when $h(i, j) = 0$ and the relay cluster of the cluster (i, j) is indicated by $relay(i, j)$.

3.3. Sensors's algorithm

The network is supposed connected, so for all cluster (i, j) considered non-empty, there is always a path from it to the sink node. Isolated clusters can't reach the sink and are considered empty even if there are not. There are three main events in the detection of empty clusters : the reception of a message *Detect* asking sensors to indicate their coordinates ; the reception of a message *Head* sent by a node to propose itself as the gateway node and ; the reception of a message *ACK* sent by the sensors to the sink node to indicate their coordinates.

– **Reception of a message** $Detect$: On the first slot, after receiving a message $Detect(-1, -1)$ from BS, the sensors of the first corona build a message $Detect(1, j)$, and broadcast it to allow sensors in other clusters to reveal their presence. Then, those that received the message $Detect$ from a neighbor cluster and have an energy higher than the threshold E_s send a message $Head$ towards their own cluster to be elected as the CH.

– **Reception of a message** $Head$: At the reception of a message $Head$, the sensor saves the identifier of the CH in its variable $gatewayNode$ if this is its first reception, otherwise it compares the residual energy E_r of $gatewayNode$ with that of the received message and stores the one that has the biggest E_r. If the E_r are equal, the one with the highest ID is chosen. A sensor that has already received a message $Head$ can't send message $Head$, because it would have received its message $Detect$ from a more distant neighbor cluster to the sink than the relay sensor that sent him the first message $Head$. Finally, the sensor that is elected in the cluster builds a message $ACK((i,j),(-1,-1))$ and sends towards the sink node which is actually its relay cluster. This process is repeated for all the other sensors untill the most distant cluster sends its message ACK.

– **Reception of a message** ACK : A sensor that receives a message ACK from the neighbor node checks whether this message is for its cluster. In this case, it sends it to his gateway node which checks if it has not already routed a message from the same cluster. If not, it broadcasts it towards its relay cluster. Otherwise, it simply ignores it.

4. Searching and filling strategic empty clusters by actuators

This section is once again inspired from [2]. Here we introduce the actuators : sensors with a mobilizer, allowing them to move on the sink's order. They can be used for many purposes, depending on the user, for example :

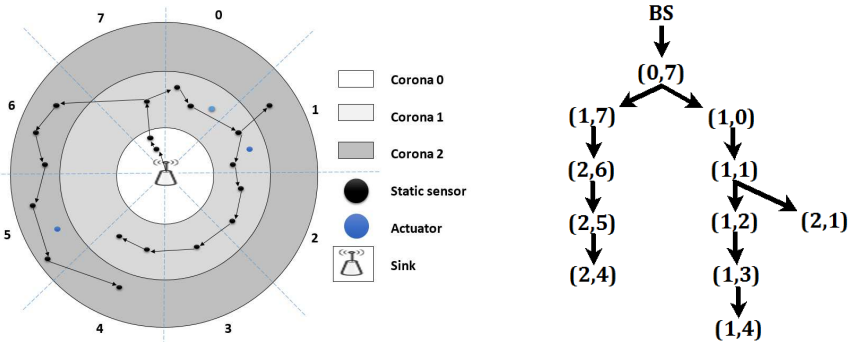
– Being the CH in a cluster where all the sensors have a small energy ;

– Collect and route information in isolated areas ;

– Connect a sub isolated connected network to the main network.

– Being sent in strategic empty clusters (purpose of this paper) to optimize the routing.

From the table $h(c,s)$ the sink knows the empty clusters. In order to know which ones it is going to fill first, it should reproduce the messages spreading tree like in figure 3b by using the two tables it has like this : Take BS as root and the first tree's leaf. As long as there's an unvisited leaf (i,j), search in the table $relay(c,s)$ the cluster that has the cluster (i,j) as relay cluster and add them as the sons of (i,j).



(a) Propagation of messages in the supervised area. (b) Messages spreading tree.

Figure 3 – The messages spreading tree obtained after the detection of the empty clusters.

Filling a strategic empty cluster has the effect of reducing the tree's height (figure 3b). The ideal one would be to reduce this size to the number of coronas of our virtual architecture. The routing will be optimal if for any cluster of corona k, the transmission of a message towards BS passes through k intermediate clusters.

4.1. Rule of detection of clusters which access can be improved

To optimize the routing to a cluster, it would be good to know whether the current access is optimizable. In figure 3, it is the case of cluster $(1, 4)$ which is at 5 intermediate clusters from BS instead of 1 if an actuator were placed in cluster $(0, 3)$.

Rule : Let $\mathbb{A}$ a message spreading tree similar to that of figure 3b, $prof(i, j)$ denotes the depth of the cluster (i, j) in the tree $\mathbb{A}$. The path from sink to (i, j) can be improved if there is another cluster (i', j') with depth $prof(i', j')$, such as $i' \geq i$ and $prof(i', j') < prof(i, j)$, i.e. (i', j') is in a corona greater or equal to (i, j)'s but in the tree $\mathbb{A}$, (i', j') appears at a depth less than (i, j)'s.

4.2. Detection of strategic empty cluster to fill in priority

To determine this priority cluster (PC), we establish for every corona a the list $L[a]$ of clusters of this corona which access can be improved (C coronas = C lists). Each list $L[a]$ contains the coordinates (a, j) of clusters of the corona a which access can be improved. It is in the form : $L[a] = [(a, j_1), (a, j_2), \ldots, (a, j_n)]$. From each list $L[a]$, *we extract the longest list* $L_s[a]$ *made of consecutive clusters of* $L[a]$. In the list $L_s[a]$, each (a, j) represents the coordinates of the clusters of corona a that follows in the message spreading tree. The coordinates (x, y) of PC are deduced from the longest sub list $L_{sp}[a]$ taken among the $L_s[a]$ extracted lists. $x = a - 1$ and y is equals to the default rounding average of j (the j are the second components items (a, j) of the list $L_{sp}[a]$).
As long as there's available actuators, it is necessary to move an actuator at the cluster (x, y), another at the cluster $(x - 1, y)$ if it is empty, ... another at the cluster $(0, y)$ if it is empty. The process can be repeated until the routing is optimal, i.e. up to $prof\mathbb{A} \leq C$.
For the example of figure 3, the determination of PC is presented in annex A.

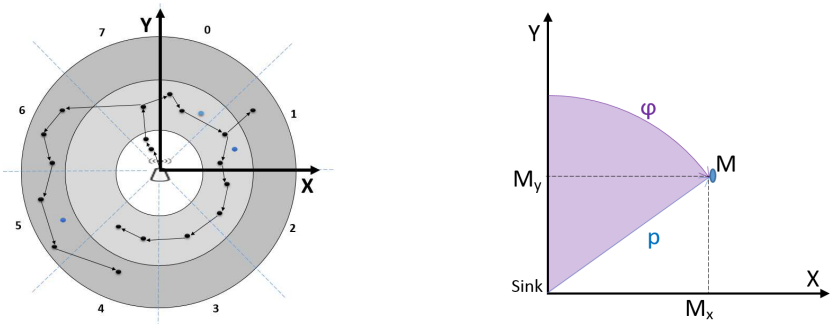
Before moving a mobile sensors, the sink node must calculate the distances from it to the target empty cluster, it is better to choose the most appropriate actuator, based on distance, residual energy, availability, etc.

5. Moving a mobile sensor

We propose to move an actuator from the cluster (x, y) to the empty cluster (j, k). The actuator will need the distance and direction to move. To simplify our calculations, we should be in a cartesian plane. For this we describe here how to transform our current coordinate system (Dynamic Coordinate System : DCS) in a Polar Coordinate System (PCS) and then in to a Cartesian Coordinate System (CCS).

5.1. Correspondence between DCS, PCS and CCS

To get the distance and the direction to follow, we must define a reference. Thus the origin of our reference will the sink. The Y axis is taken such that it coincides with the left edge of the first section (section 0). The X axis is at a quarter turn from the Y axis so that the angle $X, \widehat{sink}, Y$ is direct (figure 4a). Any point of the DCS is discoverable using p (its distance from the sink) and φ (angle measured from the Y axis) as in figure 4b.



(a) Definitio of the reference (b) PCS - CCS

Figure 4 – Correspondence between the DCS - PCS - CCS

Denote 'α' the angle of a sector and 'e' a thickness of a corona, we can state corollary 1.

Corollary 1 *Let M be a sensor of the cluster (c, s) assumed at it center. In the PCS, M has the coordinates (p, φ) where $p = c \times e + \frac{e}{2}$ and $\varphi = s \times \alpha + \frac{\alpha}{2}$;*
Let $M(p, \varphi)$ be a point in the PCS. In the CCS, M has coordinates (x, y) with $x = M_x = p\sin(\varphi)$ and $y = M_y = p\cos(\varphi)$.

5.2. Moving the actuator

Now we can start the necessary calculations (distance p and angles φ) to move the actuators.

5.2.1. Calculation of the distance (p)

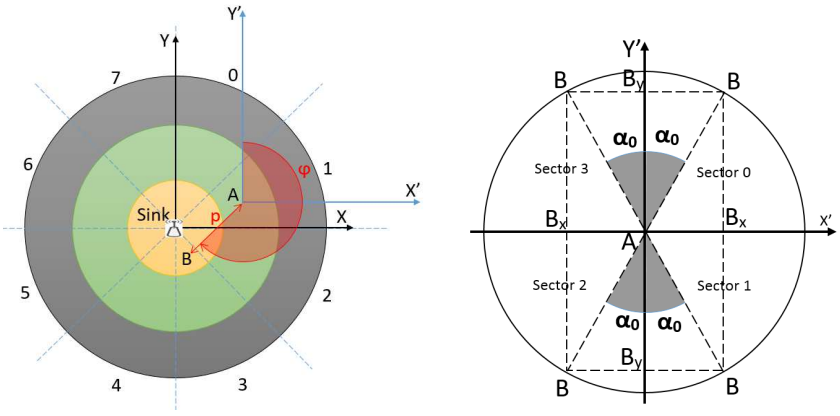
The distance between two points A and B of the plane is given by the norm of the vector $\overrightarrow{AB}$, denoted $||\overrightarrow{AB}||$ or just AB. According to figure 4b, an actuator that moves from the sink node (with cordinates $(0, 0)$) to the point M (with cordinates (x, y)) must cover the distance $p = ||\overrightarrow{sink\ M}|| = ||(x - 0, y - 0)|| = ||(x, y)||$. But $||\overrightarrow{sink\ M}||^2 = (sink\ M_x)^2 + (sink\ M_y)^2 = x^2 + y^2$. Thus $||\overrightarrow{sink\ M}|| = \sqrt{x^2 + y^2}$. So we have corollary 2.

Corollary 2 *Moving an actuator from the center of the cluster $A(c_1, s_1)$ of polar coordinates (p_1, φ_1) to the center of the cluster $B(c_2, s_2)$ of polar coordinates (p_2, φ_2) returns to move this actuator from the point $A(A_x, A_y)$ to the point $B(B_x, B_y)$ of the cartesian coordinates system on a distance $p = ||\overrightarrow{AB}||$ where $\overrightarrow{AB}(B_x - A_x, B_y - A_y)$, $p = ||\overrightarrow{AB}|| = \sqrt[2]{(B_x - A_x)^2 + (B_y - A_y)^2}$, $A_x = p_1\cos(\varphi_1)$, $A_y = p_1\sin(\varphi_1)$, $B_x = p_2\cos(\varphi_2)$ and $B_y = p_2\sin(\varphi_2)$.*

5.2.2. Calculation of the angle (φ)

To facilitate the calculation of the value of φ, let's make a change of reference.

Change of reference : We want to move a mobile sensor from a point A to a point B. For this, we define a new reference in which the base vectors are collinear with those of the previous (see figure 5a).



(a) Reference change

(b) Angle between the axis Y' and the displacement vector

Figure 5 – Angle displacement

The reference $R_1 = (sink, \vec{i}, \vec{j})$; $R_2 = (A, \vec{i}, \vec{j})$; $R_2 = t_{\overrightarrow{sink\ A}}(R_1)$ where t denotes the translation of vector $\overrightarrow{sink\ A}$. The coordinates of two points $M(x, y) \in R_1$, and $M'(x', y') \in R_2$ such that $M' = t_{\overrightarrow{sink\ A}}(M)$ are now linked by the following relations : $x' = x - x_A$ and $y' = y - y_A$.
Basis vectors of these two referenecs are pairwise collinear, that's why the angles found in one of the refeces will be equivalent in the second.

Calculation of the inclination α_0 ***formed by the displacement vector and the*** Y' ***axis :*** The displacement angle φ that we want to calculate is strongly related to the angle α_0 formed by the vector $\overrightarrow{AB}$ and the Y' axis. Figure 5b presents the different situations we may encounter. We deduced that $\sin(\alpha_0) = \frac{|B_x|}{AB} \Rightarrow \alpha_0 = \sin^{-1}\left(\frac{|B_x|}{p}\right)$ where $p = AB$.

Determination of the displacement angle φ ***:*** From figure 5b, point B can be found in one of the four sectors.

Corollary 3 *The displacement angle φ of an actuator from the point $A(A_x, A_y)$ to the point $B(B_x, B_y)$ of $(A, \vec{i}, \vec{j})$ reference is given by :*

1) if B is in the sector 0*, i.e. $B_x > 0$ and $B_y \geq 0$ then $\varphi = \alpha_0$*

2) if B is in the sector 1*, i.e. $B_x \geq 0$ and $B_y < 0$ then $\varphi = \pi - \alpha_0$*

3) if B is in the sector 2*, i.e. $B_x < 0$ and $B_y \leq 0$ then $\varphi = \pi + \alpha_0$*

4) if B is in the sector 3*, i.e. $B_x \leq 0$ and $B_y > 0$ then $\varphi = 2\pi - \alpha_0$*

For a practical example of moving a sensor, see annex B.

6. Simulation and analysis of our solution

6.1. Tools and simulation environment

Using an HP computer Intel (R) Core (TM) i7-2630QM CPU @ 2.00 GHz $\times$ 8, 8GB of RAM, running Windows 8 Professional ; a discrete event network simulator J-Sim ; and a sample of 1000 sensors randomly deployed within 10 km of the sink ; the virtual architecture has 10 coronas and 8 sectors of 45^o each. We performed repeatedly tests and averages the results. The energy model is the one adopted by many efficient contributions [13] : $E = E_{trans} + E_{recep}$. E_{trans} and E_{recep} are respectively the total energy used for transmissions in the network and receptions, knowing that each sensor has a range of 500 meters, and initial energy of 100 joules. He needs $35.28 \times 10^{-3} joule$ per transmission and $31.32 \times 10^{-3} joule$ per reception. The curves were made with version 5.0 of gnuplot software.

6.2. Analysis of the simulation results

Figure 6 compares the energy consumption of the cluster-heads and ordinary sensors when the routing is not optimized and when routing is optimized with actuators. An economy of energy is observed among both ordinary sensors and cluster-heads sensors. This increases the longevity of the network. The simulation is made for the detection of empty clusters, the election of cluster-head and routing. A slot is $78\mu s$. The great loss of energy observed at the beginning of the curve is due to the fact that the cluster-head are not elected at the beginning. It is clear that this energy loss is significantly reduced once the cluster-head are elected.

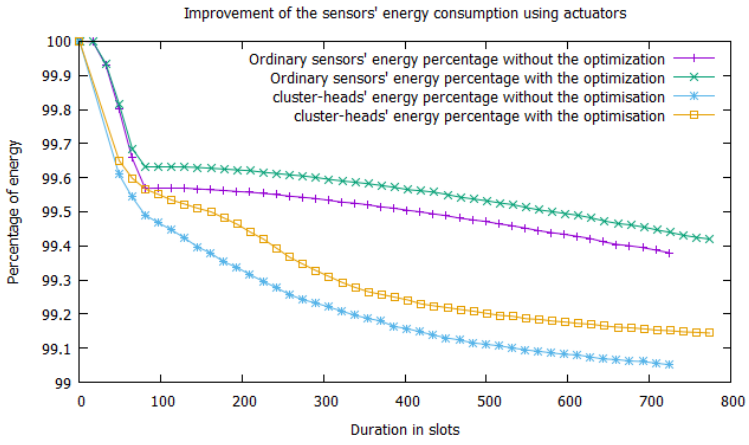


Figure 6 – Improvement of the power consumption using cluster-heads and actuators.

7. Conclusion

In this paper we have presented a virtual architecture that facilitates the management of WSN. We also introduced the gateways nodes, their election protocol and how to limit redundant messages through them. But our main aim was to optimize the routing of

collected data towards de sink. That's why we started by describing a method of detecting and filling strategic empty clusters in which we can send mobile sensors. We also present other possible utilities of the actuators, and show a new way of performing their movements to improve the routing of the collected data.

In a very soon future we intend to work on the mechanism of reelection of the clusterhead in a cluster ; the mechanism of changing relay cluster if the current relay cluster is no longer accessible and the mechanism of redirection of packets routing after the positioning of the mobile sensor.

8. Bibliographie

[1] A. WADAA, S. OLARIU, L. WILSON, M. ELTOWEISSY, K. JONES, « Training a wireless sensor network », *Mobile Networks and Applications*, Vol. 10, N° 1-2, p. 151–168, 2005.

[2] SÉBASTIEN FAYE, JEAN-FRÉDÉRIC MYOUPO, « Deployment and Management of Sparse Sensor-Actuator Network in a Virtual Architecture », *International Journal of Advanced Computer Science*, Vol. 2, N° 12, December, 2012.

[3] I. F. AKYILDIZ, WEILIAN SU, Y. SANKARASUBRAMANIAM, E. L. CAYIRCI, « A survey on sensor networks », *IEEE Communications Magazine*, Vol. 40, N° 8, p. 102–114, 2002.

[4] B. WARNEKE, M. LAST, B. LEIBOWITZ AND K. PISTER, « Smart Dust : communicating with a cubic-millimeter computer », *IEEE Computer*, Vol. 34, N° 1, p. 44–51, 2001.

[5] S. TILAK, N.B. ABU-GHAZALEH, W. HEINZELMAN, « A taxonomy of wireless micro-sensor network models », *Mobile Computing and Communications Review*, Vol. 6, N° 2, p. 28–36, 2002.

[6] MARK A. PERILLO, WENDI B. HEINZELMAN, « Wireless Sensor Network Protocols », *Algorithms and Protocols for Wireless and Mobile Networks, Eds. A. Boukerche et al., CRC Hall Publishers*, 2004.

[7] K. SOHRABI, J. GAO, V. AILAWADHI, G. J. POTTIE, « Protocols for Self-Organization of a Wireless Sensor Network », *IEEE personal communications*, Vol. 7, N° 5, p. 16–27, 2000.

[8] C. INTANAGONWIWAT AND R. GOVINDAN AND D. ESTRIN, « Directed Diffusion : a scalable and robust communication paradigm for sensor networks », *ACM Press*, p. 56–67, 2000.

[9] B. WARNEKE, K.S.J. PISTER, « MEMS for Distributed Wireless Sensor Networks », *9^{th} International Conference on Electronics, Circuits and Systems, Croatia*, Vol. 1, p. 291–294, 2002.

[10] S. FAYE, J. F. MYOUPO, « An Ultra Hierarchical Clustering-Based Secure Aggregation Protocol for Wireless Sensor Networks », *AISS : Advances in Information Sciences and Service Sciences*, Vol. 3, N° 9, p. 309 – 319, 2011.

[11] A. PERRIG, R. SZEWCZYK, V. WEN, D. CULLER, J.D. TYGAR, « SPINS : Security protocols for sensor networks », *Wireless networks*, Vol. 8, N° 5, p. 521–534, 2002.

[12] C. KARLOF, N. SASTRY AND D. WAGNER, « TinySec : A Link Layer Security Architecture for Wireless Sensor Networks », *in : Proc. of the 2nd international conference on Embedded networked sensor systems*, ACM, p. 162–175, 2004.

[13] D. WEI AND S. KAPLAN AND H. A. CHAN, « Energy Efficient Clustering Algorithms for Wireless », *in : Sensor Networks, Proceedings of IEEE Conference on Communications, Beijing*, IEEE, p. 236–240, 2008.

A. Practical example for determination of priority clusters

With the example of figure 3, the determination of priority clusters is performed as follow :
Lists construction :
$L[0] = \emptyset$; $L[1] = [(1,1),(1,2),(1,3),(1,4)]$; $L[2] = [(2,1),(2,4),(2,5)]$
Extraction of clusters sublists which follow :
$L_s[0] = \emptyset$; $L_s[1] = [(1,1),(1,2),(1,3),(1,4)]$; $L_s[2] = [(2,4),(2,5)]$
The longest sub-list : $L_{sp} = L_s[1] = [(1,1),(1,2),(1,3),(1,4)]$
Calculation of the coordinates (x, y) :
$x = 1 - 1 = 0$ and $y = floor(Average(1,2,3,4)) = floor(2.5) = 2$
Filling stategic clusters : The cluster $(0, 2)$ is free, an actuator should be sent into it. By repeating the process : We should put an actuator (if there are available) in the cluster $(0, 4)$ ...

B. Practical example of moving a mobile sensor

Let's move an actuator from the cluster $A(1, 1)$ to the cluster $B(0, 3)$. For our tests, let's suppose : the scope of the sink is $30.0m$ and the virtual architecture includes 3 coronas with $e = 10.0m$ each ; there are 8 angular sectors of $\alpha = \frac{\pi}{4}$ rad each.

Polar Coordinates : A(15.0, 1178 rad) and B(5.0, 2.749 rad)

Cartesian coordinates : $A(13858, 5740)$ and $B(1913, 4619)$

Distance : $p = ||\overrightarrow{AB}|| = 15.811$

Displacement angle : $B'_x = B_x - A_x = -11.945$ and $B'_y = B_y - A_y = -10.359$, then $\alpha_0 = 0.856$ rad. Since $B'_x < 0$ and $B'_y <= 0$ then B is in sector 2 and thus, $\varphi = \pi + \alpha_0 = 3.998$ rad or $\varphi = 229.065^{\circ}$.

Conclusion : To strengthen the area $(0, 3)$, the sink node asks the actuator of the cluster $(1, 1)$ to cover a distance $p = 15.811m$ with an angle of $\varphi = 229,065^{\circ}$.

Centre of Mass of single coverage: A comparative study with Simulated Annealing for mesh router placement in rural regions

Jean Louis Fendji Kedieng Ebongue* and Christopher Thron**

*The University of Ngaoundéré - CAMEROON

lfendji@univ-ndere.cm

** Texas A&M University Central Texas - USA

thron@ct.tamus.edu

RÉSUMÉ. Ce travail s'attaque à un problème critique dans la planification de réseaux maillés sans-fil pour zones rurales : le placement de nœuds maillés. Le but est de maximiser la couverture tout en réduisant autant que possible le nombre de nœud dans le réseau et en assurant la connectivité. Pour atteindre cet objectif, nous proposons une approche basée sur le calcul du barycentre de la zone couverte par un seul routeur. Cette approche est dix fois plus rapide que l'approche basée sur le recuit simulé. En outre, les simulations ont aussi montré une faible variation des solutions, traduisant ainsi une certaine stabilité de l'approche. Toutefois, la qualité des solutions obtenues en termes de couverture des zones d'intérêt avec le recuit simulé reste meilleure.

ABSTRACT. This paper tackles a critical issue in the planning of rural wireless mesh network (RWMN): the mesh node placement. The aim in the planning of RWMN is to maximise the coverage while keeping the number of router as few as possible and ensuring the connectivity. To achieve this, we proposed an approach based on the calculation of the centre of mass of areas covered per router. This approach is ten times more time-efficient than the simulated annealing one. In addition, the simulations results also provide a low variation of the solutions, showing some stability of the approach. However, the quality of the solution in terms of coverage of areas of interest provided by the approach based on Simulated Annealing is better.

MOTS-CLÉS : Centre de Masse, Recuit Simulé, Réseaux maillés sans fil, Placement de router maillé.

KEYWORDS: Centre of mass, Simulated Annealing, Wireless Mesh Networks, Mesh router placement.

1. Introduction

A Wireless Mesh Network (WMN) [1] is a wireless network in which nodes are connected in a mesh topology. This kind of network is an appealing cost-effective solution to bridge the digital divide observed between rural and urban regions, since it is based on off-the-shelf material especially WiFi technology.

Rural Wireless Mesh Networks (RWMN) are usually composed of one gateway which connects the network to Internet, and a set of mesh routers (MRs). The success of the planning of such networks depends on the determination of an optimal number and placement of its mesh nodes. The planning of wireless networks in rural regions is more coverage-driven than capacity-driven [2], with the aim of minimizing the overall cost of the architecture, while maximizing the coverage percentage of the area to cover.

For realistic deployment scenarios, the problem of mesh node placement is a NP-hard combinatorial optimization problem which cannot be solved in polynomial time [9], [10]. This is why metaheuristics are usually required to optimize the planning.

This paper considers the network model found in [3]. In this model, a given area to cover is decomposed into elementary areas which can be required or optional in terms of coverage and where a node can be placed or not. An extension is made to this model in order to consider the presence of obstacles that can hinder the connectivity. The aim is therefore to determine the location of mesh routers which maximizes the coverage of area of interest. To achieve this goal, a placement approach based on the calculation of the centre of mass (CM) of area covered per router is proposed. This approach is compared to the simulated annealing (SA) approach defined in [4] to solve the same problem.

The rest of the paper is organized as follows: Section 2 briefly presents related work in WMN planning. Section 3 defines the network model and formulates the placement problem. Section 4 explains the approach based on the calculation of the centre of mass of area covered alone by a router. Section 5 presents the experimental setup and discusses the results in comparison with simulated annealing ones. This paper ends with a conclusion and future work.

2. Related Work

The work in [5] provides a good overview of the planning problem in WMN. This survey classifies the planning problem according to the flexibility of the network topology: unfixed (not-predefined) and fixed (predefined). In fixed topology, all the nodes in the network have a predefined location. The problem is therefore more related

to routing protocols, channel assignment, or joint approaches. In unfixed topology, the location of at least some nodes is not predefined in the network: either the gateway(s) or the mesh routers, or both. This problem is usually assimilated to the one of facilities and locations with mesh routers representing facilities and the users to serve representing locations.

To solve the placement approach, different formulations have been proposed in the literature. They depend on the type of node considered in the planning problem: mesh routers [6], gateways(s) [7], or both [8]. Linear programming based approaches [9] have been used; but since this problem is known to be hard for real size deployment [9], search techniques and meta-heuristic are usually used [6, 10, 11, 12]. The region to be covered, usually called the universe, can be considered as continuous (a whole region), discrete (a set of predefined positions) or network (undirected weighted graph).

In [10], an approach based on simulated annealing has been proposed to solve the mesh nodes placement problem. It aims to find optimal locations of routers that maximize the network connectivity and client coverage, given a two-dimensional area with a number of fixed client nodes.

The work in [13] introduces the placement problem of mesh routers in a rural region. It has been extended later in [3], wherein a region is considered as decomposed into a set of elementary areas which may require the coverage or where a node may be placed. A placement approach based on metropolis algorithm has been therefore used.

3. Formulation of the Placement Problem

A given region is composed of areas of interest that should be covered as it is in [4]. The coverage of a region is considered as optional when this region is not of interest. A given region comprised also prohibited areas where a node cannot be placed (lake, river, road…), and a set of obstacles that could hinder the connectivity.

The area to cover is modelled as a two-dimensional irregular form in a two-dimension coordinate plane. We consider the smallest rectangle that can contain the irregular form. Therefore, we assume that this rectangle is decomposed into small square forms. Each discrete point is called elementary area (EA), which can be of one or more types: Elementary Area of Interest (EAI); Non-line-of-sight Elementary Area (NEA); or Prohibitive Elementary Area (PEA).

We define a set of two-dimensional matrices in order to characterise each EA: *Cover* indicating whether an EA requires coverage; *Place* indicating whether we can place a

node in an EA; *CoverDepth* indicating the number of routers covering an EA; and *Pathloss* indicating whether an EA contains an obstacle. Therefore, an EA at position (x, y) can be characterised by (1-4).

$$\text{Cover}(x, y) = \begin{cases} 0 \rightarrow \text{coverage not required} \\ 1 \rightarrow \text{coverage required} \end{cases} \quad (1)$$

$$\text{Place}(x, y) = \begin{cases} 0 \rightarrow \text{cannot place a node} \\ 1 \rightarrow \text{can place a node} \end{cases} \quad (2)$$

$$\text{CoverDepth}(x, y) = \begin{cases} 0 \rightarrow \text{no coverage} \\ n \rightarrow \text{covered by n routers} \end{cases} \quad (3)$$

$$\text{Pathloss}(x, y) = \begin{cases} 0 \rightarrow \text{no obstruction} \\ p \rightarrow \text{attenuation factor} = p \end{cases} \quad (4)$$

To simplify the problem, we assume that the attenuation factor of any obstacle in the line of sight between two routers is high enough to prevent any wireless link between those routers. We also assume that all routers are equipped with an omnidirectional antenna all having the same coverage radius (r). The radius is expressed as the number of EAs (r = 6 means that the radius stretches over 6 EAs).

Let p be an EA at position (x, y). If a mesh node is located in p, then the set of EAs covered by this mesh node is given by (5).

$$\forall (a, b), (x - a)^2 + (y - b)^2 < r^2 \quad (5)$$

The population is not so dense in rural regions when comparing to urban ones; thus, we consider as in [2] that the planning is coverage-driven, meaning we are more concern by the space to cover than the throughput to deliver. The mesh router placement problem in rural regions can be therefore described as the determination of a minimum set of positions, which maximizes the coverage of areas of interest, while minimising the cost of the architecture and ensuring the connectivity. This cost can be minimised just by minimising the number of routers required to cover the region.

4. Centre of mass of single coverage

4.1. Algorithm

The idea behind the approach of the centre of mass of single coverage is to reduce the area covered by multiple routers by drawing routers to the centre of mass of area they are covering alone. This approach is motivated by the fact that by moving routers to the centre of mass of their single coverage, new non-covered EAI can be reached in a relative short number of moves compared to the number of moves required by the SA approach. In fact, in the SA approach, the location where to move a selected router is chosen randomly while ensuring that $Cover$ = 1, and $Place$ = 1. The SA approach is given in Appendix 1.

Algorithm 1: Centre of mass of single coverage

Input: f : the objective function to be maximized

Output: s: the best solution found

```
1  Begin
2      s:= InitialSolution();
3      v:= f(s);
4      while (stopping condition not met) do
5          i:=selectARouter();
6          if multiple coverage of i is too large a fraction then
7              Search for an EA with CoverDepth = 0, Cover = 1, and Place = 1
8          else
9              Move i to the centre of mass of his single coverage
10         s := NewSolution(i);
11         v := f(s)
12     returns
13 End
```

4.2. Algorithm explanation

Initial Solution: The initial solution is obtained by placing routers randomly in the area to cover while ensuring that $Cover$ = 1, and $Place$ = 1. For each router we randomly select an EA until Cover(EA)=1 and Place(EA)=1be satisfied. We therefore place the current router in this EA. A minimal number of routers for covering a given region can be determined by (6). But this minimal number is not enough to cover the region since routers should overlap to ensure the connectivity, and the form of the region is irregular. We use an initial number of routers $1.5*nr_{min}$.

$$\mathrm{nr}_{\min} = \left\lceil \sum \mathrm{Cover}(x, y) \ \ /(r^2 * 3.14) \right\rceil \quad (6)$$

Single and multiple coverage: Let us consider $sCov(i)$ and $mCov(i)$ to be respectively the single coverage and the multiple coverage of router$i$. To check whether multiple coverage is too large a fraction, we use the expression in (7). In this expression, $rand(x)$ is used to provide some probability. We can remark that when $sCov(i)$is too great compared to $mCov(i)$, expression in (7) has a great probability to be not satisfied. If it is the case, the router is moved to the centre of mass of its single coverage; reducing eventually its multiple coverage. Otherwise it is relocated to another EA selected randomly. However, the EA should be one that requires coverage, which is not yet covered, and where a node can be placed.

$$\left(sCov(i) + mCov(i)\right)^2 * rand(x) < (mCov(i))^2 \quad (7)$$

Fitness function (lines 3 and 11): The evaluation of fitness function consists to count the number of covered EAI. This is done by (8) after the initialisation. Because we move only one router at the same time, we consider only the EAs of this router which are concerned by the move.

$$f = \sum \mathrm{sign}(\mathrm{CoverDepth} .* \mathrm{Cover}) \quad (8)$$

New Solution (line 10)**:** It is obtained by keeping other routers in their previously positions and considering the new position of router i.

Stopping condition: If the value of the fitness function does not improve after a certain number of iteration (nbtostop), we suppose therefore having reached the optimal.

5. Simulation results

To compare the proposed approach with SA approach, we randomly generate a region with areas of interest and prohibitive areas. We consider a grid of 100x100 with nbtostop=1000 and r=6.The unit is the size of an EA. If size (EA)=20m, the radius will be r=120m, and the grid 2km x 2km=4km². This is realistic since 802.11n routers have a theoretical outdoor transmission range of 250m. We use a number of routers between $1.5*nr_{min}$ and nr_{min} ($1.5*nr_{min}$,$1.4*nr_{min}$, $1.3*nr_{min}$, $1.2*nr_{min}$, $1.1*nr_{min}$, nr_{min}). For each number of routers, the two algorithms are run ten times. Both approaches are compared according to the CPU time used for computation, the quality of solutions in terms of coverage percentage of area of interest, and the ability to provide similar results. Tables 1 to 4 in Appendix 2 provide the results of the simulation phase conducted using Scilab 5.4.

Figure 1 provides the coverage percentage of both approaches. In this figure we can observe that the SA approach provides better solutions than the centre of mass (CM) approach in terms of coverage percentage. This can be explained by the fact that in the SA approach, when the temperature is close to the minimal one, the hop distance is reduced, allowing reaching better positions that improve the quality of the solution. But in CM approach, routers are eventually moved to their centre of mass of single coverage.

Another observation concerns the ability to provide similar results by both approaches. We observe a great difference between the best and the worst coverage percentage with the SA approach. For instance, with the number of routers $nr=1.2nr_{min}$, we observe a variation of about 8% between the maximum and the minimum coverage. But in the CM approach, for each run, the maximum is close to the third quartile while the minimum is close to the first quartile, with those quartiles close to each other. This expresses some ability of CM approach to provide similar results. Finally concerning the CPU time used, the CM approach in all configurations are in average ten times more efficient than SA approach, as we can observe in Figure 2. This is important when we are dealing with online optimisation in which we would like to observe a solution in very short time.

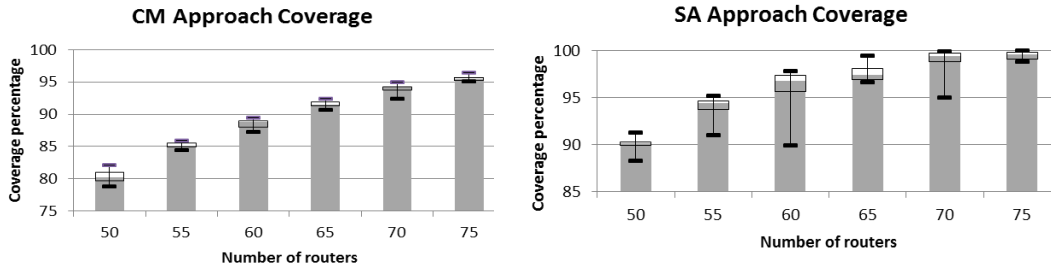


Figure 1: Coverage percentage provided by CM and SA approaches

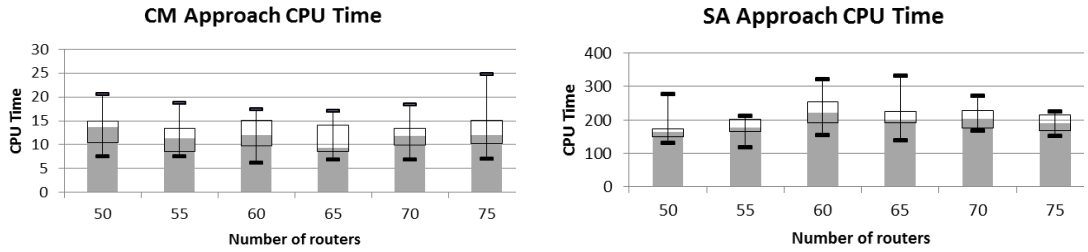


Figure 2: CPU time used by CM and SA approaches

6. Conclusion and future work

This paper has introduced a new approach based on the calculation of the centre of mass (CM) for the placement of mesh nodes in rural wireless mesh networks. This approach has been compared to simulated annealing. Simulation results have shown a rapid convergence of CM approach compared to SA. In fact CM is in average ten times faster than SA. This is suitable for online optimisation problems where convergence time should be minimised. We also observed an ability of CM approach to provide similar solutions when comparing to SA. However, SA approach provides better solutions.

Further investigation will be conducted to design a new approach combining CM and SA approaches in order to take advantage of the stability and the rapid convergence of CM approach, and the quality of solutions in terms of coverage percentage provided by SA approach. The new approach could be also used for the problem of sensor placement in wireless sensor network.

7. Bibliography

[1] I. F. Akyildiz, X. Wang, and W. Wang. Wireless mesh networks: a survey. Computer Networks 47(4) (2005), 445-487.

[2] Bernardi, G., Marina, M.K.,Talamona, F., Rykovanov, D.: IncrEase: A tool for incremental planning of rural fixed Broadband Wireless Access networks. In: IEEE Global Telecommunications Conference (GLOBECOM 2011), pp. 1013-1018.

[3] Fendji, J.L.E.K., Thron, C., Nlong, J.M.: A Metropolis Approach for Mesh Router Nodes placement in Rural Wireless Mesh Networks. Journal of Computers.10 (2), pp. 101-114, (2015).

[4] Fendji, J. L., Thron,C., and Nlong, J. M.: Simulated Annealing approach for mesh router placement in rural Wireless Mesh Networks. 7th International Conference, AFRICOMM 2015, Cotonou, Benin, December 15-16, 2015.

[5] Benyamina, D., Hafid, A.,Gendreau, M.: Wireless mesh networks design—a survey. IEEE Communications Surveys & Tutorials 14(2), pp. 299-310 (2012).

[6] Xhafa, F., Sánchez, C., Barolli, L.: Genetic algorithms for efficient placement of router nodes in wireless mesh networks. In: 24th IEEE International Conference onAdvanced Information Networking and Applications (AINA), pp. 465-472(2010).

[7] Li, F., Wang, Y., Li, X. Y., Nusairat, A., Wu, Y.: Gateway placement for throughput optimization in wireless mesh networks. Mobile Networks and Applications, 13(1-2), 198-211(2008).

[8]De Marco, G. (2009, September). MOGAMESH: A multi-objective algorithm for node placement in wireless mesh networks based on genetic algorithms. In 6th IEEE International Symposium onWireless Communication Systems(ISWCS 2009), pp. 388-392 (2009).

[9] E. Amaldi, A. Capone, M. Cesana, I. Filippini, F. Malucelli. Optimization models and methods for planning wireless mesh networks. Computer Networks 52 (2008) 2159-2171.

[10] Xhafa, F., A. Barolli, C. Sánchez, L. Barolli. A simulated annealing algorithm for router nodes placement problem in Wireless Mesh Networks.Simulation Modelling Practice and Theory, In Press, 2010.

[11] J. Wang, B. Xie, K. Cai and D.P. Agrawal. Efficient Mesh Router Placement in Wireless Mesh Networks, MASS 2007, Pisa, Italy (2007).

[12] Xhafa, F., C. Sanchez, and . L. Barolli, Genetic Algorithms for Efficient Placement of Router Nodes in Wireless Mesh Networks in Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on p. 465-472.

[13] Fendji, J.L. E. K.,Thron,C., &Nlong,J.M.: Mesh router nodes placement in rural wireless mesh networks. In M.Sellami, E.Badouel, &M.Lo (Eds.), Actes du CARI 2014 (ColloqueAfricain Sur LA Recherche en InformatiqueetMathématiquesAppliquées). Inria:Colloques CARI, pp. 265-272.

Appendix 1

Basic algorithm of Simulated Annealing

Algorithm 2: Simulated annealing

Input: f : the objective function to be minimised
Output: s: the best solution found

Begin
  $T := T_{initial}$; s:= InitialSolution(); v:= $f(s)$
  while (stopping condition not met) do
    while (equilibrium condition not met) do
      s' := GenerateSolution()
      v' := $f(s')$
      $\Delta E := v' - v$
      if $\Delta E \leq 0$ then $s := s'$
      else accept s' with probability $e^{-\frac{\Delta E}{T}}$
    Update(T)
  return s
End

Particularisation of the SA algorithm

Initialization

Routers are placed randomly in areas of interest in the region during the initialization phase.

Cooling schedule

The initial temperature T=10. A geometric update scheme with α=0.5 has been selected. When the temperature is less than T_{min}=0.01, the cooling process stops.

Move

Only one router is moved at the same time, in a randomly selected direction and distance. The movement from the current EA_a to the new EA_b is simulated if and only if *Cover*(EA_b)=1 and *Place*(EA_b)=1. Initially great moves are selected to allow a rapid convergence. The size of moves decreases with the temperature; when the temperature is close toT_{min}, the size of moves is one EA.

Fitness function

We also count the number of EAIs that are covered to evaluate the fitness function. This is done by (7) after the initialisation.

Acceptance criterion

When $C_b \geq C_a$, the coverage change is directly accepted. But when the coverage change is negative, the change is accepted with a certain probability following the Boltzmann distribution and influenced by the temperature $T$to avoid local optimum.

Equilibrium state and stopping condition

The equilibrium state is supposed to be reached if after a number (stop) of moves no solution has been accepted. The stopping condition depends on Imp and on T_{min}. At each temperatureT_i, Imp indicates whether the solution has improved. When the equilibrium state at a temperature $T_i$is reached, before decreasing the temperature we check whether the solution has improved. In case of an improvement, we decrease the temperature and move to the next iteration. But if there is no improvement or the temperature is less than T_{min}, we stop the search process and suppose having reached an optimum.

At the beginning nr_{min} routers are used. The SA algorithm is running *nRun* times at each stage. If the required coverage is satisfied, we remove one router and restart until the coverage can no longer be satisfied.

Appendix 2

Data from simulation

Routers		Run1	Run2	Run3	Run4	Run5	Run6	Run7	Run8	Run9	Run10
$1.5nr_{min}$	75	9,17	24,80	11,99	9,67	18,92	7,06	11,74	14,03	15,51	11,92
$1.4nr_{min}$	70	11,69	11,94	6,90	7,37	9,88	18,42	10,12	14,59	13,85	12,40
$1.3nr_{min}$	65	6,91	12,88	8,18	17,07	8,33	14,47	9,47	17,07	9,14	9,19
$1.2nr_{min}$	60	15,13	14,82	8,65	17,47	12,06	10,71	16,88	12,08	9,50	6,19
$1.1nr_{min}$	55	12,19	11,72	13,79	10,72	7,70	16,30	11,03	7,95	18,71	7,51
nr_{min}	50	13,84	8,36	13,36	15,29	12,48	17,44	9,70	7,50	20,56	14,03

Table 1: CM Approach CPU Time

Routers		Run1	Run2	Run3	Run4	Run5	Run6	Run7	Run8	Run9	Run10
$1.5nr_{min}$	75	196,27	223,42	222,76	224,51	174,66	186,27	151,42	195,63	165,97	165,84
$1.4nr_{min}$	70	203,54	243,02	187,52	173,54	166,77	167,72	201,57	202,68	271,51	237,25
$1.3nr_{min}$	65	192,30	216,68	192,75	168,92	201,00	138,45	281,63	198,06	229,91	333,49
$1.2nr_{min}$	60	322,46	266,62	217,05	154,74	195,34	244,87	186,72	257,85	189,71	224,98
$1.1nr_{min}$	55	204,20	167,72	212,28	119,11	165,88	206,61	197,04	174,12	179,18	154,24
nr_{min}	50	147,71	130,44	182,64	159,86	147,97	277,46	172,26	169,05	172,86	154,45

Table 2: SA Approach CPU Time

Routers		Run1	Run2	Run3	Run4	Run5	Run6	Run7	Run8	Run9	Run10
$1.5nr_{min}$	75	95,17	95,77	95,04	95,54	95,20	95,99	96,40	95,47	95,53	95,44
$1.4nr_{min}$	70	93,96	94,35	93,69	92,39	94,97	94,31	93,49	94,19	94,08	94,21
$1.3nr_{min}$	65	91,41	91,66	90,80	92,41	90,68	92,32	91,41	91,28	91,48	92,01
$1.2nr_{min}$	60	89,21	89,00	87,93	88,82	89,41	88,91	88,71	87,50	88,00	87,27

$1.1nr_{min}$	55	84,92	84,49	85,40	85,13	85,04	85,52	84,99	84,92	85,95	85,86
nr_{min}	50	80,37	79,66	80,19	81,14	82,17	80,73	78,77	78,84	81,26	79,96

Table 3: CM Approach Coverage percentage of area of interest

Routers		Run1	Run2	Run3	Run4	Run5	Run6	Run7	Run8	Run9	Run10
$1.5nr_{min}$	75	99,36	99,88	99,57	98,82	99,06	99,70	99,98	98,89	99,82	99,80
$1.4nr_{min}$	70	99,25	99,57	94,99	99,59	99,75	98,70	99,88	99,75	98,38	99,30
$1.3nr_{min}$	65	98,09	96,86	99,47	96,70	98,77	96,61	97,29	97,52	97,04	98,00
$1.2nr_{min}$	60	95,44	95,29	96,68	97,41	97,41	96,84	97,82	97,06	96,18	89,94
$1.1nr_{min}$	55	92,03	95,19	94,54	94,06	94,29	94,72	91,05	93,64	94,85	94,56
nr_{min}	50	91,28	90,62	90,19	89,89	90,32	88,27	89,41	90,21	90,18	90,00

Table 4: SA Approach Coverage percentage of area of interest

LTB-MAC

Linear Token-Based MAC protocol for Linear Sensor Network

El Hadji Malick Ndoye[1,2], Ibrahima Niang[1] Frédérique Jacquet[2] and Michel Misson[2]
[1]Laboratoire d'Informatique, Université Cheikh AntaDiop de Dakar (UCAD),B. P. 5005 Dakar, Sénégal
[2]Clermont Université / LIMOS CNRS - Complexe scientifique des Cézeaux, 63172 Aubière cedex, France
E-mail: {ndoye, jacquet, misson}@sancy.univ-bpclermont.fr,
ibrahima1.niang@ucad.edu.sn

ABSTRACT. A wireless sensor network is a large number of sensor nodes deployed in a fixed or random manner over a wide area for environmental monitoring applications. Wireless sensors communicate via wireless links and are powered by batteries. They collect and provide information to the base station usually called sink. The information collected is generally physical, chemical or biological nature. For some of these applications, as pipeline or road monitoring, wireless sensor nodes have to be deployed in a linear manner. We refer to these WSNs as Linear Sensor Networks (LSNs). Suitable MAC protocols for LSN must take account the linearity in order to ensure reliability and optimize parameters such as the end-to-end delay, the delivery ratio, the throughput, etc. In this paper, we present LTB-MAC a Linear Token Based Mac Protocol designed for linear sensor networks. We compare the protocol with CSMA/CA in terms of delay, delivery and throughput.

KEYWORDS: Wireless sensor network, linear topology, throughput, MAC protocol, CSMA/CA, RTS/CTS, token passing, end-to-end delay, delivery ratio.

1. Introduction

In LSNs, MAC protocols must effectively ensure the end-to-end delay, throughput and delivery ratio through a protected and effective access to the channel. This paper focuses on a Linear Token Based MAC protocol for linear sensor network (LTB-MAC). LTB-MAC is based on a synchronization using token generation for the access to the transmission channel. The token contains temporal informations on the periods of activity and inactivity of nodes. So, it gives to a node the right to access to the channel during an amount of time. We evaluate LTB-MAC in terms of throughput, end-to-end delay in comparison to CSMA/CA in order to show its impact on the behavior of LSN.
The rest of the paper is outlined as follows: in section 2 we present a state of the art on the MAC protocols used in linear networks. Section 3 gives the hypothesis and the network topology; section 4 presents LTB-MAC. In this section we show the principles of the token by explaining the role of temporal information related to the token We present our simulation results in section 5. Finally, we end this paper with a conclusion and perspectives in section 6.

2. State of art

The linear sensor networks are present in many monitoring applications. They are found in the surveillance of pipelines [1][2][3], mine [4][5][6], volcanoes, bridges or roads [7] [8], etc. They are characterized by a limited neighborhood and extend over long distances [9].

The major challenges of MAC protocols in LSN are therefore equitable load distribution on the nodes in the linear topology, optimizing the rate of delivery packets and the end-to-end delay, fault tolerance, energy saving, etc. Linear sensor network MAC protocols are mainly based on the contention and synchronization. DiS-MAC [10] is a MAC protocol based on time synchronization of sensor nodes linearly deployed for highway surveillance. In DiS-MAC, each node uses directional oriented antennas to reduce its transmission range to a direct neighbor in the line. This minimizes not only the interference between nodes on the same line but also between the nodes of the line on the other side of the highway. LC-MAC [11] is another MAC protocol based on time synchronization. This protocol is designed for linear networks extending over long distances to reduce the end-to-end delay while saving the energy of sensors. CMAC-T [12] is a MAC protocol for linear network designed for forest environment monitoring. It uses token propagation for nodes access to the transmission channel. WiWi [13] is a MAC protocol which synchronizes nodes by using time slots. The 802.15.4 MAC protocol [14] is the main contention MAC protocol used in linear sensor network. The L-CSMA protocol [15] is a MAC protocol based on 802.15.4 CSMA/CA and is designed for linear sensor network. It is assumed that, with CSMA/CA, the probability that a packet collides during its transport is quite high in the case of a linear topology

because of the contention and the problem of hidden terminal. The protocol presented in [16] makes a comparison of CSMA/CA protocol with RTS/CTS and without RTS/CTS in a linear sensor network.

In previous works [17][18], we introduced LTB-MAC as a Token-Based Mac Protrocol for Linear Sensor Networks in conditions where radio links are supposedly stable and identical between nodes uniformly distributed . We considered the Tworayground [19] propagation model in which the reception power for a given link remains constant for a given transmission power. We define a R-redundant LSN according to the neighborhood of the nodes. We compare different LSN using LTB-MAC in terms of throughput, delivery ratio and end-to-end delay. In this paper, we compare LTB-MAC with 802.15.4 with RTS/CTS in term of throughput and delivery ratio in a Linear Sensor Network.

3. Hypothesis and network topology

We focus on a linear sensor network where the access to the transmission channel is managed by a token generation. Three types of sensors can be defined according to the role of the sensor nodes. The basic node that is a simple node with the relay functions of aggregated data. The Token Allocator that creates the token periodically is usually located at the opposite end of the sink. In Fig. 1, it is located at the extreme left of the network. The Token Allocator is also a basic node with the particularity of having no left neighbors. The sink is the base station which aggregates and analyzes data.

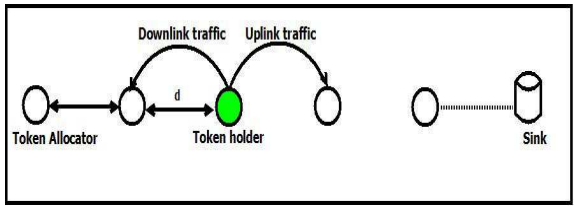


Fig. 1. Linear sensor network

In this study, we consider a LSN where the Token Allocator is located at the extreme left of the network and the sink at the right end. In this case, for a given node, we define two types of neighbors. (i) The left neighbors which are nearest to the Token Allocator. (ii) The right neighbors which are nearest of the sink. In the LSN data can transit from Token Allocator to the sink node. We refer this traffic as uplink traffic. This traffic consists of information collected by the monitoring application (physical, chemical,

environmental variables, etc.). They can also transit from the sink to the Token Allocator. This is called downlink traffic. This traffic consists of control data of the network or the application. We can also include synchronization or alert settings.

4. LTB-MAC description

In LTB-MAC, the token gives to a given node the access to the transmission channel. This is a data frame containing temporal informations on the synchronization of the nodes. So, a node is either token holder or is waiting for it. When it is the possessor of the token it accesses to the transmission channel during a defined time interval. This time interval is divided as follows as shown in [17].

Fig. 2. shows the sequence of activity time and inactivity for a defined node. Upon awakening, the node goes into reception mode of the uplink traffic of its left neighbor during T_0 and the token during T'_0. After the T'_0 period it receives the token and begins its transmission period ($T_1 + T_2 + T'_2$). After the transmission of the token, the node goes into reception mode of downlink traffic during T_3. At the end of the reception, it then goes into sleep mode to save power during T_4.

The way of propagating the token from node to node towards the sink can be seen as the passage of a shuttle [17][18] in which the nodes deposit their traffic towards sink. We define the shuttle duration (Fig. 2) as the amount of time duration of the shuttle duration (SDur) and the amount of information exchanged by a node during the shuttle passage as the shuttle payload (Splo).

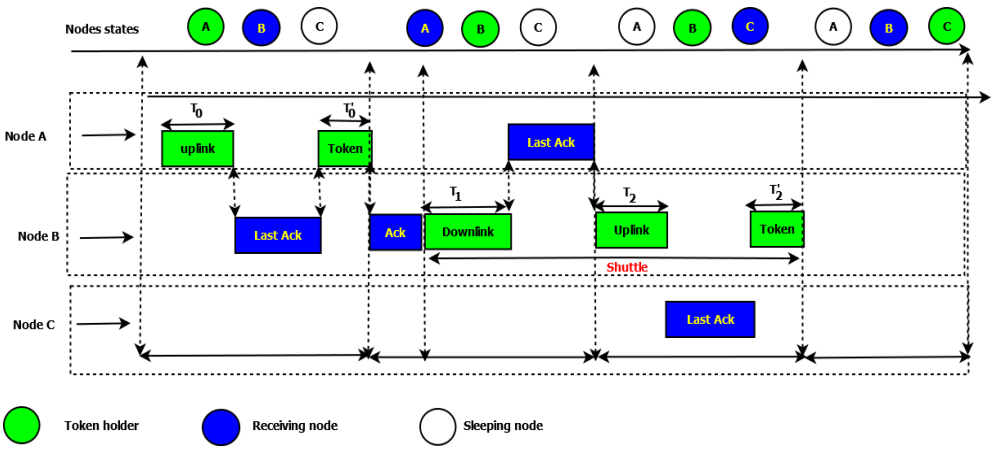


Fig. 2. Token process and shuttle propagation

5. Simulations and results

5. 1. Simulations parameters

We simulate our analysis on NS2 with version 2.32. We consider a linear sensor network of 10 nodes and a sink. Local traffic is generated pseudo-randomly per time interval and begins independently between 0 and 1 s for each node. The conditions of propagation are made so that each node has exactly two neighbors: one on the right and one on the left. To do this, let's consider a transmission power of -5 dBm and a distance equal to 90 m. The size of FIFOs is considered fixed and equal to 50 packets.

The possibility of downlink traffic is neglected assuming that the physical characteristics of the sink allow it to receive correctly traffic. In this case, the need to send resynchronization messages or alert is negligible. We focus on three performance parameters: the throughput at the sink, the end-to-end delay and the delivery rate for a given node. The throughput is the average rate of traffic received by the sink per time unit while the delivery ratio is the rate of packets delivered to the sink for a given node. In fact, it represents the ratio between the number of received packets and the number of sent packets by the node. The delivery rate depends on the overall load of the network at a time and the number of hops performed by the packets before reaching the sink. We are mainly interested in two nodes to study the delivery rate: node 1 and node 5. We compare LTB-MAC with the CSMA/CA protocol with and without RTS/CTS. Fig.3 shows the conditions of collision-free transmissions for LTB-MAC protocol and CSMA/CA with RTS/CTS.

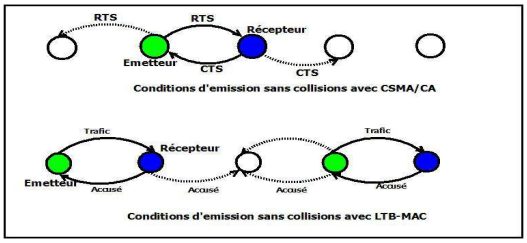


Fig. 3. Conditions of collision-free transmissions

The simulation parameters are summarized in the following table.

Table 1 Simulations parameters

Parameters	Values
Propagation Model	Tworayground
Token size	11 Bytes
Frame size	100 Bytes
Number of repetitions	100
Physical Layer	802.15.4
FIFO size	50-60 packets
Transmission Power	-5 dBm
LSN offered load	[10-100] Kbps
Simulation start time	[0-1] s
Simulation end time	[199-200] s
Shuttle duration	10, 50, 250 ms

5.2. Results

Fig. 4 shows the throughput at the sink according to the global load offered in the network for small shuttles 10, 50 and 250 ms. It shows that the LTB-MAC protocol offers better performance than the CSMA/CA with or without RTS/CTS in terms of throughput beyond a 10 ms shuttle. Indeed, for LTB-MAC the maximum received flow rate is about 40 Kbps, while it is 25 Kbps for CSMA/CA with RTS/CTS and 15 Kbps without RTS /CTS. The evolution of the throughput for LTB-MAC protocol can be divided into two phases.

- Between 8 and 40 Kbps of offered load (depending on the shuttle): during this phase the it increases as a function of the overall network load. This is explained by the fact that the network is not overloaded and therefore the FIFOs do not overflow. Thus, during the passage of the shuttle, the nodes are able to send as much data as possible.
- Between 40 and 80 Kbps the throughput is stationary. This phase corresponds to the saturation which is the consequence of the high network load. Therefore, the aggregated data during passage of the shuttle remains constant which explains that the throughput does not progress.

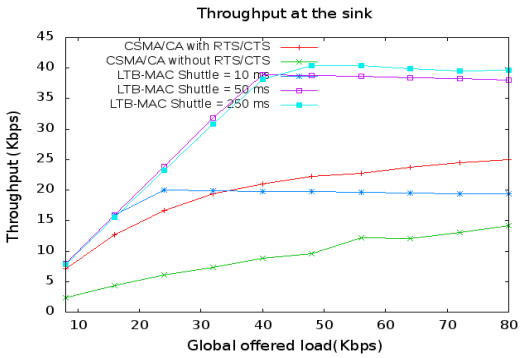


Fig. 4. Throughput at the sink

Fig. 5 and Fig. 6 show the delivery ratio as a function of the network load expressed in number of packets per second for three shuttles. For nodes 1 and 5 we can see that the packet delivery ratio at the sink is more important in the case of LTB-MAC protocol than for CSMA/CA protocols with or without RTS/CTS even in a very small shuttle of 10 ms. Indeed, for node 5, the minimum delivery ratio (maximum resp.) is 0.6 (resp. 1), whereas it is 0.11 (resp. 0.85) and 0.28 (resp. 0.25) respectively for the CSMA/CA protocols with RTS/CTS and without RTS/CTS. For node 1, we find a minimal ratio of 0.4 for LTB-MAC whereas it is 0.1 for the CSMA/CA protocols with or without RTS / CTS.

For LTB-MAC protocol we see that the evolution of the curve is divided into two parts.

- Between 10 and 60 packets per second (depending on the shuttle). In this case the delivery rate is equal to 1 because the network is not loaded. So the packets are not victims of overload of the FIFOs.
- Between 60 and 100 packets per second. In this part, the delivery rate decreases gradually as the network is loaded. This is explained by the fact that the FIFOs are overloaded causing packet drops.

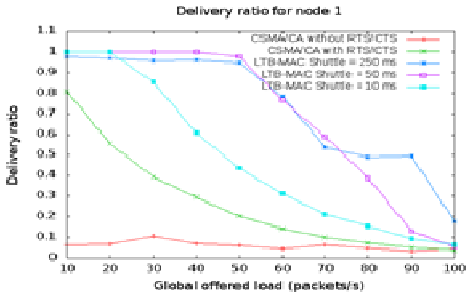


Fig. 5. Delivery ratio for node 1

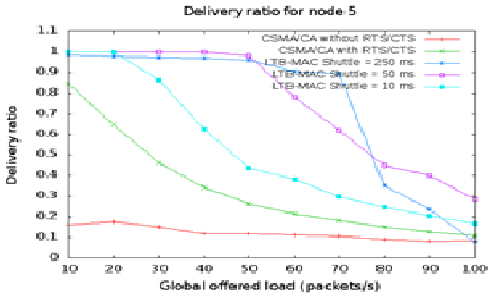


Fig. 6. Delivery ratio for node 5

6. Conclusion

In this paper present LTB-MAC based on the generation of a token that gives a node the right to access the transmission channel. It is created by the node that is at the opposite extremity of the sink known as token generator. It contains information on the activity periods of the nodes. The propagation of the token is similar to a shuttle that passes and in which the nodes deposit information to the sink. The shuttle determines the amount of information that a node can send when it is token holder. We compare LTB-MAC protocol to CSMA/CA in terms throughput, delivery ratio and end-to-end delay. We have shown, thanks to simulations, that the LTB-MAC protocol offers better performance than the CSMA/CA in wireless networks of linear sensors.

In our future work, we plan study the redundancy in sensors using LTB-MAC protocol to better optimize the performance parameters.

References

[1] S. Yoon, W. Ye, J. Heidemann, B. Littlefield, and C. Shahabi, “SWATS: Wireless sensor networks for steamflood and waterflood pipeline monitoring,” Netw. IEEE, vol. 25, no. 1, pp. 50–56, 2011.
[2] I. Jawhar, N. Mohamed, and K. Shuaib, “A framework for pipeline infrastructure monitoring using wireless sensor networks,” in Wireless Telecommunications Symposium, 2007. WTS 2007, 2007, pp. 1–7.
[3] T. T.-T. Lai, W.-J. Chen, K.-H. Li, P. Huang, and H.-H. Chu, “Triopusnet: Automating wireless sensor network deployment and replacement in pipeline monitoring,” in Information Processing in Sensor Networks (IPSN), 2012 ACM/IEEE 11th International Conference on, 2012, pp. 61–71.
[4] M. Li and Y. Liu, “Underground coal mine monitoring with wireless sensor networks,” ACM Trans. Sens. Netw., vol. 5, no. 2, pp. 1–29, Mar. 2009.
[5] M. Li and Y. Liu, “Underground Structure Monitoring with Wireless Sensor Networks,” in Proceedings of the 6th International Conference on Information Processing in Sensor Networks, New York, NY, USA, 2007, pp. 69–78.
[6] G. Werner-Allen, J. Johnson, M. Ruiz, J. Lees, and M. Welsh, “Monitoring volcanic eruptions with a wireless sensor network,” in Wireless Sensor Networks, 2005. Proceeedings of the Second European Workshop on, 2005, pp. 108–120.
[7] W. Nan, M. Qingfeng, Z. Bin, L. Tong, and M. Qinghai, “Research on Linear Wireless Sensor Networks Used for Online Monitoring of Rolling Bearing in Freight Train,” J. Phys. Conf. Ser., vol. 305, no. 1, p. 012024, Jul. 2011.
[8] M. Zimmerling, W. Dargie, and J. M. Reason, “Localized power-aware routing in linear wireless sensor networks,” in Proceedings of the 2nd ACM international conference on Context-awareness for self-managing systems, 2008, pp. 24–33.
[9] M. FITZSIMONS, “The langeled approach,” World Pipelines, vol. 5, no. 6, pp. 24–26, 2005.
[10] T. Karveli, K. Voulgaris, M. Ghavami, and A. H. Aghvami, “DiS-MAC: A MAC protocol for sensor networks used for roadside and highway monitoring,” in International Conference on Ultra Modern Telecommunications Workshops, 2009. ICUMT ’09, 2009, pp. 1–6.
[11] C. Fang, H. Liu, and L. Qian, “Lc-mac: An efficient mac protocol for the long-chain wireless sensor networks,” in Communications and Mobile Computing (CMC), 2011 Third International Conference on, 2011, pp. 495–500.
[12] T. Sun, X. J. Yan, and Y. Yan, “A Chain-type Wireless Sensor Network in Greenhouse Agriculture,” J. Comput., vol. 8, no. 9, Sep. 2013.
[13] D. De Caneva, P. L. Montessoro, and others, “A synchronous and deterministic MAC protocol for wireless communications on linear topologies,” Intl J Commun. Netw. Syst. Sci., vol. 3, no. 12, p. 925, 2010.
[14] L. S. Committee and others, “Part 15.4: wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (LR-WPANs),” IEEE Comput. Soc., 2003.
[15] C. Buratti and R. Verdone, “L-CSMA: A MAC Protocol for Multi-Hop Linear Wireless (Sensor) Networks,” 2015.
[16] E. H. M. Ndoye, F. Jacquet, M. Misson, and I. Niang, “Evaluation of RTS/CTS with unslotted CSMA/CA algorithm in linear sensor networks,” NICST, France, 2013.
[17] E. H. M. NDOYE, F. JACQUET, M. MISSON, and I. NIANG, “A Token-based MAC Protocol for Linear Sensor Networks.,” Sens. Transducers 1726-5479, vol. 189, no. 6, 2015.
[18] E. H. M. Ndoye, F. Jacquet, M. Misson, and I. Niang, “Using a token approach for the MAC Layer of Linear Sensor Neworks : Impact of the redundancy on the Throughput,” Sensornet, France, 2015.

Méthode Tabou d'allocation des slots de fréquence requis sur chaque lien d'un réseau optique flexible

Kamagaté Beman Hamidja [*] ─ Babri Michel** ─ Gooré Bi Tra[**] ─ Brou Konan Marcelin[**]

*.Ecole Doctorale Polytechnique de l'Institut National Polytechnique Houphouët Boigny de Yamoussoukro (EDP/INP-HB).Laboratoire de Recherche en Informatique et Télécommunication (LaRIT)/ E-mail : beman2017@gmail.com

**.l'Institut National Polytechnique Houphouët Boigny de Yamoussoukro (INP-HB)

RÉSUMÉ. Les réseaux optiques élastiques constituent à n'en point douter une solution prometteuse face à la croissance exponentielle du trafic généré par les réseaux de télécommunication. Ils allient la flexibilité à la granularité plus fine des ressources optiques pour se positionner comme une meilleure solution que le réseau conventionnel WDM. Cependant la multiplicité des ressources et la possibilité d'avoir plusieurs niveaux de modulation avec l'utilisation de l'OFDM rend plus difficile l'allocation des ressources aux requêtes des clients. Ce présent travail tend par l'utilisation judicieuse d'une méta-heuristique Tabou d'apporter une contribution à l'utilisation optimale des ressources optiques par la minimisation du nombre de slots de fréquence nécessaires sur chaque lien optique.

MOTS-CLÉS : réseau optique élastique/flexible, OFDM, RMSA, méthode Tabou, Slot de fréquence

ABSTRACT. The elastic optical networks represent a promising solution to the exponential growth in traffic generated by the telecommunication network. It combines flexibility with the finest granularity of optical resources to position itself as a better solution than conventional WDM network. However the multiplicity of resources and the ability to have multiple modulation level with the use of OFDM make harder allocating resources to client requests. This present work tends to use a Tabu meta-heuristic to contribute to the optimal utilization of optical resources by minimizing the number of frequency slots required on each optical link.

KEYWORDS: Elastic/Flexible optical network, OFDM, RMSA, Tabu search, Frequency slot.

1. Introduction

Le réseau de communication optique conventionnel WDM (Wavelength Division Multiplexing) basé sur la grille fixe de 50 GHz de l'Union Internationale des Télécommunications(UIT) est inadapté aux nouvelles exigences des télécommunications. Son spectre de fréquence manque de flexibilité et le signal optique subit des détériorations lorsque la capacité de celui-ci devient très grande comme l'exige les besoins de transmissions actuelles. C'est pourquoi le nouveau paradigme de réseau de communication optique basé sur la technologie de multiplexage de division orthogonale des fréquences(OFDM) connu sous le vocable de réseau optique élastique ou réseau optique flexible s'avère être un candidat prometteur pour les futures générations de réseaux de communication optiques à haute capacité. Le réseau optique flexible fournit des bandes passantes hétérogènes en fonction des débits requis par les requêtes et apporte une plus grande flexibilité dans l'utilisation du spectre de fréquence grâce à l'architecture SLICE [1]. Cette architecture est constituée de transpondeurs et de commutateurs à débits variables, reliés par des liens en fibre optique.

Cependant l'un des principaux défis à relever pour ce réseau est de trouver pour une requête donnée, la route, le format de modulation et des slots de fréquence appropriés. Ceci doit se faire dans le respect des contraintes de contiguïté et de continuité auxquelles sont soumis les slots de fréquence alloués à la requête. Ce problème est connu sous le nom de Routage, Modulation et Allocation de Spectre de Fréquence (RMSA)[2],[3]. Lorsque la modulation est fixe, il se réduit au Routage et Allocation du Spectre de Fréquence (RSA). Le RSA est l'équivalent du problème RWA (Routing and Wavelength Assignment) dans les réseaux WDM. Ce problème comme plusieurs études [3, 4] l'attestent est réputé NP-difficile. Pour un réseau de grande taille, il est difficile de trouver une solution exacte en un temps raisonnable. L'un des champs d'étude du RMSA est l'Allocation Minimum de Slots de Fréquence(AMSF) dans la phase de planification du réseau qui consiste à minimiser le nombre de slots de fréquence nécessaires sur chaque lien. Cela participe à l'optimisation de l'utilisation des slots de fréquence. Notre travail va donc consister à apporter une contribution à la réduction du nombre de slots de fréquence requis sur chaque lien pour le traitement d'un ensemble de requête connu d'avance. Pour cela, la suite de ce travail se structure de la façon suivante. Dans la section 2, nous présentons le contexte technologique ainsi que des approches de résolution de l'AMSF existantes. La section 3 est consacrée à notre contribution qui est une approche basée sur la méta-heuristique Tabou. Puis nous terminons dans la section 4 par une simulation et l'analyse des résultats obtenus.

2. Travaux liés

2.1. Contexte technologique

Dans les réseaux optiques élastiques, le spectre optique est subdivisé en plusieurs slots de fréquence. Chaque slot de fréquence a une bande passante fixée généralement à 12,5 GHz. Un ou plusieurs slots de fréquence adjacents peuvent être alloués à une requête en fonction de sa capacité et du niveau de modulation adapté à la portée de son signal optique. La formule (1) indique le nombre de slots de fréquence requis par une requête de capacité C_i en Gb/s

$$N_i = \left\lceil \frac{C_i}{M_i * F_{slot}} \right\rceil \quad (1)$$

$M * C_{slot}$ représente la capacité d'un slot de fréquence. M (b/s/Hz) est le niveau de modulation en nombre de bits par symbole et représente l'efficacité du format de modulation choisi. F_{slot} est la bande passante d'un slot de fréquence en GHz. M peut prendre les valeurs 1,2,3 ou 4 selon que le format de modulation est BPSK, QPSK, 8-QAM ou 16-QAM. Le nombre de slots de fréquence requis par une requête est donc fonction du format de modulation. De plus un lien optique est constitué d'un nombre connu de slots de fréquence qui sont chacun identifiés par un numéro d'indice qui est nombre entier positif. Par exemple un lien optique de spectre 5 THz contient 400 slots de fréquence, les slots de fréquence peuvent être numérotés de 1 à 400. L'utilisation de la technologie de multiplexage OFDM dans les réseaux optiques élastiques autorise le chevauchement du spectre des slots de fréquence alloués à un même signal optique. L'ensemble de ces slots de fréquence constitue un canal unique sans bande de garde entre les slots de fréquence, ce qui n'est pas le cas du réseau conventionnel WDM où chaque longueur d'onde est séparée d'une autre par un intervalle rigide de 50 Ghz.

2.2. Approches de résolution existantes

Conceptuellement, le réseau est représenté par un graphe G(V,E) orienté. V est un ensemble de nœuds qui sont des transpondeurs ou des commutateurs de bandes passantes variables. E est un ensemble de liens en fibre optique. Chaque requête R_i doit avoir un chemin physique P_i constitué de liens en fibre optique interconnectés par des nœuds. A chaque requête R_i est alloué un ensemble de slots de fréquence contigus identifié par l'indice du premier slot de fréquence noté $\varphi(R_i)$. Le format de modulation qui détermine le nombre de slots de fréquence N_i de la requête R_i dépend de la longueur du chemin physique P_i[2].

L'Allocation Minimum de Slots de Fréquence(AMSF) consiste à déterminer le nombre minimal de slots de fréquence pour satisfaire une instance de requêtes connue d'avance. Elle se fait en respectant les contraintes de continuité, de contiguïté et non chevauchement des slots de fréquence. C'est un problème d'optimisation linéaire en

nombre entier dont l'objectif est de minimiser l'indice maximal de slots de fréquence alloués sur chaque lien optique. Plusieurs travaux [5,6,7] proposent des modèles de programmation linéaire en nombre entiers et utilisent des méthodes exactes de résolution tels que *Branch and Bound(BB), Branch and Price(BP), Column Generation (CG*). Bien que ces méthodes donnent des solutions optimales, elles ne sont pas utilisables pour les réseaux de grandes tailles, puisqu'il s'agit d'un problème NP-difficile. C'est pourquoi d'autres auteurs utilisent des heuristiques et des méta-heuristiques. Ces dernières permettent d'avoir des solutions approchées, en un temps raisonnable. Parmi ces heuristiques, on peut citer le *Most Subcarriers First* (MSF) et le *Longest Path First(LPF)* proposés dans [8] qui sont des heuristiques qui résolvent séparément le routage et l'allocation des slots de fréquence. Dans la même approche de résolution séparée ou séquentielle du routage et de l'allocation, Wang et *al* [5] proposent le *Shortest Path with Spectrum Reuse*(SPSR) qui utilise les plus courts chemins disjoints afin de favoriser la réutilisation des slots de fréquence et le *Balanced Load Spectrum Allocation (BLSA)* qui choisit le chemin le moins congestionné. En ce qui concerne les méta-heuristiques, une méthode basée sur les algorithmes génétiques a été proposée dans[9] et plus récemment une approche basée sur la recherche Tabou plus spécifiquement pour les réseaux optiques multicast a été proposé dans [10]. Toutes ces approches sur-citées procèdent par une séparation du problème en deux sous-problèmes, celui du routage, souvent de la modulation et celui de l'allocation. Si cette séparation simplifie la complexité du problème, elle n'est pas de nature à préserver l'optimalité des solutions.
C'est pourquoi nous proposons une méta-heuristique basée sur la recherche Tabou qui prend en compte la résolution simultanée du routage, de la modulation et de l'allocation.

3. Méthode tabou de résolution de l'AMSF

La méthode Tabou [11] proposée par F.Glover en 1986, est une méta-heuristique itérative qui, à partir d'une solution initiale, construit de nouvelles solutions. Chaque solution notée Γ conduit à un ensemble de solutions voisines notées $N(\Gamma)$. Pour traiter l'AMSF par la méthode Tabou, nous procédons comme suit. Etant donnée une instance de requêtes $\Delta = \{R_1, R_2, R_3, \dots, R_m\}$, on associe à chaque requête $R_i \in \Delta$ un triplet$(P_i, \varphi(R_i), N_i)$. Une solution est un ensemble de triplets noté :
Γ=$\{(P_1, \varphi(R_1), N_1), (P_2, \varphi(R_2), N_2) \dots (P_m, \varphi(R_m), N_m)\}$ et a un coût $C(\Gamma)$ définie par la somme des degrés de conflits des requêtes de l'instance Δ. La formule (2) donne ce coût :

$$C(\Gamma) = \sum_i^{|\Delta|} C_i \qquad (2)$$

Le degré de conflits d'une requête R_i noté C_i est le nombre de requêtes dont les chemins partagent au moins un lien avec celui de R_i et auxquelles sont allouées au moins un slot de fréquence identique à ceux alloués à la requête R_i. C'est-à-dire chacune de ces requêtes a une plage de slots de fréquence allouée qui se chevauche avec celle de R_i sur au moins un lien. Prenons comme exemple illustratif une solution définie par l'ensemble

suivant : $\Gamma = \{(P_1, \varphi(R_1), N_1), (P_2, \varphi(R_2), N_2), (P_3, \varphi(R_3), N_3)\}$. Dans cette solution, si R_1 est en conflit avec R_2 et R_3 alors $C_1 = 2$. Par contre si R_1 n'est en conflit avec aucune des requêtes appartenant à la solution Γ alors $C_1 = 0$ et par la suite si au aucune requête n'est en conflit avec les autres alors $C(\Gamma) = 0$. Dans ce cas la solution Γ est faisable.

La valeur initiale d'indice de slot de fréquence maximale MS doit être suffisamment grande pour que toute requête R_i de Δ puisse être traitée en lui associant son triplet. Intuitivement MS est l'indice du dernier slot de fréquence occupé sur un lien optique en supposant que toutes les requêtes traversent ce lien. MS initiale se calcule avec la formule (3)

$$MS = \sum_{i=1}^{|\Delta|} N_i + (|\Delta| - 1) * BG \qquad (3)$$

Dans cette formule N_i se calcule avec la formule (1) en fixant le niveau de modulation M à 1 peu importe la longueur du chemin de la requête R_i .BG représente la bande de garde entre deux connexions adjacentes et permet d'éviter les interférences. BG est constitué d'un nombre fixe de slots de fréquence.

La recherche avec Tabou proposée va donc consister à améliorer la valeur initiale MS en cherchant des solutions d'allocations faisables, c'est-à-dire de coût nul. On rentre dans la procédure Tabou avec cette valeur initiale, lors de cette procédure les slots de fréquence alloués aux requêtes auront des indices compris dans l'intervalle$[1,\ Ms]$. Lorsqu'une solution faisable est trouvée, on décrémente MS de 1 et on cherche une nouvelle solution faisable avec la nouvelle valeur de MS. L'algorithme s'arrête lorsqu'un nombre maximal d'itération (NbItMax) prédéfini est atteint ou lorsqu'une certaine valeur de MS ne permet pas d'avoir une solution faisable. Le nombre d'indice maximal de slots de fréquence sur chaque lien devient finalement la dernière valeur de MS incrémenté de 1. Pour chaque requête, on calcule k-plus courts chemins avec l'algorithme de Yen[12]. Puis on construit une solution initiale Γ_{ini} en attribuant de façon aléatoire à chaque R_i un de ses k-plus courts chemins. Le nombre de slots de fréquence requis par cette requête est calculé en utilisant la formule (1). Ce qui permet de choisir le niveau de modulation adapté à la longueur du chemin. Le choix du premier slot de fréquence $\varphi(R_i)$se fait aussi de façon aléatoire dans l'intervalle$[1,\ Ms]$. On construit l'espace des solutions voisines $N(\Gamma)$ à la solution Γ en générant une liste de conflit dérivée de la solution initiale. Cette liste de conflit (LC) contient les requêtes R_i en conflits avec d'autres requêtes de la solution Γ ($C_i > 0$). On désigne une tête de la liste LC notée R_t. A partir de cette tête de liste, on génère une solution voisine Γ' de la solution Γ en allouant une nouvelle valeur à la tête de liste R_t tout en laissant les valeurs des autres requêtes de Γ inchangées. La Liste de Conflit (LC) doit aussi être mise à jour. Pour cela, on crée une liste Θ_f constituée des requêtes appartenant à Γ et qui sont en conflit avec la tête de liste R_t. On calcule le degré de conflit de chaque $R_x \in \Theta_f$ par rapport à Γ'. Si le degré de conflit de R_x vaut zéro, cela signifie que la requête R_x n'est en conflit avec aucune des requêtes de Γ' et donc, on peut retirer R_x de la liste de conflit LC. Si le degré de conflits R_t vaut zéro, R_t est retirée de la liste LC et une autre requête de LC devient la tête de liste, dans le cas contraire cela

signifie que R_t est en conflit avec une ou plusieurs requêtes de la solution Γ'. Représentons par Θ'_f cet ensemble de requêtes. Si $R_y \in \Theta'_f$ est déjà dans LC, R_y passe en tête de la liste LC. Les autres requêtes de Θ'_f sont insérées une à une dans la liste LC, avec cette méthode la tête de liste R_t est renouvelée pour chaque itération et la procédure de recherche guide vers un nouveau voisin en se basant sur la nouvelle tête de la liste LC. Cette procédure s'accompagne de deux règles : la Règle Tabou (RT) et la Règle d'Aspiration (RA). Ces deux règles ainsi les mécanismes de création de la solution initiale et de mise à jour de la liste de conflits sont explicités dans l'annexe 1.

Algorithme : Pseudo code de l'algorithme tabou de recherche du MS minimale

```
DEBUT
Calculer le nombre initial d'indice de slot MS avec la formule (3)
Solution =faisable
Tant que (solution==faisable) Faire
    Construire la solution initial Γini ;
    Créer la liste LC à partir de Γini
    Γ =Γini ; NbIt=0 ;// Nombre d'itération
    Tant que (C(Γ) ≠ 0 et NbIt ≠ NbItMax) Faire
        Rt= tête de liste de la liste de conflit LC ;
        Construire N(Γ) en se basant sur Rt;
        Appliquer les règles RT et RA pour trouver la meilleure C(Γ′)
        pour chaque Γ′ ∈ N(Γ) ;
        Γ = Γ′ ;
        Mettre à jour la liste LC ; la liste tabou et NbIt ;
    Fin tant que
    Si C(Γ) = 0
        Solution = faisable ;
        MS=MS-1 ;
    Sinon   solution =infaisable
Fin tant que
MS=MS+1 ;
FIN
```

4. Simulation et analyse des résultats

Les simulations ont été réalisées avec la topologie du réseau NSFNET de 14 nœuds. Le matériel utilisé est un PC de processeur 2,16 Ghz (Dual core) et de mémoire RAM 4 Go. Les codes ont été implémentés avec JAVA (IDE Eclipse). Pour ces simulations, nous avons fixé le nombre maximal d'itération à 100 et considéré des requêtes dont les capacités varient entre 20Gb/s et 100Gb/s. Nous avons procédé dans un premier temps à la comparaison de notre méthode à une autre méta-heuristique basée sur l'algorithme génétique et proposée dans [9]. Le critère de performance est l'indice maximal de slot de fréquence minimum nécessaire sur chaque lien. La figure 1 illustre les résultats. De 20 à 200 requêtes, les deux méthodes ont pratiquement le même indice maximal de slots de fréquence. Par contre au-delà de 200 requêtes comme l'indique la figure 1, notre méthode fournit un résultat meilleur que la méthode génétique. Dans la deuxième simulation illustrée par la figure 2, nous procédons à la comparaison de notre méthode avec deux autres heuristiques, le *MSF* et le *LPF*[2]. Le critère de comparaison est le nombre moyen de slots de fréquence alloué pour 100 requêtes dont les capacités varient entre 20 Gb/s et

100 Gb/s. Il ressort que notre méthode fournit le meilleur résultat car comme l'indique la figure 2, le nombre moyen de slots de fréquence utilisé avec notre méthode est plus petit que le nombre obtenu avec les deux autres heuristiques.

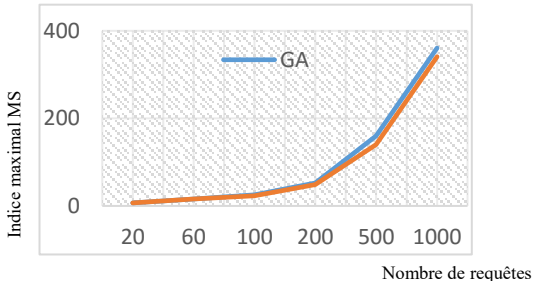


Figure 1 : Indice maximal de slots de fréquence en fonction du nombre de requêtes

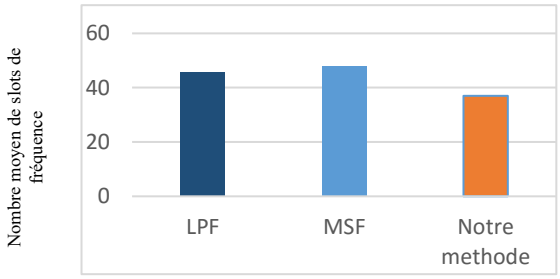


Figure 2 : Nombre moyen de slots fréquence alloués selon la méthode utilisée

5. Conclusion

Dans cet article nous avons proposé une approche de réduction du nombre de slots de fréquence nécessaire à allouer à une instance de requêtes en phase de planification d'un réseau optique flexible. Cette approche basée sur la recherche Tabou permet d'avoir des résultats performants comparativement à d'autres méthodes telles que l'approche basée sur l'algorithme génétique et d'autres heuristiques. Notre approche contribue à l'utilisation optimale des ressources des réseaux optiques flexibles.

Les perspectives sont de vérifier les performances en termes de temps d'obtention des résultats et la prise en compte des trafics dynamiques qui interviennent en phase d'exploitation (opérationnelle) des réseaux.

6. Bibliographie

[1] JINNO, Masahiko, TAKARA, Hidehiko, KOZICKI, Bartlomiej, et al. Spectrum-efficient and scalable elastic optical path network: architecture, benefits, and enabling technologies. Communications Magazine, IEEE, 2009, vol. 47, no 11, p. 66-73.

[2] CHRISTODOULOPOULOS, Konstantinos, TOMKOS, I., et VARVARIGOS, E. A. Elastic bandwidth allocation in flexible OFDM-based optical networks.Journal of Lightwave Technology, 2011, vol. 29, no 9, p. 1354-1366.

[3] ZHOU, Xiang, LU, Wei, GONG, Long, et al. Dynamic RMSA in elastic optical networks with an adaptive genetic algorithm. In : Global Communications Conference (GLOBECOM), 2012 IEEE. IEEE, 2012. p. 2912-2917.

[4] KLINKOWSKI, Miroslaw et WALKOWIAK, Krzysztof. Routing and spectrum assignment in spectrum sliced elastic optical path network. IEEE Communications Letters, 2011, vol. 15, no 8, p. 884-886.

[5] WANG, Yang, CAO, Xiaojun, et PAN, Yi. A study of the routing and spectrum allocation in spectrum-sliced elastic optical path networks. In : INFOCOM, 2011 Proceedings IEEE. IEEE, 2011. p. 1503-1511

[6] OTKIEWICZ, Mateusz, PIÓRO, Michał, RUIZ, Marc, et al. Optimization models for flexgrid elastic optical networks. In : Transparent Optical Networks (ICTON), 2013 15th International Conference on. IEEE, 2013. p. 1-4.

[7] KLINKOWSKI, Mirosław, PIÓRO, Michał, ZOTKIEWICZ, Mateusz, et al.Spectrum allocation problem in elastic optical networks-a branch-and-price approach,2015

[8] CHRISTODOULOPOULOS, Kostas, TOMKOS, Ioannis, et VARVARIGOS, Emmanouel A. Routing and spectrum allocation in OFDM-based optical networks with elastic bandwidth allocation. In : Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE. IEEE, 2010. p. 1-6

[9] GONG, Long, ZHOU, Xiang, LU, Wei, et al. A two-population based evolutionary approach for optimizing routing, modulation and spectrum assignments (RMSA) in O-OFDM networks. IEEE communications letters, 2012, vol. 16, no 9, p. 1520-1523.

[10] GOSCIEN, Roza, KLINKOWSKI, Mirosław, et WALKOWIAK, Krzysztof. A tabu search algorithm for routing and spectrum allocation in elastic optical networks. In : Transparent Optical Networks (ICTON), 2014 16th International Conference on. IEEE, 2014. p. 1-4.

[11] GLOVER, Fred et LAGUNA, Manuel. Tabu Search∗. Springer New York, 2013.

[12] YEN, Jin Y. Finding the k shortest loopless paths in a network. management Science, 1971, vol. 17, no 11, p. 712-716.

Annexe 1 : Mécanismes de notre proposition

- Construction de la solution initiale Γ_{ini}

Pour construire la solution initiale, on attribue à chaque requête R_i un de ses k-plus courts chemins de façon aléatoire. Puis on calcule N_i, le nombre de slots de fréquence nécessaire à chaque requête avec la formule (1). Ce qui permet de choisir le niveau de modulation adapté à la longueur du chemin. Enfin, on choisit pour chaque requête R_i, l'indice du premier slot occupé $\varphi(R_i)$ de façon aléatoire dans l'intervalle$[1, MS]$.

- Construction des solutions voisines $N(\Gamma)$ à une solution Γ

On crée d'abord une liste de conflits dénommée LC constitués des requêtes R_i tel que $C_i > 0$. On désigne ensuite une tête de liste R_t, appartenant à la liste de conflits LC. Enfin on génère une solution voisine Γ' en allouant une nouvelle valeur $(P_t, \varphi(R_t), N_t)$ à la tête de liste R_t, les valeurs des autres requêtes de Γ restent inchangées.

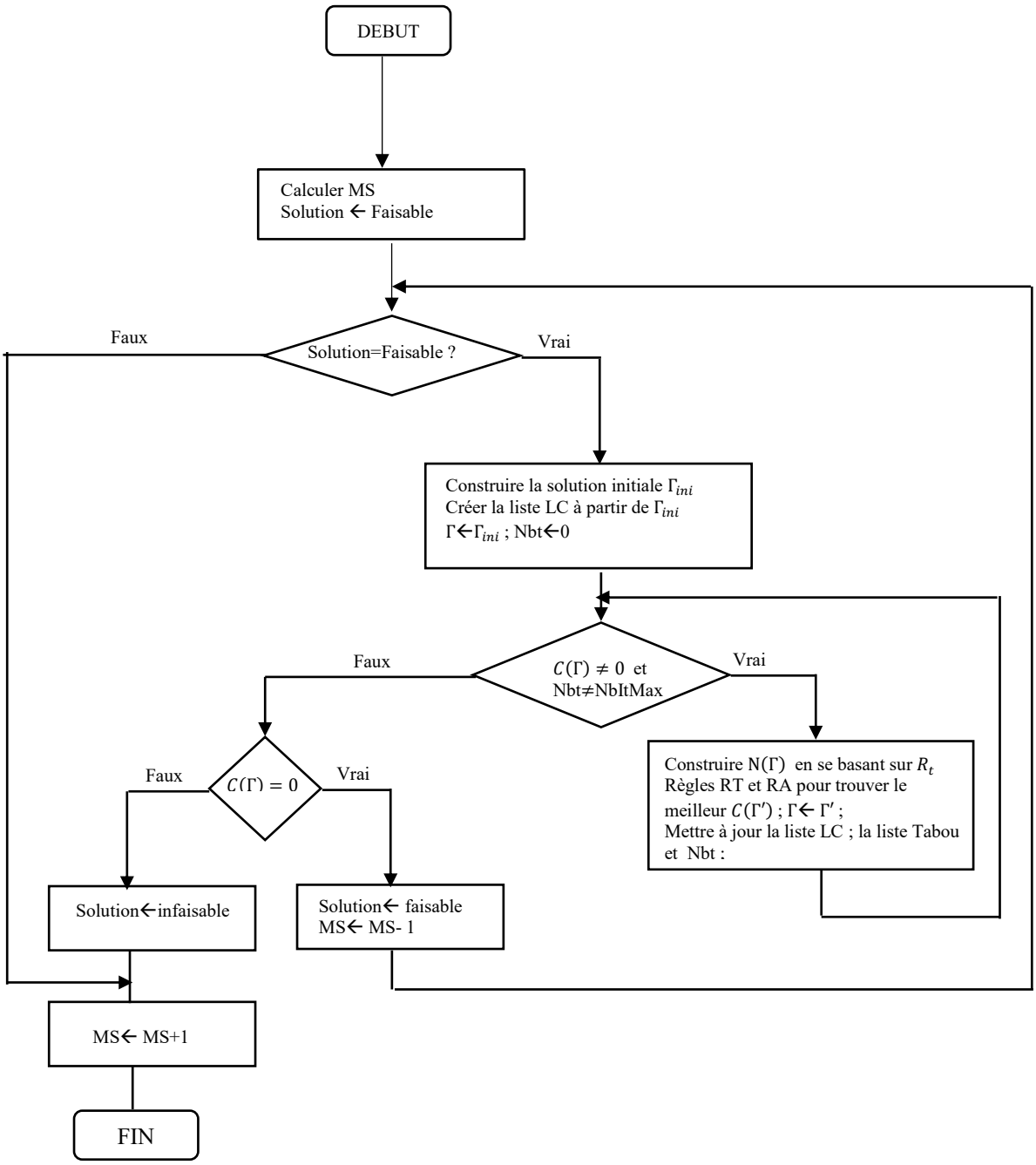
- Mise à jour de la liste de conflits LC

On crée une liste auxiliaire notée Θ constituée des requêtes R_i de Γ qui sont en conflits avec la tête de liste R_t c'est-à-dire : $\Theta = \{R_i \in \Gamma/, \quad P_i \cap P_t \neq \emptyset\}$. Ensuite pour chaque connexion de Θ, on calcule son degré de conflits C_i par rapport à Γ'. Si $C_i = 0$, alors on retire R_i de la liste de conflits LC. De plus si $C_t = 0$ par rapport à Γ', on doit retirer R_t de la liste de conflits et choisir une autre tête de liste de façon aléatoire.

- Règle de la méthode Tabou

Il y a deux règles, la Règle Tabou (RT) et la Règle d'Aspiration (RA).
RT : Dans le but d'éviter des solutions déjà visitées. Une liste Tabou est utilisée pour mémoriser les têtes de listes déjà retirées de la liste des conflits. Cette règle consiste à étiqueter une solution Γ' qui conduit à une tête de liste Tabou. Et cette solution étiquetée ne peut être sélectionnée comme la prochaine solution courante.
RA : La règle d'aspiration est utilisée pour enfreindre la règle Tabou quand une bonne solution existe parmi les solutions étiquetée. Pour chaque itération si le coût de la solution étiquetée est meilleur que le coût des solutions parmi toutes les itérations, l'étiquette Tabou est retirée par cette règle.

Annexe 2 : Organigramme de l'algorithme Tabou de recherche du MS minimale



Evidential HMM Based Facial Expression Recognition in Medical Videos

Arnaud Ahouandjinou[1, 2], Eugène C. Ezin[2, 3], Kokou Assogba[4], Cina Motamed[1], Mikael. A. Mousse[2], Bethel C. A. R. K. Atohoun[1]

[1]Laboratoire d'Informatique Signal Image de la Côte d'opale (LISIC)

Université du Littoral de la Côte d'Opale(ULCO), Bat 2, 50 Rue F. Buisson, 62228 Calais, France

[2]Institut de Formation et de Recherche en Informatique (IFRI), UAC, Bénin

[3]Institut de Mathématiques et de Sciences Physiques (IMSP), Université d'Abomey-Calavi (UAC)

[4]Ecole Polytechnique d'Abomey-Calavi (EPAC), Université d'Abomey-Calavi (UAC)

01 BP 2764 Cotonou, Bénin

ahouandjinou@lisic.univ-littoral.fr

ABSTRACT. A great challenge of practical significance in a recent research topic is to develop computer vision system which can automatically recognize a variety of facial expressions. Such an automated system enables to detect faces, analyzes and interprets facial expressions in a scene although the accomplishment of this task is rather strenuous. There are several related problems: detection of an image segment as a face, extraction of the facial expression information, classification of the expression (e.g., in emotion categories) and their recognition. In this paper, we proposed system that performs facial expression recognition using an Evidential Hidden Markov (Ev-HMM) model in order to manage efficiently the constraints related to facial expression recognition problem. An application of this method as part of improving the monitoring system in medical intensive care units is carried out through to analysis and interpretation of the patient face behavior. The experimental results are very exciting and have shown a promise of our automatic recognition system.

KEYWORDS: Face Detection, Facial Expression Information Extraction, Facial Expression Recognition, Hidden Markov Model, Transferable Belief Model Framework (TB¨M).

RÉSUMÉ. Un défi majeur et d'application importante sur un axe de recherche très actuel est le développement de système automatisé en vision par ordinateur de reconnaître une variété d'expressions faciales. Un tel système intelligent permet de détecter les visages, d'analyser et d'interpréter les expressions faciales dans une vidéo bien que la mise en œuvre d'un tel système est une tâche est plutôt ardue. Des aspects connexes à la réalisation de ce système que sont, la détection d'un segment d'image comme un visage, l'extraction de l'information de l'expression du visage, la classification de l'expression (par exemple, dans les catégories de l'émotion) et enfin la reconnaissance de cette dernière doivent être traités. Dans cet article, nous avons proposé un système de reconnaissance d'expressions faciales en utilisant un modèle de Markov caché évidentiel afin de gérer de façon efficiente les contraintes de reconnaissance d'expression faciale. Une application de cette méthode dans le cadre de l'amélioration du système de surveillance dans les unités de soins intensifs médicaux est effectuée par le biais d'une analyse et d'interprétation du comportement du visage du patient. Les résultats expérimentaux sont très intéressants et ont montré une promesse de notre système de reconnaissance automatique.

MOTS-CLÉS : Détection de face, Extraction d'information d'expression faciale, Reconnaissance d'expression faciale, Modèle de Croyance Transférable**.**

1. Introduction

Facial Expression Recognition (FER) in video scenes is an important topic in computer vision, impacting important applications in areas such as video conferencing, forensics, biomedical applications such as pre or post surgical path planning or clinical improvement prediction, machine vision [1]. The most expressive way humans display emotions is through facial expressions and this latter provides cues about facial behavior. The aim of facial expression recognition methods is to build a system for the classification of facial expressions from continuous video input automatically. Furthermore, development of an automated system that recognizes facial expression is rather a difficult task. There are three main related problems for facial recognition system: detection of an image segment as a face, extraction of the facial expression information, classification of the expression (e.g., in emotion categories) and facial expression recognition. In this paper, we propose a system that detects the face while analyzing and interpreting the behavior of the face of a human in a medical video. Indeed, this system contributed significantly to the recognition of interest events (critical health) that can improve the quality of the patient monitoring system in ICU [3]. An original application is proposed in order to assess the impact of the proposed method for patient monitoring in medical ICUS in cardiology section. Three main contributions can be noted in this work:

i) the first deals with the efficiency of facial expression recognition system based on robust approach by using an evidential HMM. This extension of the HMM allows to take into account at the same time several constraints of the system like physiognomic variability of the human, environment situation-dependent, timing of facial expressions that is a critical factor in the interpretation of expressions. The power of the proposed model lies in the ability and the potential of what the reasoning framework as transferable belief model. ii) The second contribution is related to the combination of facial expression information that uses the maximum intensity of the mouth on the one hand and on the other the maximum intensity of the eyes. iii) Finally, the field of applications is the originality of this work. For this, analysis and understanding of the scene in a video was not done in medical environment. In addition to this, a scenario such as "*fields the pain"* and "*anxious*" in a patient had never been studied.

The paper is organized as follows. In the first section, we describe our proposed method for facial expressions recognition in images sequences. To this end, we present at first, the face detection technique in image and then we explain how the facial expressions features are extracted. Finally, the last part of this section is to expose our robust and flexible algorithm for facial expressions recognition. Thus, after a briefly overview on Transferable Belief Model (TBM) framework, the main steps of evidential hidden Markov model for facial expressions recognition are presented. Section 3 is devoted for applying our approach to recognize facial expressions in medical video. The performance analysis of our method is done by comparing some experimental results with a baseline algorithm applied to various databases in section 4.

2. Proposed Method for facial expression recognition

Detailed review of existing methods on facial expression is seen in [5, 8]. A thorough study of the state of the art of existing methods on facial expression is

proposed by Maja Pantic in IEEE transactions on pattern analysis and machine intelligence [8]. Since the mid-1970s, many methods have been proposed for facial expression analysis and their recognition from either static facial images or image sequences. Among these, we have the approaches based on active contours, robust appearance filter, probabilistic tracking, adaptive active appearance model and active appearance model [1]. The aim of this section is to explore the issues in design and implementation of a system that could perform automated facial expression analysis. In general, three main steps can be distinguished for solving this problem. First step, before the analysis of facial expression, the face must be detected in a scene. The next step is to devise mechanisms for extracting the facial expression information from the observed facial image or image sequence. The final step is to define some sets of categories, which will be used for facial expression classification and/or facial expression interpretation, and to devise the mechanism of categorization. To this end, most facial expression recognition systems focus on only six basic expressions (i.e., joy, surprise, anger, sadness, fear, and disgust) proposed in the work of Darwin at the beginning [9] and more recently Ekman [9]. In everyday life, however, these six basic expressions occur relatively infrequently, and emotion or intent is more often communicated by subtle changes in one or two discrete features, such as tightening of the lips which may communicate anger. Facial expression recognition or human emotion analysis remains a very daunting task.

2.1. Face Regions Detection

In response to real-time system development and the homogenous processing system for facial expression recognition, we used Hidden Markov Model (HMM) to detect face in video sequence. HMM consists of two interrelated processes: (1) an underlying, unobservable Markov chain with a finite number of states, a state transition probability matrix and an initial state probability distribution and (2) a set of probability density functions associated with each state. The used approach is based on the technique proposed by A. Nefian and Monson Hayes III in [4]. This technique involves the extraction of the face features in order to detect it. Each face image of width W and height H is divided into overlapping blocks of height L and width W. The amount of overlap between consecutive blocks is P. T is the number of observations which denotes the number of blocks extracted from each face. T is generated using equation 1:

$$T = \frac{H - L}{L - P} + 1 \tag{1}$$

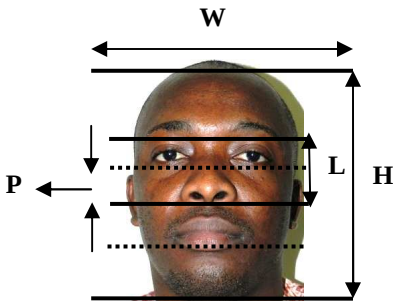


Figure 1 is an illustration of face image parameterization and blocks extractions.

Particular facial regions such as: hair, forehead, eyes, nose and mouth come in a natural order from top to bottom, even if the images are taken under small rotations in the image plane. Each of these facial regions is assigned to a state from the left to the right topology of HMM. Note that, the state structure of the face model and the non-zero transitions probabilities are shown in Figure 2.

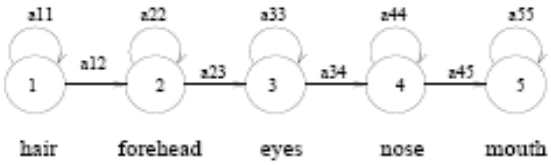


Figure 2: Left to right HMM for face recognition [4].

Two main steps are used by HMM to detect and recognize faces. Among these steps, we have training the face model and their recognition. For the training step, we use an HMM face model to represent each individual in the database. A set of five images representing different instances of the same face are used to train each HMM. In the recognition phase, the probability of the observation vector given each HMM face model is computed after extracting the observation vectors as in the training phase. A face image t is recognized as face k if:

$$P\left(O^{(t)}\left|\lambda_{\mathrm{k}}\right.\right)=Max_{n}P\left(O^{(t)}\left|\lambda_{n}\right.\right) \tag{2}$$

After the face detection, the facial expression recognition system performs the mouth and the eye region feature extraction using the pixel intensity code value to recognize facial expression in images sequences.

2.2. Feature Extraction Process from Eye and Mouth Region

This work exploits the temporal intensity change of expressions in videos for facial expression recognition through the HMM. Considering the intensity scale of the different facial expressions, each person has his/her own maximal intensity of displaying a particular facial action. We combine the Mouth region intensity Code Value namely MICV [1] and the Eye region Intensity Code Value namely EICV as features for facial expression recognition.

In this section, we describe how we compute the eye and the mouth region intensity coded value (EICV/MICV). The E/MICV for eye and the mouth region which characterizes the intensity variations between blocks that corresponds respectively to the eye and the mouth region in a video frame is computed using a simple procedure that divides a mouth region into blocks and creates a code called EICV and MICV which represents the intensity difference between blocks in a frame. Eq. (3) illustrates the generation of proposed MICV feature [1]. i and j represent the i_{th} and j_{th} blocks in a frame. MICV is generated using Equation 3 [7]:

$$y\left[(i-1)25+\mathrm{j}-\frac{\mathrm{i}(\mathrm{i}+1)}{2}\right]=\begin{cases}1 \text{ if } \mathrm{x}(\mathrm{i})>x(j)\\ 0 \text{ Otherwise}\end{cases} \quad (3)$$

$$1\le i\le 25, 2\le j\le 25 \text{ and } \mathrm{i}<\mathrm{j}$$

Where *x(i), x(j)* are the average intensities of the i_{th} and j_{th} blocks respectively. To generate the MICV, for example, the frame is divided into 5 x 5 blocks to generate the feature vector. Figure 3 shows the detected mouth region and the 5 x 5 representation of mouth region.

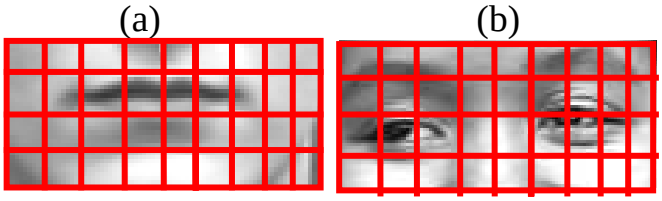


Figure 3: (a) representation of mouth region and (b) the same for eye region

To generate the MICV [1], for example, the frame is divided into 5 x 5 blocks to generate the feature vector. Fig. 3a and 3b show the detected eye and mouth region with both their representation. Each block in a frame is compared with every other block to generate EICV and MICV using "Eq. (4)". For example, if the image is divided into 5 x 5 blocks, then "Eq. (4)" generates 300 dimensional feature vectors. First element in the feature vector compares the intensity of 1st and 2nd block; second element compares the intensity of 1st and 3rd block and so on. The distance or error between the two comparison codes $p=(P_1,P_2,P_3,.....,P_n)$ and $q=(q_1,q_2,q_3,.....,q_n)$ can be calculated using equation 4:

$$d=\sum_{k=1}^{n}(p_k \oplus q_k) \quad (4)$$

2.3. Robust and Flexible Facial Expression Recognition

The proposed system for facial expressions recognition is using an evidential Hidden Markov Model (EvHMM) developed by E. Ramasso in [6] and first introduced by [9]. This version of HMM is based on an extension of probabilistic reasoning framework to the evidential. This new reasoning framework is very generic and powerful to develop tools to support any type of application with better management of uncertainty and imperfect data. In addition, it is possible to combine information with the careful fusion rules and operators. In this work, we proposed a new and robust approach for event recognition in videos sequences. A substantial benefit of belief functions is their versatility and efficiency in the information fusion process. Transferable Belief Model is a very suitable tool for information combination as it takes into account the nature and quality of sources to provide noisy information [6]. Another advantage of the reasoning part also lies in its ability to manage the imperfections of the data in order to estimate the best accuracy recognition system. In the related work, we noted that facial expression recognition from still image has less precision with respect to video sequence because a single image offers much less information than a sequence of images for expression recognition processing.

Feature classification is performed in the last stage of our automatic facial expression analysis system. Hidden Markov Models (HMMs) have been widely used to model the temporal behaviors of facial expressions from image sequences. This work exploits HMM to recognize facial expression. Three basic facial actions are defined (neutral, smile, eyes closed and raised eyebrows) and five emotional or facial expressions (neutral, happy, anxious, painful, and disgusted) can be recognized by the system. For each facial expression, we use an Evidential HMM for training the model and afterwards to recognize its. We define five HMMs to recognize the facial expression which are *"neutral", "happy", "disgust", "pain" , and "anxious"*. Facial expressions features such as EICV and MICV are computed in probabilistic quantity. And then, we have combined in the belief mass two main information estimated on the eye region features named (Eye Intensity Probability Value: EIPV) and the mouth region features called (Mouth intensity Probability Value: MIPV). The result of EIPV and MIPV combination design Facial Expression Code Value (FECV) is given as input to estimate Ev-HMM parameter from the learning step. The remainder of this section describes the HMM learning process and the recognition of facial expressions through two steps. These steps are implemented using the beliefs parameter input in probabilistic HTK toolkit for the expression recognition.

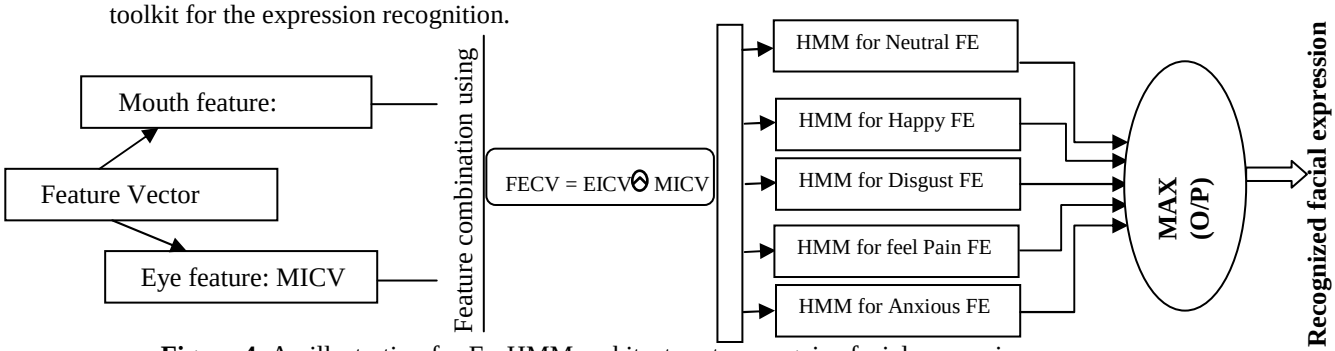


Figure 4: An illustration for Ev-HMM architecture to recognize facial expressions

3. Applying Method to Recognize Facial Expression in Medical Video

In this section, we present on one hand, an appliance of evidential Markov model for facial expression recognition and on the other hand, the experimental results on real-world facial expression dataset. In addition, we described the used datasets and presented the experimental results of even the performance analysis of the proposed approach compared it to other existing methods. Our Algorithms have been implemented using Matlab, C/C++ using OpenCV library.

3.1. Tested Data Setup

In order to test the algorithms described in the previous sections we use two different databases, a database collected by us and the Cohn-Kanade [10] AU code facial expression database. Full details of this database are given in [10]. The second test database is ours. The data collection method is described in detail in [3], our database has been collected from the experimental video-surveillance system that we installed in the cardiology department at the hospital (have collected roughly 47 videos sequences for three activities with 1500 frames/sequences. An observation sequence is recorded

every one and a half second from the 25fps video. The duration of the video sequences is 300 seconds with an average length of circa 90 seconds. In this database, we have the subjects that were instructed to display facial expressions corresponding to the five types of emotions such as *"neutral", "happy", " disgust", "pain"* , and *"anxious"*. Four basic actions (***neutral, smile, eyes closed and raised eyebrows***) detected over the face feature extraction step are used through like input data of the Evidential HMM to recognize these facial expressions.

3.2. Experiments Results 1: *Using Cohn-Kanade AU database*

All the tests of the algorithms are performed on a set of five persons, each one displaying five sequences of each of the five emotions, and always coming back or not to a neutral state between each emotion sequence. The sampling rate of the video sequence was 30 Hz, and a typical emotion sequence is about 150 samples long (5s). Figure 5 in appendix, shows one frame of each emotion for each subject. We used the sequences from a set as test sequences and the remaining sequences were used as training sequences. In this case, we performed person dependent experiments, in which part of the data for each subject was used as training data, and another part as test data. Table 1 show the recognition rate of the test for each HMM version. Note that the results obtained with this database are much better than the ones obtained with our database. This is because in this case we have more training data. Furthermore, it is observed that among the five expressions *"happy"* expression is well (98% recognition rate) recognized than the others (between 70% and 85% recognition rate). It can also be seen that the evidential HMM with temporal constraints, achieves the best recognition rate (and improves it in some cases) compared to the other used version HMM, even though the input is segmented as continuous video. The other expressions are greatly confused with one another other. See illustration results in Table 1 in Appendix.

3.3. Experiments Results 2: *Using Medical videos database*

Our experimental data were collected in an open recording scenario, where the patient was asked to display the expression corresponding to the emotion being induced. This is a simulation process for generating facial expressions in medical context. Although we are aware that this assumption does not take into account all the constraints related to the real conditions of facial expressions data collection, we think that, the experimental result achieved shows involved significantly the technological progress. For complex and highly sensitive applications such as patient monitoring in medical UCIs, power, robustness and efficiency of the proposed model stands out and improves very significant way the performance of the expressions recognition system facial. The specific facial expressions recognition rate to the medical context such as feel the pain and the patient is in anxious condition depends on the performance of the facial feature extraction system for the detection of facial expressions basic such as smile, eyes closed, eyebrows raised and finally neutral. In average, the best results of facial expression recognition were obtained using Ev-HMM. The temporal layer assumption gives a significant improvement in recognition rate comparing with standard probabilistic HMM. In appendix, find in Table 2 & 3, the results reporting the facial expression recognition rate reached depending on the various kinds of HMM we tested. In this used case of Ev-HMM, *"Happy"* was detected with over 96% accuracy and Disgust with over 83% accuracy. Whereas, the patient's behavior like fells the pain and anxious state are recognized at respectively 78 % and 70%.

5. Conclusion

We have developed in this work, a computer vision system that automatically recognizes a series of complex facial expressions. Our recognition system applied to psychological research in medical field. In the first instance, the proposed approach has been tested on a generic [14] database of facial expression to assess the system is performance and efficiency. More specifically, the proposed system was used to recognize the patient's specific behaviors closely linked to his facial expressions and emotions (resentment pain and mental state of anguish) in cardiological ICUs. A Robust and powerful approach for automatic facial recognition expression using HMM in belief framework is presented. The proposed work is able to detect human faces over extracting face features using HMM tool by segmenting face from the real time video. Among the facials expressions, *happy* and *disgust* expressions has been recognized with an accuracy of 96% but expressions *neutral* and *disgust* cannot be distinguished well. Thank to our method, because it provides better rate recognition with complex expressions in a medical environment such as the issue of pain and patient anxious are not easy to recognize. Nevertheless, our system has allowed us to recognize these two expressions with a rate of about 83% on average. Hence the future work aims to apply the feature extracted in this work to the forehead and noise region and also considering more number of expressions. In addition, we think to take into account a generic maximal intensity for all people because that is the lack in current model, each person has his/her own maximal intensity of displaying a particular facial action.

6. References

[1] A. Punitha, M. K. Geetha, HMM Based Real Time Facial Expression Recognition, *International Journal of Emerging Technology and Advanced Engineering*, Volume 3, Special Issue 1, January 2013.

[2] A. R. M. S. Ahouandjinou, C. Motamed, E. Ezin,Credal Human Activity Recognition Based-HMM by Combining Hierarchical and Temporal Reasoning, *The fifth International Conference on ImageProcessing Theory, Tools and Applications IPTA'15*, November 10-13,Orléans – France, 2015.

[3] A. R. M. S. Ahouandjinou, C. Motamed, and E. Ezin, Activity Recognition Based on Temporal HMM for Visual Medical Monitoring Using a Multi camera System, *Special issue of ARIMA Journal, Special issue of CARI'14*, Volume 21 – 2015.

[4] A. V. Nefian, Hidden markov model for Face Recognition, *Phd thesis*, Georgia Institute of Technology, 1999.

[5] B. Fasel, and J. Luettin, Automatic facial expression analysis: a survey, *Pattern Recognition* 36, 2003.

[6] E. Ramasso. Contribution of belief functions to Hidden Markov Models., In IEEE Workshop on Machine Learning and Signal Processing, pages 1–6, Grenoble, France, October 2009.

[7] J. Lien. Automatic recognition of facial expressions using hidden Markov models and estimation of expression intensity. *PhD thesis*, Carnegie Mellon University, 1998.

[8] M. Pantic, and L. Rothkrantz, Automatic analysis of facial expressions: the state of the art, *IEEE Trans. Pattern Analysis and Machine Intelligence* 22 (12), 2000.

[9] P. Ekman. Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. Psychological Bulletin, 115(2):268–287, 1994.

[10] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis, 2000.

7. Appendices

7.1. Evidential HMM training and Facial Expression Recognition

In this work, TBM is used to combine the mouth/eye features parameter of facial expression recognition in single parameter named Facial Expression Code Value (FECV). For HMM training step, FECV is computed after the EPIV and MIPV are extracted using respectively eye and mouth region intensity from each frame in the video sequence and is given as input to estimate the parameters of Ev-HMM. We propose to use a Credal version of HMM algorithm proposed in my previous paper [2] in order to handle the spatial and temporal variability and also the uncertainty existing over the machine learning task [6]. To this end, regrouping components into states is made automatically by maximizing likelihood, and a relevant regrouping implies a better recognition of states. Given observations sets how to adjust the HMM parameters to maximize the training set likelihood? Facial Expression Recognition step concerns the test of the data against the model built. The Ev-HMM classification scheme used in this approach is shown in Fig. 4. Initially, separate Ev-HMMs are used for each expression. FECV is fed as input into the Ev-HMM. Finally, the maximum output obtained is considered as the output expression. Upon completion of learning step, the properly so called recognition stage is carried out.

7.2. Results Performance Analysis: *Cohn-Kanade AU database*

See in following Figure 5, the examples of images from the video sequences used in the experiment.

Figure 5: Used data of Cohn-Kanade AU database

Facial expressions/HMM Model	Classic HMM	Temporal HMM	Hierarchical HMM	Evidential HMM
Neutral	70,00	70,00	72,00	80,00
Happy	80,00	85,00	85,00	98,00
Disgust	60,00	62,00	63,00	70,00
Surprise	70,00	80,00	80,00	85,00

Table I: Facial Expression recognition rate for Cohn-Kanade AU database (average in %)

7.3. Results Performance Analysis: *Medical videos database*

See in following Figure 6, the examples of images from the medical video sequences used in the experiment.

Figure 6: Our used data gathered from experiment video surveillance system in UCIs.

Facial expressions/HMM Model	Classic HMM	Temporal HMM	Hierarchical HMM	Evidential HMM
Neutral	70,00	75,00	78,00	80,00
Happy	76,00	80,00	80,00	85,00
Disgust	65,00	70,00	72,00	96,00

Table II: Facial Expression recognition rate for our test database (average in %)

Facial expressions/HMM Model	Classic HMM	Temporal HMM	Hierarchical HMM	Evidential HMM
Neutral	52,00	55,00	55,00	80,00
Happy	60,00	80,00	80,00	96,00
Disgust	53,00	70,00	70,00	83,00
The patient feels pain	54,00	55,00	55,00	78,00
The patient Anxiety	50,00	52,00	55,00	70,00

Table III: Facial Expression recognition rate for our test database (average in %)

8. Acknowledgments

The authors would like to thank Professor Hippolyte AGBOTON who is in charge of the cardiology section research team. We also acknowledge each individual (Fréjus LALEYE, Ulrich AKPACA and Arcadius ABRAHAM) appearing in our face database.

Tatouage vidéo dynamique et robuste basé sur l'insertion multi-fréquentielle

Sabrine Mourou, Asma Kerbiche et Ezzedine Zagrouba

Laboratoire RIADI- Equipe de recherche en Systèmes Intelligents en Imagerie et Vision Artificielle SIIVA

Institut Supérieur d'Informatique, Université Tunis El Manar
2 Rue Abou Raihane Bayrouni, 2080, Ariana
TUNISIE

mourou_sabrine@live.fr - asma.kerbiche@gmail.com - ezzeddine.zagrouba@fsm.rnu.tn

RÉSUMÉ. Dans ce papier, nous proposons une nouvelle approche de tatouage vidéo robuste et invisible basée sur l'insertion dynamique et multi-fréquentielle. Dans cette approche, la marque est insérée dans les images qui présentent un fort changement de plan détectées à partir de la vidéo originale. Le choix de l'insertion dans ces images est basé sur leur robustesse face aux attaques vidéo les plus importantes telles que la collusion et la suppression d'images. En plus, pour maximiser la robustesse contre les attaques usuelles nous avons opté pour une insertion multi-fréquentielle basée sur les trois transformées DWT, SVD et ACP. Les résultats expérimentaux, ont montrés que notre approche a permis d'obtenir une meilleure invisibilité et robustesse face aux maximum d'attaques.

ABSTRACT. In this paper, we propose a new robust and imperceptible video watermarking approach based on the dynamic multi-frequency insertion. This approach inserts the mark in images that have many plan changes detected from the original video. The choice of the insertion on these frames is based on their robustness against the most important video's attacks such as collusion and frame suppression. In addition, to maximize the robustness against usual attacks, we opted for a multi-frequency insertion based on the three transforms DWT, SVD and PCA. Experimental results showed that this approach allows obtaining a good invisibility and robustness against the maximum of attacks.

MOTS-CLÉS : Tatouage Vidéo, Robustesse, Invisibilité, Ondelette, SVD, ACP.

KEYWORDS: Video watermarking, Robustness, Invisibility, DWT, SVD, PCA.

1. Introduction

L'échange et la portabilité des donnés multimédia (texte, image, vidéo etc.) deviennent de plus en plus fréquents et accessibles à tout le monde et ce, grâce à l'avancement technologique matériel (disque dur externe, mémoire flash) et logiciel (techniques de compression) sans oublier Internet et les réseaux d'une façon générale. Au cours de cette évolution, les techniques d'espionnage et les fraudes se sont multipliées en parallèle. En effet, les documents numériques sont menacés par le piratage, la modification et le copiage illégal. Pour satisfaire ce besoin, le tatouage est apparu pour garantir l'authenticité et assurer un accès autorisé. Dans ce contexte, nous avons développé une méthode de tatouage vidéo qui maximise le compromis robustesse-invisibilité et qui est efficace contre les attaques illicites d'une image fixe et les attaques spécifiques aux vidéos. Dans la section suivante, un état de l'art du tatouage vidéo sera brièvement présenté. Ensuite, l'approche de tatouage vidéo basée sur l'insertion dynamique et multi fréquentielle sera détaillée. Les évaluations expérimentales feront l'objet de la section 4. Enfin, les conclusions ainsi que quelques perspectives seront présentées.

2. Etat de l'art du tatouage vidéo

Plusieurs algorithmes de tatouage vidéo ont été développés. En effet, nous avons choisi de classifiées ces méthodes en se basant sur deux critères principaux qui sont le type d'insertion et le domaine d'insertion. En ce qui concerne le premier critère, nous avons distingué deux classes. La première classe présente les méthodes basées sur l'insertion statique [1] où la signature est insérée dans toutes les images composant la vidéo. La deuxième classe présente les méthodes basées sur l'insertion dynamique [2] où la signature est insérée dans quelques images de la vidéo. La table 1 présente une étude comparative de ces deux classes.

Classes	**Robustesse**					**Invisibilité**
	Compression	Permutation	Suppression	Insertion	Collusion	
Statique	Robuste	-	-	-	-	+
Dynamique	Robuste	Robuste	Robuste	Robuste	Robuste	+

Table 1. *Tableau comparatif des différentes méthodes selon le type d'insertion*

Concernant le deuxième critère, nous distinguons quatre classes : spatiale, mono-fréquentielle, multi-fréquentielle et hybride. L'insertion dans le domaine spatial [3] se

fait directement sur les valeurs de pixels de différentes images de la vidéo, contrairement à l'insertion mono-fréquentielle [4, 5, 6, 7] ou la marque est insérée sur les coefficients d'une seule transformée appliquée sur les images de la vidéo. Concernant l'insertion multi-fréquentielle [8, 9] la marque est ajoutée sur les coefficients de plusieurs transformées appliquées. Et pour finir, l'insertion hybride [10] consiste à insérer la marque sur les coefficients dans une combinaison entre le domaine spatial et le domaine fréquentiel qui peut être mono-fréquentiel ou multi-fréquentiel. Une étude comparative de ces classes est présentée dans la Table 2.

Classes		**Transformations géométriques**	**Cropping**	**Filtrage**	**Bruitage**	**Compression**	**Invisibilité**
Spatiale [LSB]		-	-	-	-	-	+
Mono	DWT	Robuste	-	Robuste	Robuste	-	+
	SVD	-	-	Robuste	Robuste	Robuste	+
	DCT	Robuste	-	-	Robuste	Robuste	+
	PCA	Robuste	Robuste	Robuste	-	-	+
Multi	DWT-DCT-SVD	Robuste	Robuste	Robuste	Robuste	Robuste	++
	DWT-DFT-SVD	Robuste	Robuste	Robuste	Robuste	Robuste	++
Hybride [LSB-DWT]		-	-	-	Robuste	Robuste	+

Table 2. *Tableau comparatif des différentes méthodes selon le domaine d'insertion*

3. Approche proposée

En se basant sur l'état de l'art, nous avons constaté que chaque méthode de tatouage vidéo présente ses propres avantages et inconvénients. Cependant, nous avons remarqué que l'insertion dans le domaine multi-fréquentiel assure une robustesse face aux différentes catégories d'attaques avec une meilleure invisibilité. En plus, l'insertion dynamique comme étant cible d'insertion permet une robustesse aux attaques les plus importantes telles que la collusion, la permutation d'images, la suppression d'images et la compression. L'idée que nous avons conçue consiste alors à profiter des avantages de ces deux classes en tatouant les images de la vidéo qui présentent un changement de scène en utilisant une insertion multi-fréquentielle. Le schéma général de cette approche est présenté dans (figure 1.(a)) et se décompose en sept étapes.

- **Détection des changements de scène** : nous avons choisi comme une insertion dynamique, l'insertion dans les images qui présentent un changement de scène par rapport aux autres images composant la vidéo. Pour la détection de ces images nous

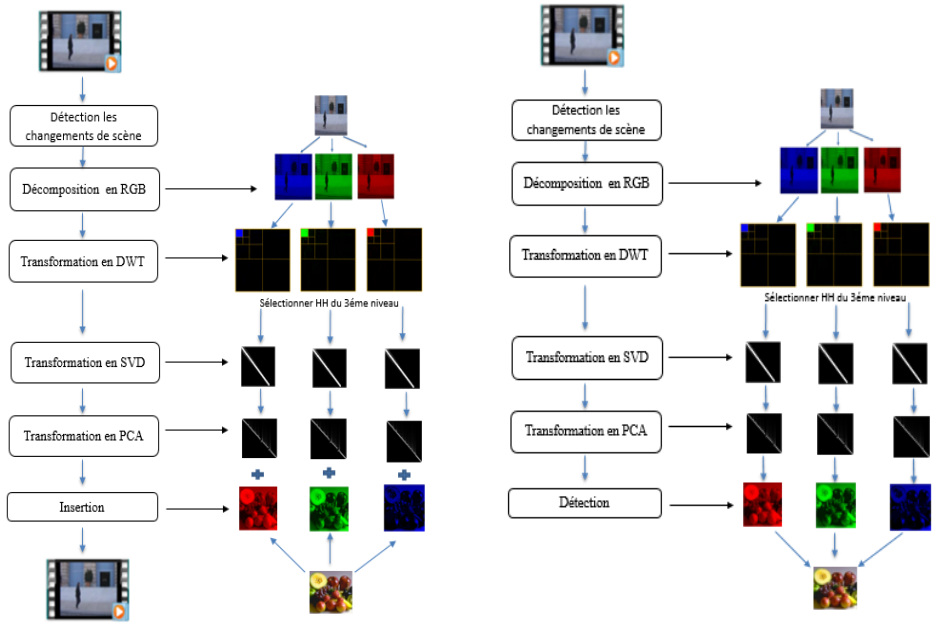
avons utilisé l'algorithme [12] qui est basée sur la détection de changement de scène par les histogrammes. En effet, afin de détecter les changements, une différence est calculée entre les deux histogrammes des deux images successives.

- **Décomposition en RGB :** Chaque image résultante est décomposée en trois composantes RVB (rouge, vert et bleu). L'insertion de la marque se fait sur chaque image couleur R, V et B.
- **Transformation en ondelette :** Nous appliquons, par la suite, une décomposition en ondelette jusqu'au 3ème niveau sur chaque image couleur et nous choisissons la sous bande diagonale de haute fréquence, de chaque image couleur pour l'insertion.
- **Décomposition en valeurs singulières :** Après avoir obtenu les sous blocs (HH3R(1);HH3V (2);HH3B(3)) de chaque composante de l'image, nous appliquons la transformée SVD sur chaque sous blocs. Nous choisissons par la suite la matrice S de chaque sous bande HH de chaque composante couleur.
- **Transformation avec l'ACP :** Dans cette étape, nous appliquons la transformée ACP sur chaque matrice (SR, SV, SB) résultante de l'étape précédente. Comme résultat, nous obtenons les coefficients principaux des composantes principales (YR, YV, YB).
- **Insertion :** Cette étape consiste à insérer la marque dans les images. Pour cela, nous décomposons la marque en trois composantes couleurs : rouge, vert et bleu (WR, WV, WB). Par la suite, nous additionnons chaque composante couleur de la marque avec chaque bloc résultant de la dernière transformation ayant la même couleur.

$$\begin{aligned} Y^T_R &= Y_R + W_R * \text{Alpha} \\ Y^T_V &= Y_V + W_V * \text{Alpha} \quad (1) \\ Y^T_B &= Y_B + W_B * \text{Alpha} \end{aligned}$$

- **Transformation inverse :** Dans cette étape, nous effectuons l'inverse de chaque transformée : PCA puis SVD et enfin DWT et après reconstruction des images de la vidéo nous obtenons notre vidéo tatouée.

Concernant l'étape de détection, elle se décompose en différentes étapes (figure 1.(b)) dont les quatre premières sont identiques à celle de l'insertion. En effet, après la détection des images qui présentent un changement de scène dans la vidéo tatouée et originale et l'application des trois transformées DWT-SVD-ACP sur les trois composantes couleurs (R, V et B) de ces images, nous appliquons une soustraction de chaque bloc résultant des transformations sur la vidéo tatouée avec chaque bloc résultant des transformations de la vidéo originale pour extraire la marque.



(a) (b)

Figure 1. *Schéma général de l'approche proposée, (a) Schéma général de l'insertion, (b) Schéma général de la détection.*

4. Résultats expérimentaux et étude comparative

Pour évaluer notre approche proposée, nous avons choisi deux séquences vidéo couleur qui sont la séquence de Stefan et la séquence Granguardia en se basant sur les deux contraintes invisibilité et robustesse. Ensuite, nous avons comparé nos résultat obtenus avec d'autres méthodes existantes qui sont : la méthode de Saurbh & al [1] qui ont utilisé une insertion multi-fréquentielle basée sur DWT-PCA, la méthode de Al katib & al [11] qui ont utilisé une insertion basée sur DWT-SVD et la méthode Masoumi & al [2] basée sur l'insertion dynamique DWT.

4.1. Invisibilité

Concernant l'invisibilité, la vidéo originale et celle tatouée sont invisible à l'œil nue. Afin de prouver cette invisibilité, nous avons calculé les valeurs de PSNR (Peak Signal Noise Ratio) et nous l'avons comparé avec d'autres méthodes existantes. Les résultats

obtenues (56.194dB « Granguardia » et 53dB « Stefan ») prouvent une meilleure qualité visuelle (figure 2).

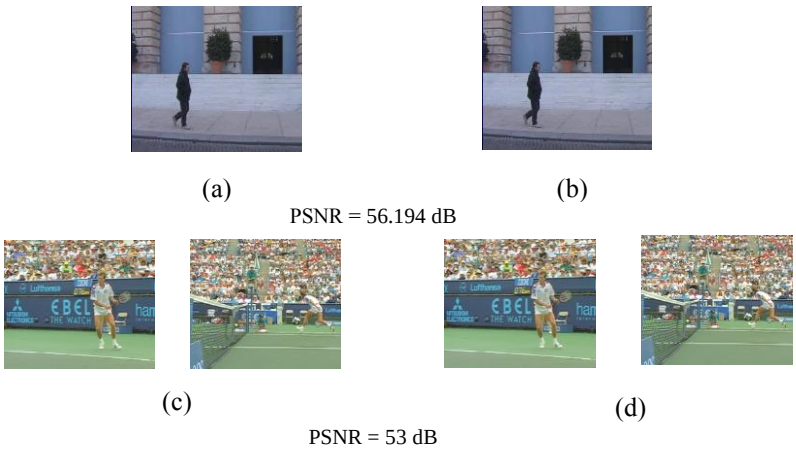


Figure 2. *(a) et (c) Image originale, (b) et (d) Image tatouée*

En comparant notre approche avec la méthode de Masoumi & al. [2], Saurbh & al [1] et Al katib & al [11], nous constatons que notre algorithme permet d'obtenir une meilleure qualité visuelle de la vidéo tatouée comme le montre la Table 3

	Méthode proposée	**Saurbh & al [1]**	**Al katib & al [11]**	**Masoumi & al [2]**
PSNR	56.194 dB	45.41 dB	48.13 dB	36.77 dB

Table 3. *Tableau comparatif d'invisibilité entre notre approche et d'autres approches existantes.*

4.2. Robustesse

Concernant la robustesse, nous avons testé la robustesse de notre approche contre plusieurs attaques : compression MPEG-4, collusion, permutation, suppression, bruit, filtrage, cropping, transformation géométrique, les résultats obtenus sont illustrés dans la Table 4.

Méthodes	Transformation géométriques	Filtrage	Bruit	Cropping	Permutation	Suppression	Collusion	Compression
Méthode Proposée	Rotation Changement d'échelle Translation	Moyenneur Gaussien	Sel et poivre Speckle Gaussien Poisson	Robuste	Robuste	Robuste	Robuste	Robuste
[1]	Rotation Changement d'échelle Translation	-	Sel et poivre	-	-	-	-	-
[2]	-	-	-	-	Robuste	Robuste	Robuste	Robuste
[11]	Rotation		Sel et poivre Gaussien			Robuste		Robuste

Table 4. *Tableau comparatif de robustesse entre notre approche et d'autres approches existantes.*

D'après le tableau comparatif, notre approche proposée a vérifié une meilleure robustesse face aux attaques testées. En effet, elle est robuste face au plus importantes attaques qui touche le flux vidéo.

5. Conclusion

Dans ce papier, nous avons proposé une nouvelle méthode de tatouage vidéo basée sur l'insertion multi-fréquentielle en utilisant les trois transformée DWT-SVD-ACP dans les images qui présentent un fort changement de plan dans la vidéo. Les résultats expérimentaux obtenus ont prouvé que notre algorithme permet d'obtenir une meilleure invisibilité et une forte robustesse contre les plus importantes attaques qui touche le flux vidéo comme les transformations géométriques, le bruit, le filtrage, le Cropping, la suppression et la permutation d'image, la collusion, et la compression MPEG-4. Comme perspective à ce travail, nous proposons d'améliorer notre approche en insérant la marque dans des régions d'intérêt spécifiques, qu'on détectera, ce qui d'après la littérature [13] peut apporter plus de robustesse en utilisant le domaine multi fréquentielle.

6. Bibliographie et biographie

[1] Phadtare Saurabh, Dhebe Pooja, Bobade Sharayu and Jawalkar Nishigandha, *Video Watermarking Scheme Based on DWT and PCA for Copyright Protection*, Journal of Computer Engineering, 9(4): 18-24, Mars April 2013.

[2] Majid Masoumi and Shervin Amiri, *A Blind Video Watermarking Scheme Based on 3D Discrete Wavelet Transform*, International Journal of Innovation, Management and Technology, 3(4): 487-490, 2012.

[3] M.George, J.-v.Chouinard and N. Georganas, *Digital watermarking of images and video using direct sequence spread spectrum techniques*, Electrical and Computer Engineering, IEEE Canadian Conference, vol.1, pages 116-121, May - 1999.

[4] Saket Kumar, Ashutosh Gupta, Ankur Chandwani, Gaurav Yadav and Rashmi Swarnkar, *RGB Image Watermarking on Video Frames using DWT,* IEEE 5th International Conference-Confluence The Next Generation Information Technology Summit (Confluence), pages 675-680, 2014.

[5] Ruizhen Liu and Tieniu Tan, *An SVD based watermarking scheme for protecting rightful ownership*, IEEE Circuits & Systems Society, 4(1):121-128, 2002

[6] Sonjoy Deb Roy, Xin Li, Yonatan Shoshan, Alexander Fish, Orly Yadid-Pecht, *Hardware Implementation of a Digital Watermarking System for Video Authentication*, IEEE Transactions on Circuits and Systems for Video Technology, 23(2):289-301, 2013.

[7] Hanane H.Mirza, Hien D.Thai, Yasunori Nagata and Zensho Nakao, *Digital Video Watermarking Based on Principal Component Analysis*, IEEE Innovative Computing, Information and Control, pages 290-294, 5-7 Septembre 2007.

[8] Nandeesh B, Lohit S Meti, Manjunath G K, *A Robust Non-Blind Watermarking Technique for Color Video Based on Combined DWT-DFT Transforms and SVD Technique*, Information Technology and Computer Science, pages 59-65, 2014.

[9] P. Satyanarayana, C. N. Sujatha, *Analysis of Robustness of Hybrid Video Watermarking against Multiple Attacks,* International Journal of Computer Applications, 118(2):12-19, May 2015

[10] Hui-Yu Huang, Cheng-Han Yang and Wen-Hsing Hsu, *A video watermarking technique based on pseudo 3-D DCT and Quantization Index Modulation*, IEEE Transactions on Information Forensics and Security, 5(4):625-637, 2014.

[11] Tahani Al-Khatib, Ali Al-Haj, Lama Rajab and Hiba Mohammed, *A Robust Video Watermarking Algorithm*, Journal of Computer Science, 4(11): 910-915, 2008.

[12] Kintu Patel, Mukesh Tiwari and Jaikaran Singh, *Video Watermarking using Abrupt Scene Change Detection*, International Journal of Computer Technology and Electronics Engineering, Volume 1, Issue 2, pages 187-190.

[13] Asma Kerbiche, Ezzedine Zagrouba, Saoussen Ben Jabra, *Tatouage video robuste basé sur les regions d'intérêt*, CARI, Volume 1, 2012.

Dynamic Pruning for Tree-based Ensembles

Mostafa EL HABIB DAHO, Mohammed El Amine LAZOUNI, Mohammed Amine CHIKH

Département d'Informatique
Université de Tlemcen
Ville Tlemcen
Pays Algérie
mostafa.elhabibdaho@mail.univ-tlemcen.dz

ABSTRACT. In this paper, we propose a new dynamic pruning based on trees selection in ensembles methods. This algorithm allows, for each test instance, the selection of the best trees in the forest. This approach is tested on 10 databases from the UCI Machine Learning Repository. Results show that using a few best trees selected by our proposed pruning method, we can improve the performance of each dataset compared to classical ensembles methods and pruning techniques.

KEYWORDS : Random Forests, Bagging, Random Subspaces, Sub_RF, Trees selection, Dynamic Ensemble Pruning, Supervised Classification.

1. Introduction

The principle of ensemble methods (for example [6]) is to build a collection of predictors, and then aggregate all of their predictions. In classification, aggregation returns, for example, a majority vote among the classes provided by each individual predictor.

In this work, tree-based ensemble methods are used. They consist of a set of prediction trees ; each one being capable of producing a response when presented a sub-set of variables. For classification problems, the response takes the form of a class (label).

Using the sets of trees, a significant improvement in prediction compared with the conventional techniques (like CART) is believed to be obtained. Response of each tree depends on the subset of independently selected variables. One of the most used tree-based ensemble methods called RF (Random Forest)[4].

Despite the efficiency of the random forests, several researchers have tried to improve the accuracy using only the best trees of the forest. This improved method is called Trees Selection or Pruning. There are two kinds of Pruning : Static Pruning where a subset of trees is selected once for the whole test set, and Dynamic Pruning where the selection is made for each test sample individually at prediction time.

In this paper, the main interest is therefore to study the ability of tree selection on a modified version of random forests (called Sub_RF) by selecting the best ensemble of trees. Our new proposed method for tree selection attempts at improving accuracy. For that, this work has been framed as follows : in section 2, methods that we use in our algorithm are introduced. After that, related works to the method we made for ensemble pruning is discussed. Then, our results obtained on some benchmarks from the UCI Machine Learning Repository are exposed. At last, a general summary is given that highlights the main properties of our technique.

2. Methods

2.1. Random Forest

In random forests, Breiman proposes to use the Bagging [5], but for each data set generated, the growth of the tree is processed with a random selection of explanatory variables at each node. The word Bagging is a contraction of Bootstrap and Aggregating [1]. The idea of Bagging, is that by applying the basic rule on different bootstrap samples, we modify the predictions, and so we eventually build a collection of various predictors. The aggregation step then allows to obtain a powerful predictor.

The Random Forests algorithm - Random Input (Forest-RI) [4] is one of the most popular achievements of research devoted to the aggregation of randomized trees. Synthesizing the approaches developed respectively by [5] and [1], it generates a set of trees doubly disrupted using a randomization operating both at the training sample and internal partitions. Each tree is thus generated at first from a subsample (a bootstrap sample) of the complete training set, similar to the techniques of bagging. Then the tree is constructed using the CART methodology with the difference that at each node the selection of the best split based on the Gini index is performed not on the complete set of attributes M but on a randomly selected subset of it. During the prediction phase, the individual to be

1. A bootstrap sample L is obtained by randomly drawing n observations with replacement from the training sample L_n, each observation has probability $1/n$ to be pulled.

classified is spread in every tree of the forest and labelled according to the CART rules. The whole forest prediction is provided by a simple majority vote of the class assignments of individual trees.

In addition to building a predictor, the algorithm of Random Forests-RI calculates an estimate of its generalization error : the error Out-Of-Bag (OOB). This error was already calculated by the Bagging algorithm ; hence, the presence of "Bag". The calculation procedure of this error is as follows : From a training set "A" of "X" examples , bootstraps samples are generated by drawing "X" samples with replacement from "A". In average, for each bootstrap sample 63.2% are unique examples of "A", the rest being duplicates. So for each sub base, 1/3 samples of "A" are not selected and are called OOB samples. They will be used in internal evaluation of the forest (estimated error classification generalization of forest) or as a measure to calculate the variable of importance to use it in variable selection.

2.2. Subspaces Random Forest

In this method, the creation of a set of classifiers is made by using the method SubBag [17] for the generation of training samples. The classifiers are decision trees generated by using the Forest-RI algorithm [4].This algorithm of tree ensemble creation is called Sub_RF (Subspaces Random Forest) [7].

Algorithm 1 Pseudo code of the Sub_RF algorithm (LearnSubRF)

Input : The Training set L, Number of Random Trees N, SubSpace size S.
Output : $TreesEnsemble$
Process :
for $i = 1 \rightarrow N$ **do**
$T^i \leftarrow BootstrapSample(T)$
$T^i \leftarrow SelectRandomSubSpaces(T^i, S)$
$C^i \leftarrow ConstructRF\_tree(T^i)$
$E \leftarrow E \cup \{C^i\}$
end for
$ReturnE$

The function $ConstructRF\_tree()$ allows to create trees using the principle of random forests.

3. Related works

Ensemble selection algorithms (also called pruning algorithms) aim at finding the best subset, among the set of all hypotheses space, which may optimize the computation time (as in static Pruning) and / or improve performances (dynamic pruning). The main aim of this experimental work is to fundamentally apply ensemble selection methods for selecting best classifiers from a random forest which is generated using the method SubBag. There exist several studies in the literature that we discuss below according to their types (static or dynamic).

3.1. Static Pruning

Static pruning consists in creating a set of classifiers (random forest or other) and then selecting a part of this set (the best classifiers) that performs as well as, or better than, the original ensemble. The selected set will be used for the classification of test instances. Many researchers have shown in their studies on the tree selection in a random forest, that better subsets of decision trees can be obtained by using sub-optimal methods of classifier selection [29] [20] [26] [15] [3]. Their results affirm that an induction algorithm of classical random forests is not the best approach to produce well performing tree-based classifiers.

Among the most recent works, in this regard, we find the article of Zhao et al. [27] where the authors propose a fast pruning method compared with the existing methods. Their idea is to create a prediction table where each row of the table contains a database instance and each column a classifier. The proposed algorithm chose the best combination of classifiers that minimizes the error.

[13] in their article, propose a heuristic that respects the compromise accuracy / diversity for the evaluation of the contribution of each classifier and thus, choose the best subset. Their results show that the subset chosen by their algorithm EPIC (for Ensemble pruning via indivdual contribution ordering) outperforms the original set.
Other studies present classifiers selection as an optimization problem where we had to look for the best solution in the space. Most of the proposed algorithms have used optimization algorithms such as greedy search [8] [16] [18], hill climbing [25] or even genetic algorithms [28].

In [11], the authors have presented an entropy-inspired ordering ensemble pruning algorithm exploiting an alternative definition of the margin of ensemble methods. This pruning strategy considers the smallest margin instances as the most significant in building reliable classifiers. The algorithm combines best classifiers, which classify correctly smallest margin, for future decisions.

3.2. Dynamic Pruning

Dynamic pruning (also called dynamic ensemble selection or instance-based ensemble selection) aims at selecting the best subset of classifiers dynamically (ie : for each test example) from the original set. The selected classifiers are aggregated afterwards by a majority vote. The subset should lead to a greater accuracy compared to the whole set. This type of selection is best suited for offline problems where we privilege accuracy over computation time because there is an additional cost in the testing phase.

[24] and [10] are said to be among the first authors who were interested in dynamic selection. Their methods consist in using for each instance of the test base, the best classifiers of its neighborhood (using KNN). Authors propose two methods to calculate the performance of classifiers. The first is OLA (Overall local Accuracy) ; this metric calculates the rate of correct classifications of each classifier on instances of the neighborhood. The second metric is called LCA (Local Class Accuracy), it allows to calculate, for each classifier, the rate of correct classification of examples in the neighborhood that have the same given class for the test instance. Best Classifiers are combined to classify this instance.

Two other approaches, dynamic selection (DS) and dynamic voting (DV) have been proposed by [19]. DS uses the same principle as OLA [24] but by weighting selected classifiers by their distance. DV does not use KNN but rather all the classifiers weighted by their local competence. An approach between DS and DV was introduced by [21]

where the author proposed to select the 50% best classifiers and then combining them using DV.

Among the most recent works, one may find that of [12]. The authors proposed four different versions of a method called KNORA (K-nearest Oracle). The proposed algorithms use the KNN to select neighbors of each test instance.

[14] modelled the pruning as a multi-label problem called IBEP-MLC (Instance-Based Ensemble Pruning via Multi-label Classification). The idea proposed by the authors is to add, for each instance of the training set, a label with each classifier. If the instance is well classified, a positive label is given (+), otherwise it is a negative one (-). The classification of a new instance is made by taking the classifiers with a positive label in its neighborhood.

In [23] authors developed a probabilistic model method for calculating the classifier competence. The competences calculated for a validation set are generalized to an entire feature space by constructing a competence function based on a potential function model or regression. Three systems based on a dynamic classifier selection and dynamic ensemble selections (DES) were constructed using the method developed.

In [9], they have proposed a dynamic classifier selection strategy for One-vs-One scheme that tries to avoid the non-competent classifiers when their output is probably not of interest. This method considers the neighborhood of each instance to decide which classifier may correctly classify this instance.

4. Proposed Method

It has been noticed that all the works previously cited, in the section dynamic pruning, are based on KNN for the choice of the neighborhood, which is an additional parameter to adjust. Noting that this method is not effective if we do not use all the space of attributes (case of RSM or SubBag). Indeed, two instances may be far in the complete space and close in a part of it.

As a solution to this problem, a method based on a different notion of neighborhood is suggested. In this work, the nodes of the trees are used as a heuristic neighborhood. Indeed, two instances are adjacent if they pass through the same nodes in a given tree. Our algorithm involves three steps :

– Generation of a random tree-based ensemble using Sub_RF method [7].

– For each tree in the forest, the classification of its OOB elements (with this tree) is launched and their paths are saved (step (1) in the Algorithm 4).

– To classify a new instance, the score of each tree for this instance should be calculated and process to a majority vote among the K-best trees. The score of the tree is calculated based on the correct classification of its OOB weighted by their distance with this instance (step (2) in the Algorithm 4.).

For a test instance, the score of a tree, is a value comprised between 0 and 1. A score equal to "1" means that the tree is very efficient and will ensure a correct classification for this test instance. A tree with a score equal to "0", has a high chance to give a false classification for the instance.

The principle of calculating the score of a tree, for an instance, is very simple. It is based on a Boolean function which weights the distance between the test instance and each OOB of this tree. This function returns "1" if the element OOB was well classified by the tree, otherwise "0".

A distance between a test instance and an OOB equal to "1" means they have gone together through all the nodes of the tree. A distance very close to zero means that the two elements have gone through different paths.

The notion of neighborhood based paths was introduced by Vens and Costa in [22]. It is about calculating communes nodes between an OOB and a given instance considering all the paths and not only leafs. The distance of an OOB compared to an instance is a fraction of the number of nodes traversed together over the maximum depth between this two paths.

5. Results and interpretations

To test our algorithm, ten databases from the UCI Machine Learning Repository [2] were used. Databases which have been used in our experiments are described in the Table 1.

Our experiments are to implement seven different ensembles : Sub_RF, Sub_RF with Static Pruning, Sub_RF with Dynamic Pruning, Sub_RF with OEP, Bagging with OEP, Randomized trees with OEP and RF with OEP. The goal is to visualize and study the evolution of the error rate of each method and subsets obtained during the process of tree selection.

First, each database has been divided into two sub-data sets, one for learning and the other for test (using 5-fold cross validation). The separation of the data was carried out by random draw from the whole set.

Databases	Inst	Features	Cl
Breast	699	9	2
Ecoli	366	7	8
Habermann	306	3	2
Isolet	7797	617	26
Liver	345	6	2
Pendigits	10992	16	10
Pima	768	8	2
Segmentation	2310	19	7
Vehicle	846	18	4
Yeast	1484	8	10

Tableau 1. *Used databases*

As it has been already explained, our method uses bootstrapping to generate the bag. OOB will be used for selecting classifiers. Several works in the literature bulk have shown that a number of attributes equal to $\sqrt{M}$ is a good compromise to produce an efficient forest [4] [3].

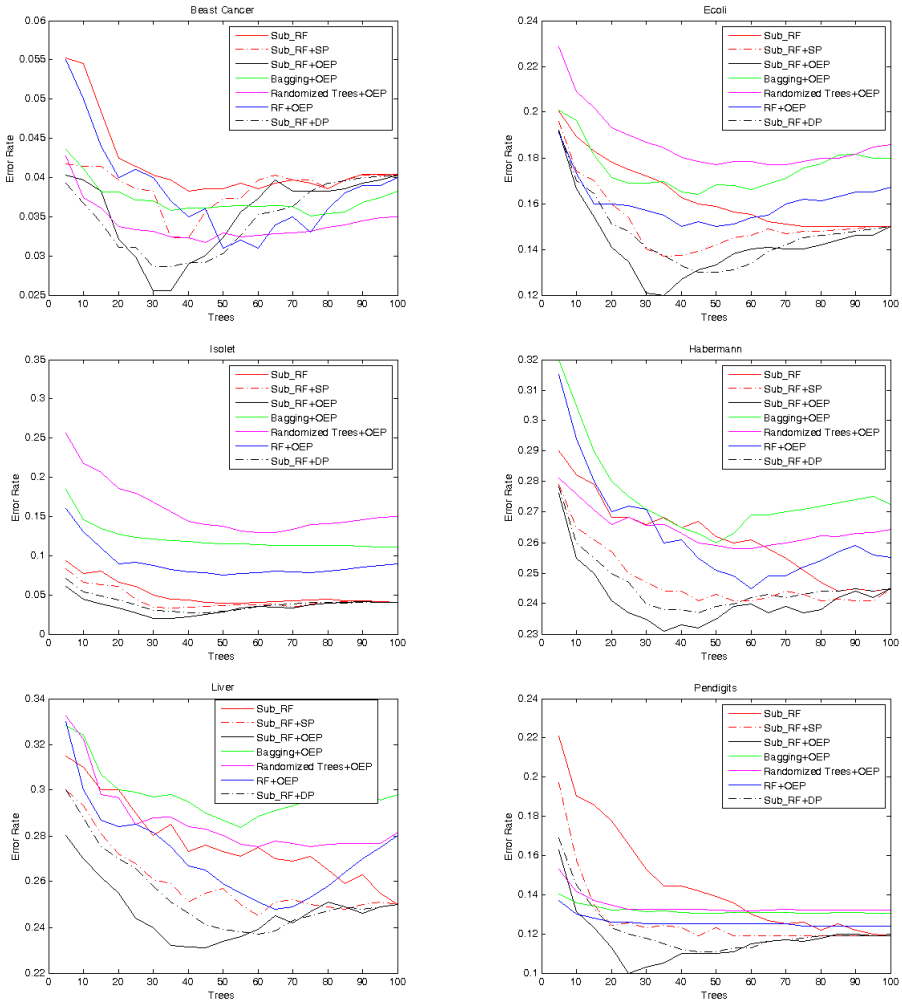
In this experiments, a comparison of our proposed dynamic pruning method OEP (for Out of bag-based Ensemble Pruning), Static Pruning (SP) and Dynamic Pruning (DP) applied on Random Trees (which uses only one random feature), Random forests (RF), Bagging and Sub_RF was established. Groups of selections were organized to which, each time, five trees to the group where added. In the first experiment, a random tree selection for Sub_RF, where trees are selected and aggregated according to their order of appea-

rance and without condition, was processed. For the Static pruning, the OOB database is used like a validation database and the performance of each tree is calculated based on the correct classification rate of its OOB. At each stage, the K-best trees are selected for the classification of the test set. OEP Algorithm is used with all cited methods and compared with the Dynamic Pruning algorithm based on KNN used with Sub_RF (Sub_RF+DP in the figures).

Fig.1 show error rates of different combinations as the number of selected trees increases. It may be observed that our algorithm of dynamic pruning OEP gives best result between 20 and 50 trees for all databases. The best results are obtained with the forest generated by the Sub-RF algorithm. This can be explained by the fact that, unlike Bagging and RF, the Sub-RF trees are very different since they do not use all attributes and, unlike the Random Trees, they choose the best variable. Sub_RF thus provides overall the best tradeoff in terms of randomization in the context of our dynamic pruning algorithm. OEP seems to gives better results than the static pruning and dynamic pruning methods that use KNN : it leads globally a lower error rate than all methods and it also reaches its optimum for a smaller set of trees. Therefore, the neighborhood based on tree nodes is more efficient if we do not use the whole attribute space.

6. Conclusion

To put it in a nutshell ; in this paper, a new instance-based ensemble pruning method which uses the neighborhood in the tree has been essentially hypothesized. This method has, in fact, proven effective on trees that do not use all the attribute space. For this, it sounds quite important to investigate the efficiency of a method of generating tree which is very similar to SubBag (Sub_RF) and gives better results compared to conventional random forests. For that reason, our approach on ten UCI databases was experimentally tested. Results display that our suggested approach is almost competitive with pruning methods (static and dynamic) which are based on KNN.


Beast Cancer
Ecoli
Isolet
Habermann
Liver
Pendigits
Sub_RF
Sub_RF+SP
Sub_RF+OEP
Bagging+OEP
Randomized Trees+OEP
RF+OEP
Sub_RF+DP
Error Rate
Trees

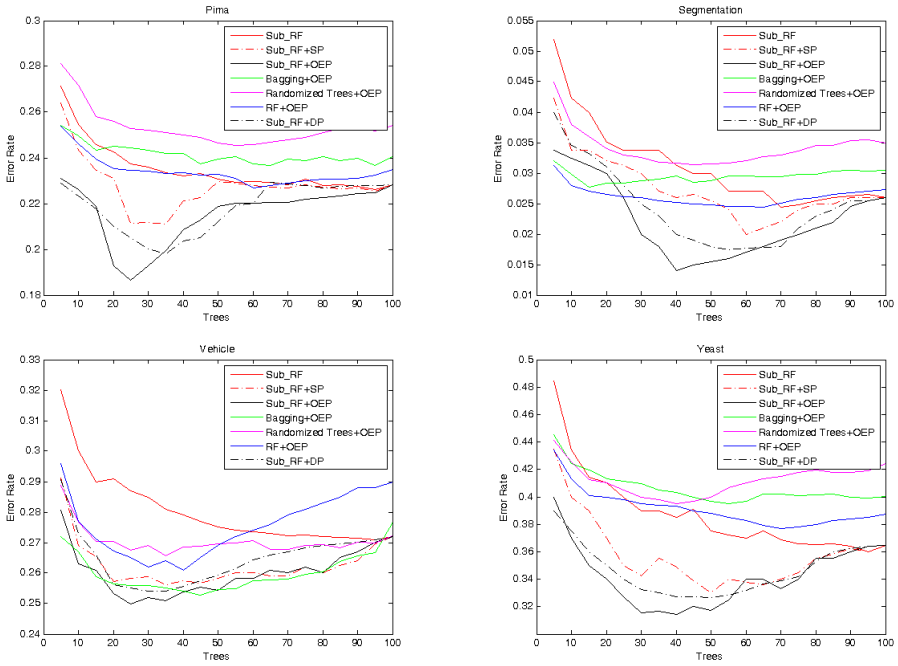


Figure 1. *Error rates of different algorithms*

7. Bibliographie

Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7) :1545–1588, 1997.

K. Bache and M. Lichman. UCI machine learning repository, 2013.

Simon Bernard, Laurent Heutte, and Sebastien Adam. Influence of hyperparameters on random forest accuracy. In *Multiple Classifier Systems*, volume 5519 of *Lecture Notes in Computer Science*, pages 171–180. Springer, 2009.

L. Breiman. Random forests. *Machine Learning*, 45 :5–32, 2001.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2) :123–140, 1996.

T. Dietterich. Ensemble methods in machine learning. *Lecture Notes in Computer Science*, 1857 :1–15, 2000.

Mostafa EL HABIB DAHO and Mohammed El Amine CHIKH. Combining bootstrapping samples, random subspaces and random forests to build classifiers. *Journal of Medical Imaging and Health Informatics*, 5(3) :539–544, 2015.

Wei Fan, Haixun Wang, Philip S. Yu, and Sheng Ma. Is random model better ? on its accuracy and efficiency. In *Third IEEE Int. Conf. on Data Mining*, ICDM '03, pages 51–, Washington, DC, USA, 2003. IEEE Computer Society.

Mikel Galar, Alberto Fernández, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. Dynamic classifier selection for one-vs-one strategy : Avoiding non-competent classifiers. *Pattern Recognition*, 46(12) :3412–3424, December 2013.

Giorgio Giacinto and Fabio Roli. Adaptive selection of image classifiers. In *Image Analysis and Processing*, volume 1310, pages 38–45. Springer Berlin Heidelberg, 1997.

Li Guo and Samia Boukir. Margin-based ordered aggregation for ensemble pruning. *Pattern Recognition Letters*, 34(6) :603–609, April 2013.

Albert H.R. Ko, Robert Sabourin, Alceu Souza Britto, and Jr. From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognition*, 41(5) :1718 – 1731, 2008.

Zhenyu Lu, Xindong Wu, Xingquan Zhu, and Josh Bongard. Ensemble pruning via individual contribution ordering. In *16th ACM SIGKDD Int. conf. on Knowledge discovery and data mining*, KDD '10, pages 871–880, New York, NY, USA, 2010. ACM.

F. Markatopoulou, G. Tsoumakas, and L. Vlahavas. Instance-based ensemble pruning via multi-label classification. In *22nd IEEE Int. Conf. on Tools with Artificial Intelligence*, volume 1, pages 401–408, 2010.

Gonzalo Martinez-Munoz, Alberto Suarez, Gonzalo Martínez-muñoz, and Alberto Suárez. Using boosting to prune bagging ensembles. *Pattern Recognition Letters*, 28 :156–165, 2007.

Gonzalo Martínez-muñoz and Alberto Suárez. Aggregation ordering in bagging. In *Int. Conf. on Artificial Intelligence and Applications*, pages 258–263. Acta Press, 2004.

Panče Panov and Sašo Džeroski. Combining bagging and random subspaces to create better ensembles. In *7th Int. Conf. on Intelligent data analysis*, IDA'07, pages 118–129, Berlin, Heidelberg, 2007. Springer-Verlag.

Ioannis Partalas, Grigorios Tsoumakas, and Ioannis Vlahavas. A study on greedy algorithms for ensemble pruning. Technical report, Aristotle University of Thessaloniki, 2012.

Seppo J. Puuronen, Vagan Terziyan, Artyom Katasonov, and Alexey Tsymbal. Dynamic integration of multiple data mining techniques in a knowledge discovery management system. *Data Mining and Knowledge Discovery : Theory, Tools, and Technology*, 3695 :128–139, 1999.

Grigorios Tsoumakas, Lefteris Angelis, and Ioannis Vlahavas. Selective fusion of heterogeneous classifiers. *Intelligent Data Analysis*, 9(6) :511–525, November 2005.

Alexey Tsymbal. Decision committee learning with dynamic integration of classifiers. In *East-European Conf. on Advances in Databases and Information Systems*, ADBIS-DASFAA '00, pages 265–278, London, UK, UK, 2000. Springer-Verlag.

Celine Vens and Fabrizio Costa. Random forest based feature induction. *IEEE Int. Conf. on Data Mining*, 0 :744–753, 2011.

Tomasz Woloszynski and Marek Kurzynski. A probabilistic model of classifier competence for dynamic ensemble selection. *Pattern Recognition*, 44(1011) :2656 – 2668, 2011.

Kevin Woods, W. Philip Kegelmeyer, Jr., and Kevin Bowyer. Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4) :405–410, April 1997.

Ying Yang, Kevin Korb, KaiMing Ting, and GeoffreyI. Webb. Ensemble selection for superparent-one-dependence estimators. In *Advances in Artificial Intelligence*, volume 3809, pages 102–112. 2005.

Yi Zhang, Samuel Burer, and W. Nick Street. Ensemble pruning via semi-definite programming. *Machine Learning Research*, 7 :1315–1338, December 2006.

Qiang-Li Zhao, Yan-Huang Jiang, and Ming Xu. A fast ensemble pruning algorithm based on pattern mining process. *Data Mining and Knowledge Discovery*, 19(2) :277–292, October 2009.

Z.-H. Zhou, W. Tang, Zhi hua Zhou, and Wei Tang. Selective ensemble of decision trees. In *Lecture Notes in Artificial Intelligence*, pages 476–483. Springer, 2003.

Zhi-Hua Zhou, Jianxin Wu, and Wei Tang. Ensembling neural networks : many could be better than all. *Artificial Intelligence*, 137(1-2) :239–263, May 2002.

Fast Polygons Fusion for Multi-Views Moving Object Detection from Overlapping Cameras

Mikaël A. Mousse[1,2], Cina Motamed[1] and Eugène C. Ezin[2]

[1]Laboratoire d'Informatique Signal et Image de la Côte d'Opale
Université du Littoral Côte d'Opale, France
E-mail : {mousse, motamed}@lisic.univ-littoral.fr
[2]Unité de Recherche en Informatique et Sciences Appliquées
Institut de Mathématiques et de Sciences Physiques, Bénin
E-mail : eugene.ezin@imsp-uac.org

RÉSUMÉ. Dans cet article, nous proposons un algorithme de fusion rapide de polygones pour la détection d'objets mobiles par le biais dans un réseau de caméras. Après la détection des pixels de premier plan de chaque camera, nous approximons les contours détectés par des polygones. Ces polygones sont projetées dans le plan de référence. Après cela nous proposons une approche de fusion efficiente dans le but d'obtenir une détection multi caméras. Les Différents résultats sur un jeu de données publique sont présentés et analysés. La détection des objets mobiles à travers la vue de chaque caméra est obtenue en utilisant un algorithme basé sur le codebook.

ABSTRACT. In this paper, we propose a fast polygons fusion algorithm to address the problem of moving object detection from overlapping cameras. Once the foreground pixels are detected in each view, we approximate their contours with polygons and project them into the reference plane. After this, we propose an efficient fusion approach to fuse polygons in order to obtain a multi-views foreground area. The different results on open video dataset are presented and analyzed. Each foreground information is obtained by using a codebook based moving object detection algorithm.

MOTS-CLÉS : Détection d'objet, Codebook, Caméras avec vues chevauchant, Fusion d'informations

KEYWORDS : Motion detection, Codebook, Overlapping camera, Information fusion

1. Introduction

In computer vision community, the use of multi-camera takes a lot of scope. Indeed, motivations are multiple and concern various domains as the monitoring and surveillance of significant protected sites, the control and estimation of flows (car parks, airports, ports, and motorways). Because of the fast evolution in the fields of data processing, communications and instrumentation, such applications become possible. These kind of systems require more cameras to cover overall field-of-view. They reduce the effects of objects dynamic occlusion and improve the accuracy of estimation of foreground zone.

According to Xu et al., existing multi-camera surveillance systems may be classified into three categories [6]. The system in the first category fuses low-level information. In this category, multi camera surveillance systems detect and/or track in a single camera view. They switch to another camera when the systems predict that the current camera will not have a good view. In the second category, system extracts features and/or even tracks targets in each individual camera. After this, we integrate all features and tracks in order to obtain the global estimates.These systems are of intermediate-level information fusion. The system in the third category fuses high-level information. In these systems, individual cameras don't extract features but provide foreground bitmap information to the fusion center. Detection and/or tracking are performed by a fusion center [1, 2, 3, 6, 7]. This paper will focus on the approaches in the third category. In this category some algorithms have been proposed. Authors in [2] proposed to use a planar homographic occupancy constraint to combine foreground likelihood images from different views in order to resolve occlusions and to determine regions on the ground plane that are occupied by people. In [3], authors extended the ground plane to a set of planes parallel to it, but at some heights off the ground plane to reduce false positives and missed detections. The foreground intensity bitmaps from each individual camera are warped to the reference image by authors in [1] and the set of scene planes are at the height of people heads. The head tops are detected by applying intensity correlation to aligned frames from the different cameras. This work is able to handle highly crowded scenes. Yang et al. detect objects by finding visual hulls of the binary foreground images from multiple cameras [7]. These methods fully utilize the visual cues from multiple cameras and are robust in coping with occlusion. However the pixel-wise homographic transformation at image level slows down the processing speed. In order to overcome this drawback, Xu et al. proposed an object detection approach via homography mapping of foreground polygons from multiple camera [6]. They approximate the contour of each foreground region with a polygon and only transmit and project the vertices of the polygons. The foreground regions are detected by using Gaussian mixture model. After the projection of the polygons vertices, they rebuilt each foreground map in the reference image by considering as foreground pixels all pixels lying inside a polygon. The multi-view object detection is obtained by considering the pixels which have been detected to be a foreground pixels in n different polygons (n is the number of cameras). This method provides good results [6]. In [5], authors also propose an algorithm based on polygons fusion for moving object extraction.

In this work, we propose a new strategy based on polygons which reduces the complexity of polygons fusion. Indeed a major challenge in computer vision is to get a real time system. Then it is important to reduce the complexity of each part of a computer vision system. This paper consists of four sections. The second section presents the po-

lygons fusion approach. The third section presents experimental results. Conclusion and future works are presented in section four.

2. Polygons Fusion Approach

In this section, we present our fusion approach for moving object detection in a multi camera system. The goal of our algorithm is to extract the relevant vertices of the polygons associated with the various objects in each view.

Let us consider a scene observed by n ($n \geq 2$) cameras with overlapping views. The multi-view moving object detection in the ground plane is the intersection of the single views foreground polygons projection. Then using our approach, we automatically detect the vertices of the polygons resulting from intersections. In our approach, the multi-view foreground map is represented by a codebook $\mathcal{D} = \{d_1, d_2, .., d_L\}$ and each codewords d_i, $i = 1, ..., L$ represents a polygons resulting from intersection. Each codewords consists of two vectors. The first vector $index_i$ contains the identifiers of the polygons which form the intersection whereas the second $content_i$ contains the vertices of the polygon resulting from intersection. In this part, we call vector a sequence container representing arrays that can change in size.

Firstly, we consider two camera views and we project the vertices of their polygons in the ground plane by using the principles of the homography. So if $\mathcal{V}_1 = (v_{11}, v_{21}, ..., v_{k1})$ is original polygon in the single view plan, the projected polygons becomes $\mathcal{V}_1 = (v'_{11}, v'_{21}, ..., v'_{k1})$ with $v'_{11}, v'_{21}, ..., v'_{k1}$ which are respectively the projection of $v_{11}, v_{21}, ..., v_{k1}$. Among the two views, we select one view and we seek its projected points which belong to a projected polygon from the view of the second camera. When we find a point which verify this condition, we compare it to the current codebook to determine which codeword d_m (if any) it matches (m is the matching codeword's index). To determine which codeword will be the best match, we create a vector from different polygons identifiers (belonging polygon identifier, origin polygon identifier) and compare it to the first vector of each codewords. Two vectors will be considered as equivalent if all the elements in one of the vector is necessarily in the second. If there is no match, we create a new codeword d_k by setting $index_k$ to the vector issues from different polygons identifiers (belonging polygon identifier, origin polygon identifier) and creating $content_k$ in which we insert the point. This part is resume in Algorithm 1. In this algorithm the two views V_1, V_2 are considered and we select the view V_2 in order to seek its projected points which belong to a projected polygon from the view of the second camera.

After that in each codeword d_i, we have $index_i$ which contains the information about polygons which form the intersection and $content_i$ which contains one point of the intersection. From this point we rebuild the intersection. For that we update the codeword d_i. We rewrite the projected polygon to which the point belongs by taking this vertex as the first point of the polygon. For example, if $\mathcal{V}_1 = (v'_{11}, v'_{21}, ..., v'_{k1})$ represents the projected polygon and v'_{31} is the point then the rewriting gives $\mathcal{V}_1 = (v'_{31}, v'_{41}, ..., v'_{k1}, v'_{11}, v'_{21})$. By using this polygon, we check from the first segment if in the ground plane a segment has

1. $inCod(index_n, \langle id_k, id_i \rangle)$ returns true if $index_n$ contains only id_k and id_i.

Algorithm 1: Codebook initialization

$L \leftarrow 0$ ($\leftarrow$ means assignment), $\mathcal{D} \leftarrow \emptyset$ (empty set)
for *each projected polygons* id_i *of the view* V_2 **do**
  for *each each vertex* $v_{j\_id_i}$ *of the polygon* id_i **do**
    if $v_{j\_id_i}$ *is inside the projected polygon* id_k *of the view* V_1 **then**
      Find the codeword d_m in $\mathcal{D} = \{d_n | 1 \leq n \leq L\}$ matching to $v_{j\_id_i}$ based on condition (a)
        (a) inCod [1]$(index_n, \langle id_k, id_i \rangle) = true$
    if $\mathcal{V} = \emptyset$ *or there is no match* **then**
      $L \leftarrow L + 1$
      create codeword d_L by setting parameter $index_L \leftarrow (id_k, id_i)$ and $content_L \leftarrow (v_{j\_id_i})$

an intersection with any segment of the second polygon of the codeword. If we don't find an intersection then we update codeword by adding at the end of the vector $content_i$ the point at the second end of said segment and the initial polygon is considered as default polygon during this part of the process. But if we find an intersection, we add two points at the end of the vector $content_i$: the first point is the intersection and the second is the point of the segment from the second polygon which belongs to the first polygon of codeword. In this case, the second polygon becomes default polygon. We repeat the search for intersection between segments from two different polygons by using the default polygon segment that comes from the last point which is inserted into $content_i$ until obtaining the first point of the codeword. We take care to avoid to include this point again. After realizing these instructions on each codeword, our codebook contains information about the polygons that form intersections using the two chosen views and the vertices of polygons representing these intersections. This part is resume in Algorithm 2.

For each of the remaining cameras (if any remain), we rebuild the codebook. In order to perform it, we consider the contents of the vector $content$ of each codeword of the immediately previous codebook as the vertices of a polygon and the concatenation of the contents of the vector $index$ of the codeword as the identifier of this polygon. All polygons from this codebook are considered as part of an imaginary camera view. And we use the process for codebook building for two different views (process which is explained in previous algorithms (Algorithm 2 and Algorithm 3)) to build the new codebook by using our imaginary camera view and the new input camera view.

Using this method we obtain automatically the vertices of the polygons resulting from intersection. The multi view moving objects detection are then obtained by set as foreground the pixels which are inside these polygons. The ray casting algorithm proposed by Sutherland et al. [8] has been used in order to resolve point-in-polygon problem.

1. default segment is the segment which is obtained by considering in default_polygon, default_point and the vertex that follows its.

Algorithm 2: Extraction of intersection vertices

for *each codeword* d_n *(with* $index_n = (id_k, id_i)$ *and* $content_n = (v_{j\_id_i})$*) of codebook* $\mathcal{D}$ **do**
  Rewrite the projected polygon id_i by taking $v_{j\_id_i}$ as the first point of the polygon.
  default_point $\leftarrow v_{j\_id_i}$, default_polygon_id $\leftarrow id_i$.
  repeat
    if *the default segment*[2] *has an intersection with an other segment from a second polygon forming* c_i **then**
      default_polygon_id $\leftarrow \overline{default\_polygon\_id}$ (identifier of the second polygon which forming c_i).
      intersect_point $\leftarrow$ intersection of the two segments.
      default_point $\leftarrow$ point of the segment from the default_polygon_id which belongs to $\overline{default\_polygon\_id}$.
      if *intersect_point* $\neq v_{j\_id_i}$ **then**
        update d_n by inserting intersect_point at the end of $content_n$.
      if *default_point* $\neq v_{j\_id_i}$ *and default_point* $\neq$ *intersect* **then**
        update d_n by inserting default_point at the end of $content_n$.
    else
      default_point $\leftarrow$ point at the second end of said segment.
      if *default_point* $\neq v_{j\_id_i}$ **then**
        update d_n by inserting default_point at the end of $content_n$.
  until *default_point* $= v_{j\_id_i}$

3. Experimental Results and Performance Evaluation

In this section, we present the performance of the proposed approach. Firstly we present the experimental environment and results. After that we present and analyze the performance of our system.

3.1. Experimental Results

For the validation of our algorithm, we tested it on video sequence that contains significant lighting variation, dynamic occlusion and scene activity. Both qualitative and quantitative evaluations have been carried out by using the PETS'2001 dataset[3]. We selected sequence "**Dataset 1**" which are also used in other researches works. The size of each frame is (768×576). The experiment environment is Intel® Core i7 CPU L 640 @ $2.13GHz \times 4$ processor with 4GB memory and the programming language is C++.

During our experiment, we use foreground pixels detection algorithm presented in Mousse et al. [4] for each single view foreground pixels extraction The foreground polygons is obtained by finding the convex hull of the foreground pixels. Each region can be approximated by a polygon. The polygon is obtained by finding the convex hull of all contours detected in threshold image. The convex hull or convex envelope of a set X of points in the Euclidean plane or Euclidean space is the smallest convex set that contains X. For instance, when X is a bounded subset of the plane, the convex hull may be vi-

3. Available online at `http://www.cvg.reading.ac.uk/PETS2001/pets2001-dataset.html`

sualized as the shape enclosed by a rubber band stretched around X. Some segmentation results are presented in Figure 1.

3.2. Performance Evaluation and Discussion

Xu et al. demonstrated the efficiency of the use of single views polygons and of their intersections in a ground plane for multi-view objects detection [6]. Our experiment results also confirm that the polygon projection results are very close to those from the bitmap projection. Due to this, we only evaluate the performance of our system by using the processing time as metric. Xu et al. proved that their algorithm faster than the existing algorithms [6]. So the discussion about the processing time of our proposed algorithm is done by comparing its with the processing time of Xu et al.'s algorithm. Then, the overall

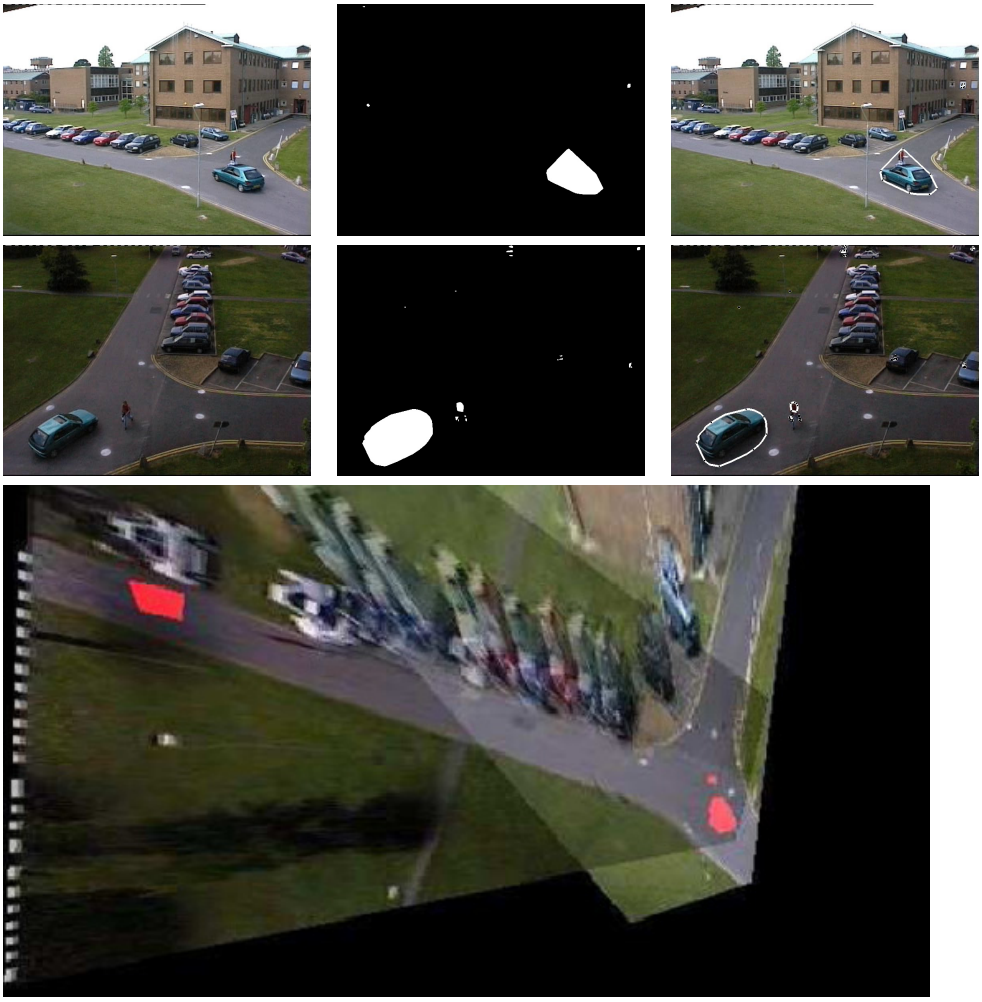


Figure 1 : The first two rows show each camera views. In these rows, the first column presents the original frame, the second column shows the foreground maps in single view and the third column presents a foreground approximation with polygons The third row shows the segmentation result using a multi-view informations.

Tableau 1 : Global performance evaluation.

Score	Xu et al Algorithm [6]	Proposed algorithm
Processing times (f/s)	65.82	**73.97**

execution time of the two algorithms. It is expressed in frames per second. Regarding the comparison of overall performance, the obtained values are reported in Table 1. According to these values we can conclude that our proposed algorithm is faster than algorithm suggested by Xu et al. The difference between the two execution times will increase when the number of cameras will increase and/or the number of objects observed by several cameras will become much larger. In fact with more cameras and/or more objects we will obtain more polygons. The complexity of the fusion process strongly depends on the number of cameras and/or the number of foreground objects.

4. Conclusion

In this work, we have proposed a fast algorithm for object detection by using overlapping cameras. In each camera, we use an improvement of codebook based algorithm to get foreground pixels. The single moving object detection algorithm integrates superpixels segmentation in original codebook and extends its on pixel level. In order to obtain the multi-view moving object detection, we propose a fusion approach which enables to determine quickly the polygons resulting from intersection of single views polygons. The experiment results have shown that the use of our fusion method reduces the computational complexity of multi-view moving object detection.

5. Bibliographie

Eshel, R., Moses, Y. : Homography based multiple camera detection and tracking of people in a dense crowd. *18th IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

Khan, S.M., Shah, M. : A multi-view approach to tracking people in crowded scenes using a planar homography constraint. *9th European Conference on Computer Vision*, 2006.

Khan, S.M., Shah, M. : Tracking multiple occluding people by localizing on multiple scene planes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31*, 2009.

Mousse, M.A. , Ezin, E.C. , Motamed, C. : Foreground-background segmentation based on codebook and edge detector. *10th International Conference on Signal Image Technology & Internet Based Systems*, 2014.

Mousse, M.A., Motamed, C., Ezin, E.C. : Fast moving object detection from overlapping cameras. *International Conference on Informatics in Control, Automation and Robotics*, 2015.

Xu, M., Ren, J., Chen, D., Smith, J., Wang, G. : Real-time detection via homography mapping of foreground polygons from multiple cameras. 18th IEEE International Conference on Image Processing, 2011.

Yang, D.B., Gonzalez-Banos, H.H., Guibas, L.J. : Counting people in crowds with a real-time network of simple image sensors. *9th IEEE International Conference on Computer Vision*, 2003.

Sutherland, I.E., Sproull, R.F., Schumacker, R. A. : A characterization of ten hidden surface algorithms. *ACM Computing Surveys (CSUR)*, 1974.

A multi-agent model based on Tabu Search for the permutation flow shop problem minimizing total flowtime

Soumaya Ben Arfa* — Olfa Belkahla Driss**

* Stratégies d'Optimisation et Informatique intelligentE (SOIE) High Institute of Management University of Tunis 41, Street of Liberty Bouchoucha-City CP-2000-Bardo Tunis, Higher Business School of Tunis, University of Manouba
Tunisia
benarfa.soumaya@gmail.com

** Stratégies d'Optimisation et Informatique intelligentE (SOIE) High Institute of Management University of Tunis 41, Street of Liberty Bouchoucha-City CP-2000-Bardo Tunis, Higher Business School of Tunis, University of Manouba
Tunisia
olfa.belkahla@isg.rnu.tn

ABSTRACT. In this paper, we treat the permutation flowshop scheduling problem with total flowtime minimization. We have propose a multi-agent model using tabu search method for solving this type of problem. Our proposed model MA.TF.PFS « Multi-Agent model to minimize Total Flowtime in Permutation Flow Shop » is composed by two classes of agents which are the supervisor agent and n job agents. The supervisor agent generates an initial solution and each job agent has a key role, it is a scheduler looking for a neighbor solution to improve the current solution by tabu search metaheuristic. Computational results show that the MA.TF.PFS is performant and it is significantly better than the BES (LR) method and three of other metaheuristics.

RÉSUMÉ. Dans cet article, nous traitons le problème d'ordonnancement d'atelier de type flow shop de permutation avec la minimisation de temps d'écoulement total. Nous proposons un modèle Multi-Agents en utilisant la méthode de recherche tabou pour résoudre ce type de problème. Notre modèle proposé MA.TF.PFS « Multi-Agent model to minimize Total Flowtime in Permutation Flow Shop » est composé de deux classes d'agents : Un agent superviseur et n agents jobs. L'agent superviseur génère une solution initiale et l'agent job a un rôle primordial, c'est un ordonnanceur qui cherche une solution voisine pour améliorer la solution courante en utilisant la métaheuristique recherche tabou. Les résultats obtenus montrent que MA.TF.PFS est performant et il est nettement meilleur que la méthode BES (LR) et trois autres métaheuristiques.

KEYWORDS : Multi-agent systems, Scheduling, Permutation flow shop, Total flowtime, Tabu search.

MOTS-CLÉS : Système multi-agents, Ordonnancement, Flow shop de permutation, Temps d'écoulement total, Recherche tabou

1. Introduction

The Permutation Flow shop Scheduling Problem (PFSP) is an important manufacturing system widely existing in industrial environments. So it can be described as follows: n different jobs are processed on m machines, where jobs on each machine follows the same order. The makespan or the minimization of total completion time, is considered to be the traditional criterion. Nowadays, the minimization of total flowtime has become an interesting topic in the scheduling literature. The PFSP with total flowtime criterion has proved to be NP-complete [6], even with two machines. However, so far no method seems to be the best for total flowtime minimization, including mathematical methods [1] [7], many heuristics and metaheuristics have been proposed. Most researches [5] [11][9][10] have been devoted to developing heuristic algorithms to obtain good solutions. Liu and Reeves [8] proposed an effective method LR(x) to generate the initial solution for their composite heuristics, by which new best solutions were found for nearly all 120 benchmark instances [13]. At the same time, many heuristics have been [5] integrated NEH insertion method as well as the pairwise exchange strategy in their algorithm. Indeed, we are looking for faster solutions leading to the development of several metaheuristics. Rajendran and Ziegler proposed two ant colony algorithms called M-MMAS and PACO [13]. The Particle Swarm Optimization algorithm called PSOvns where a SPV (the smallest position value) rule and VNS (variable neighborhood search) local search were applied proposed by [14]. Some of the most recent are the artificial bee colony algorithm and a discrete differential evolution algorithm illustrated by [15]. Dong et al. [3] proposed a Multi-Restart Iterated Local Search algorithm called MRSILS. Nowadays, they showed that the works are clearly superior to the heuristics addressed by Liu and Reeves unless for 100 benchmark instances by Taillard. All works that have been done for solving this type of problem are centralized, but for the minimization of Makespan, [2] used multi-agent systems proving best results. Based on these results, we suggest a model based on multiagent paradigm. The remainder of this paper is structured as follows: in section 2, we briefly describe the formulation of PFSP with total flowtime minimisation. We describe in details the Multi-Agent model based on Tabu Search in section 3. Section 4 contains the adaptation of the different elements of the Tabu Search. In section 5, we propose the dynamic of MA.TF.PFS. In section 6, experimental results are proposed. Finally, section 7 concludes the paper and suggests some future studies.

2. Problem Formulation

The PFSP with minimizing total flow time can be formally defined as follows: A set of jobs N= 1, 2 . . ., n available at time zero must be processed on m machines, where n $\geq$ 1 and m$\geq$1. The processing time for job i on the machine j is noted by $p_{i,j}$. C_i denotes the completion time of job i, where the completion time for job i on the machine j noted by $C(i,j)$ whether π (π_1,π_2,...π_n) a permutation, which represents the completion time of job π_i on the machines j. It Can be calculated as follows:

$C(\pi_1, 1) = p_{\pi_1,1}$,

$C(\pi_i, 1) = C(\pi_1, 1) + p_{\pi_1,1}$ for i=2, ..., n,

$C(\pi_1, j) = C(\pi_1, j-1) + p_{\pi_1,j}$ for j=2, ..., n,

$C(\pi_i, j) = \max C(\pi_{i-1}, j), C(\pi_1, j-1) + p_{\pi_1,1}$,

For $i = 2, ..., n; j = 2, ...m$.

Since ready times are zero, the flow time $C(\pi_i)$ is equivalent to the completion time $C(\pi_i, m)$. As a result, The PFSP with the total flow time is to find a π^* permutation throughout all Π permutations so that:

$$\sum_{i=1}^{n} C(\pi_i^*, m) \leq \sum_{i=1}^{n} C(\pi_i, m), \qquad \forall \pi \in \Pi \qquad [1]$$

3. The Multi-Agent model based on Tabu Search for PFSP

The different solving approaches that exist in the literature are all centralized architectures. They are sometimes ineffective given the difficulty of the problem. That's why we turned to the solving distribution by the use of Multi-Agent Systems [4]. So this type of system offers some parallel architectures that save computation time when solving difficult or large problems. We present in this section our multi-agent model named Multi-Agent model to minimize Total Flow in Permutation Flow Shop (MA.TF.PFS) as illustrated in Figure 1. The model consists of two types of agents: one Supervisor agent and n Job Agents where n is the number of jobs. Each agent in our model has its own static and dynamic knowledge and its own behavior. This behavior depends on its state, it can be: satisfied, unsatisfied or gives priority to the processing of messages. In addition, each agent has some acquaintances, the agents knows with which to communicate. In the remainder of this section, we show the different types of agents.

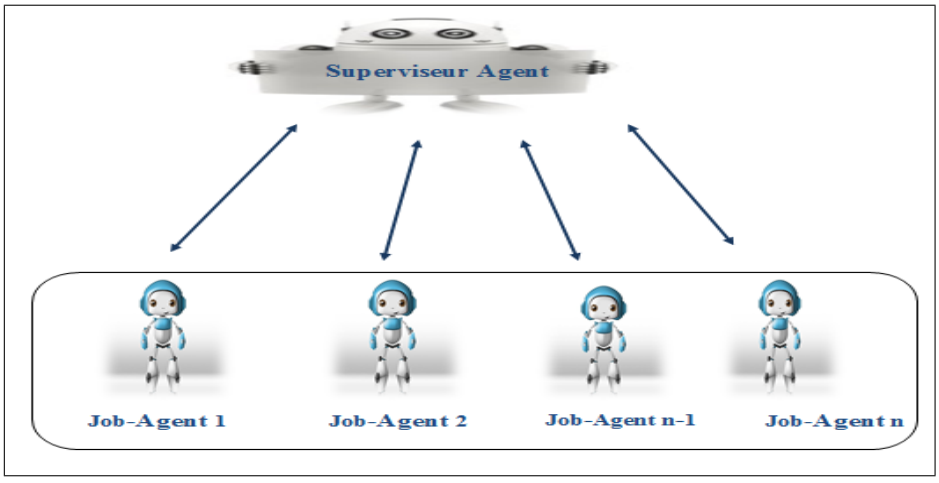


Figure 1. *MA.TF.PFS model*

3.1. The Supervisor Agent

The Supervisor Agent contains the core of the tabu search algorithm. It aims to launch the program, generating an initial solution and create as many agents as jobs named Job-Agent. In our approach, the Supervisor Agent is a regulator, it can communicate with all the job agents with the overall goal of finding an optimal schedule minimizing the total

flow time. As the number of iterations has reached the maximum number, the Supervisor Agent is not satisfied. Otherwise, it provides the best solution to the user. The Supervisor Agent has as acquaintances all agents in the system. Its static knowledge includes:

– The system operations to be performed and their respective durations on different machines;

– The size of the tabu list ;

– The maximal number of iterations allowed;

– The initial solution S0 from which the optimization process begin.

Its dynamic knowledge consists of:

– The tabu list elements;

– The current solution and the total flow time ;

– The neighbours and their solution ;

– The number of iterations ;

– The best solution found by the tabu search until the current iteration and its total flowtime.

3.2. The Job Agent

In our model, we have n Job Agents (n is the number of jobs of the flow shop problem). Each Job-Agent has an important role, it is a scheduler that is looking for a neighbor solution in order to improve the current solution. Each agent has its own dynamic and static knowledge and his own behavior. This behavior depends on its state: satisfied or unsatisfied and gives priority to the processing of messages. Furthermore, each agent has acquaintances. Its static knowledge includes:

– The diversification requirement ;

– The execution times of each job on all machines;

Its dynamic knowledge consists of:

– The common solution sent by the Supervisor Agent;

– The tabu list elements;

– The best solution met for each Job-Agent.

4. Global dynamic

In our model MA.TF.PFS, the global optimization process is carried out by collaboration between the Supervisor-Agent and the Job-Agents. The Supervisor-Agent knows the problem to solve. So, it generates an initial solution and then tries to improve it with applying the Tabu search method. Once the initial solution is determined, it is considered as a common solution. The Supervisor-Agent sends a message to the n Job-Agents each of which contains the current solution and the total flow time. In a parallel manner, the Job-Agents look for other neighbor solutions using a smart search in the diversification phase to get rid the local optimum. After the total flow time calculations of all neighbor solutions, they will be sent to the Supervisor-Agent. It chooses the best non-tabu neighbor with a minimum Total Flow Time to start a new iteration and inserts its total flow time in the tabu list. The above process continues until the stopping rule is satisfied. At this point, the Supervisor-Agent kills all the Job-Agents and displays to the user the best scheduling

found and its total flow time. We can see in Algorithm 1 the used tabu search algorithm.

Algorithm 1 The used tabu search algorithm

```
List-tabu← ∅
Nbr-iter ← 0
Current-schedule← Receive-initial-schedule(initial-schedule, supervisor-agent)
Best-schedule ← Current-schedule
Best-TFT ← Courrent-Total-Flowtime
while (Nbr-iteration≤ Nbr-iter-max) do /*the Nbr-iter-max can vary between 10-100*/
    Diversification
    List-tabu← add-in-List-Tabu( Best-TFT) /*the list tabu = 50 */
    Nbr-iter ++
end while
```

Despite the effectiveness of tabu search method in solving permutation flow shop scheduling problems, certain limitations have been detected. In fact, the main inconvenience is summed up in the absence of an effective diversification technique that encourages the search process to examine unvisited regions, as the best solutions at the local level are not necessarily good solutions globally. In our model, the Job-Agent is responsible for the diversification phase. Indeed, we implement a research method at Job-Agent level to get better neighbor solutions. Hence the research method is proposed in Algo 2. At each iteration diversification called iter-div, the Job-Agent moves its job to another position in the current solution and choose the best among them. Once the variable nbr-iter-div exceeds a predetermined number of iterations, called Threshold-Div, then the Job-agent sends the best solution 'Best-Sol' found and the Best Total Flow Time 'Best-TFT' to Supervisor-Agent.

Algorithm 2 The research method of Diversification

```
Nbr-iter-div = 0
List-tabu = ∅
while (Nbr-iter-div≤ Threshold-Div) and (current-TFT ≤ Best-TFT) do
    Current-position← Insert-moves (another-position);
    List-tabu ← Add(current-position);
    Nbr-iter-div ++
end while
Send (Best-Sol, Best-TFT, Supervisor-Agent)
```

5. Experimental results

In this section, MA.TF.PFS is compared with the best method (BES (LR) refers to the best performing heuristic as investigated by [8]); the two ant colony algorithms (M-MMAS and PACO) by [12]; the Particle Swarm Optimization algorithm with local search (PSOvns) by [15]; the a Discrete Artificial Bee Colony algorithm (DABC) by [14]; and the

Multi-Restart Iterated Local Search algorithm (MRSILS) by [3]. The proposed approach is implemented in the JADE platform and tested on a core i3 2.5 Mhz with 4GB RAM and we use the Taillard's instances [13]. We solve 110 problems in which the number of jobs between 20 and 200 and the number of resources varies between 5 and 20. Experimental results are presented in Table 1 by calculating the Relative Percentage Deviation (RPD) of the obtained results. RPD is calculated by the following equation:

$$RPD = \frac{Obt_{sol} - Best_{sol}}{Best_{sol}} * 100 \quad [2]$$

Hence Obt_{sol} is the solution yielded by a combination of factors for a given instance and the $Best_{sol}$ given by all combinations of factors for an instance. From table1, it can be concluded that the average performance of MA.TF.PFS is better than BES (LR), M-MMAS, PACO, and PSOvns. With respect to the rest of the methods, BES (LR) is outperformed by other algorithms. However, BES(LR) is rather simple and easily implemented compared to other algorithms. Therefore regarding the average performance, it seems that our model is effective in solving flow shop problems with the total flowtime criterion compared with the existing algorithms. According to the results presented in Table 1, we notice that MMAS, PACO, PSOvns, DABC and MRSILS approaches have not solved the problems of large size such as n=200. On the other hand, our approach is effective if the problem size increases. In table 2, we present only instances to that the best solution (bold values) are given by MA.TF.PFS. We also remark that the results obtained by MA.TF.PFS are better performing on 37.4 % of instances. So with n=200, we can see that the proposed model MA.TF.PFS provided the optimal solution for 11 instances out of 20.

Table 1. *Average relative percentage deviation over the best solutions*

instances	BES(LR)	M-MMAS	PACO	PSOvns	DABC	MRSILS	MA.TF.PFS
20x5	1.361	0.197	0.454	0.000	0.006	0.006	0.088
20x10	1.433	0.049	0.323	0.002	0.000	0.000	0.329
20x20	1.019	0.118	0.732	2.260	0.000	0.000	0.284
50x5	1.835	1.413	1.227	0.526	0.162	0.031	1.025
50x10	2.906	1.908	1.644	0.666	0.050	0.083	1.572
50x20	2.709	1.600	1.289	2.155	0.019	0.158	0.863
100x5	1.067	0.918	1.136	0.310	0.198	0.005	0.369
100x10	2.156	1.746	1.402	0.689	0.245	0.005	1.742
100x20	3.263	1.991	1.733	1.612	0.156	0.046	1.725
average	1.972	1.104	1.104	0.913	0.142	0.029	0.887

Table 2. *Best solutions obtained by MA.TF.PFS on Taillard's benchmarks*

Problem	N/M	BES(LR)	M-MMAS	PACO	PSOvns	DABC	MRSILS	MA.TF.PFS
Ta002		15446	15151	15214	15151	15151	15151	**15151**
Ta003		13676	13416	13403	13301	13301	13301	**13301**
Ta004		15750	15486	15505	15447	15447	15447	**15447**
Ta005	20x5	13633	13529	13529	13529	13529	13529	**13529**
Ta008		13968	13968	14042	13948	13948	13948	**13948**
Ta009		14456	14317	14383	14295	14295	14295	**14295**
Ta010		13036	12968	13021	12943	12943	12943	**12943**
Ta011		21207	20980	20958	20911	20911	20911	**20911**
Ta013	20x10	20072	19833	19968	19833	19833	19833	**19833**
Ta017		18723	18376	18377	18363	18363	18363	**18363**
Ta019		20561	20330	20330	20330	20330	20330	**20330**
Ta022		31918	31604	31597	32659	31587	31587	**31587**
Ta027	20x20	33449	33038	32922	33733	32922	32922	**32922**
Ta028		32611	32444	32533	33008	32412	32412	**32412**
Ta029		33625	33623	34446	33600	33600	33600	**33600**
Ta033	50x5	64378	64166	64149	63577	63162	63241	**63162**
Ta038		65582	64863	65123	64638	64381	64578	**64381**
Ta056	50x20	124061	122369	122262	123217	120850	121083	**120850**
Ta057		126363	125609	125351	125586	123043	123084	**123043**
Ta059		125318	126582	123646	124932	121872	122111	**121872**
Ta066		235793	236225	236409	234082	234017	233651	**233651**
Ta068	100x5	235171	234813	234579	232755	232238	232167	**232167**
Ta069		251291	252384	253325	249959	249884	248999	**248999**
Ta079	100x10	312175	309664	305376	305605	304457	304026	**304026**
Ta091		1063976	-	-	-	-	-	**1062859**
Ta092		1049076	-	-	-	-	-	**1040604**
Ta094	200x10	1051335	-	-	-	-	-	**1048682**
Ta095		1055823	-	-	-	-	-	**1052832**
Ta097		1071471	-	-	-	-	-	**1052832**
Ta099		1045183	-	-	-	-	-	**1043902**
Ta100		1044888	-	-	-	-	-	**1038016**
Ta103		1297768	-	-	-	-	-	**1254529**
Ta105	200x20	1255708	-	-	-	-	-	**1236246**
Ta109		1259311	-	-	-	-	-	**1237428**
Ta110		1273354	-	-	-	-	-	**1253075**

6. Conclusion and future works

In this paper, we have proposed a multi-agent approach by using tabu search method to solve the permutation Flow Shop scheduling problem with minimizing total flow time. The model MA.TF.PFS (Multi-Agent model to minimize Total Flowtime in Permutation Flow Shop) provides good results and allows to solve large size problems. It is competitive with other successful methods. In the future works, we are planning to make some modifications in order to enhance the performance of our model. We can reinvest our work to study Flow shop using other optimization criteria. Another interesting work field would be to adapt our model for multi-objective scheduling problems.

7. References

[1] BANSAL.S. P, "Minimizing the sum of completion times of n jobs over m machines in a flowshop - A branch and bound approach ", *AIIE Transactions* , vol. 9, num. 306-311, 1977.

[2] BELKAHLA.DRISS.O, BARGAOUI.H, " Multi-Agent Model based on Tabu Search for the Permutation Flow Shop Scheduling Problem", *Advances in Distributed Computing And Artificial Intelligence Journal*, 2014.

[3] DONG. X, CHEN .P, HUANG.H , NOWAK.M, " A multi-restart iterated local search algorithm for the permutation flow shop problem minimizing total flow time", *Computers Operations Research*, 2013.

[4] FERBER.J, "Les systèmes multi-agents vers une intelligence collective", *InterEditions*, 1995.

[5] FRAMINAN.JM, LEISTEN.R, "An efficient constructive heuristic for flowtime minimization in permutation flow shops", *Omega*, 2003.

[6] GAREY.MR, JOHNSON.DS, SETHI.R "The complexity of flow shop and job shop scheduling", *Mathematics of Operations Research*, vol. 1 num. 117-29, 1976.

[7] IGNALL.E, SCHRAGE.L, "Application of the branch and bound technique to some flowshop scheduling problem", *Operations Research*, vol. 13 num. 400-412, 1965.

[8] LIU.JY, REEVES.CR, "Constructive and composite heuristic solutions to the $P//\sum Ci$ scheduling problem", *European Journal of Operational Research*, 2001.

[9] LI.XP, WU.C, "An efficient constructive heuristic for permutation flow shops to minimize total flow time ", *Chinese Journal of Electronics*, 2005.

[10] LAHA.D, SARIN. SC, "A heuristic to minimize total flowtime in permutation flowshop", *Omega*,2009.

[11] RAJENDRAN.C, ZIEGLER.H, "An efficient heuristic for scheduling in a flow shop to minimize total weighted flow time of jobs", *European Journal of Operational Research*, vol. 38 num. 103-129 ,1997.

[12] RAJENDRAN.C, ZIEGLER.C, "Ant-colony algorithms for permutation flowshop scheduling to minimize makespan/total flowtime of jobs", *European Journal of Operational Research*, 2004.

[13] TAILLARD.E, "Benchmarks for basic scheduling problems", *European Journal of Operational Research*, vol. 64 num. 78−285 , 1993.

[14] TASGETIREN.M, PAN.Q.K,, SUGANTHAN.P,, CHEN.AH.L "A discrete artificial bee colony algorithm for the total flowtime minimization in permutation flow shops", *Information Sciences*, 2011.

[15] TASGETIREN.M.F, LIANG.Y.C, SEVKLI.M, GENCYILMAZ.G "A particle swarm optimization algorithm for makespan and total flowtime minimization in the permutation flowshop sequencing problem", *European Journal of Operational Research*, vol. 177 num. 1930-1947,2007.

Formation de coalitions A-core : S-NRB

Pascal François FAYE[1,2], Mbaye SENE[1], Samir AKNINE[2]

[1] LID UCAD, SENEGAL, [2] LIRIS UCBL, FRANCE
pfmfaye@hotmail.com, ngalagne@yahoo.com, samir.aknine@univ-lyon1.fr

RÉSUMÉ. Nous proposons un mécanisme décentralisé de formation de coalitions stables dans un contexte de tâches stochastiques qui tient compte des incertitudes sur, les dépendances, les conflits de préférences et la disponibilité imprévisible des agents. Nous ne supposons aucune connaissance a priori sur les coalitions stables à former et qu'il n'est pas possible de calculer de façon centralisée ces coalitions avant le début de l'exécution des tâches en raison de ces incertitudes sur les agents et des contraintes de temps des tâches stochastiques. Nous appelons *S-NRB (Sequential Non-Return Broadcast)*, ce mécanisme qui permet de former des coalitions dites *A-core*, c'est-à-dire, *Core-stable* et Auto-stables. *S-NRB* combine les méthodes de la théorie des jeux avec les lois de probabilités pour atteindre les coalitions stables requises. L'analyse théorique et les expérimentations montrent que *S-NRB* surmonte dynamiquement les incertitudes des agents et des tâches et favorise l'établissement des coalitions *A-core*.

ABSTRACT. We focus on devising an efficient parallel and decentralized coalition formation mechanism dealing with uncertainties on agent's dependencies, on agent's conflicts, on agents' preferences and on agents' availability when stable agents' teamwork (coalitions) are required to achieve stochastic tasks. We assume no prior knowledge on stable coalitions to form and we consider it is not possible to compute in a centralized manner these coalitions before tasks' achievement due to agents' uncertainties and the time constraint of stochastic tasks. We propose a coalition formation mechanism called S-NRB (Sequential Non-Return Broadcast). The main propertY we emphasize is the A-Core (*Core-stable* and Auto-stables coalitions). So, S-NRB combines methods of game theory and the laws of probability to reach the required stable coalitions. The theoretical analysis and the experiments show how S-NRB dynamically overcomes uncertainties on agents and on tasks.

MOTS-CLÉS : Agents, Coalition, Stabilité

KEYWORDS : Agents, Coalition, Stability

1. Introduction

Les réseaux de Smart-phones, de robots et de capteurs qui améliorent la sécurité et les prises de décision des humains, sont de plus en plus étudiés. De nombreux mécanismes de formation de coalitions ont été étudiés dans le domaine des agents [2], cependant, un ensemble d'aspects intéressants qui sont sous-jacents à l'utilisation des réseaux ad-hoc pour le déploiement des agents, restent faiblement explorés [7] [8]. [1] propose une application utilisant les réseaux de capteurs pour la surveillance sans tenir compte de la possibilité d'indisponibilité de certains capteurs et lorsque les informations sont complètes. [9] considère la dynamique des tâches mais avec des agents homogènes et coopératifs. Leur méthode de formation de coalitions est basée sur les MDP mais sans tenir compte de la possibilité d'indisponibilité des agents. Il utilise des connaissances a priori sur les préférences des agents et sur la dynamique des tâches pour guider le choix des coalitions à former sans garantir leurs stabilités. Notre contribution principale est la proposition d'un mécanisme de formation de coalitions *core stables* qui est adaptatif, décentralisé et asynchrone. Plus en détail, nous proposons un mécanisme de formation de coalitions appelé *S-NRB* qui se fonde sur des négociations multilatérales et qui tient compte - des incertitudes sur les dépendances qui existent entre les agents, - des préférences des agents et de leur disponibilité afin que les coalitions requises soient stables pour l'exécution des tâches stochastiques. En distribuant le contrôle et les prises de décisions des agents, *S-NRB* garantit une convergence vers des coalitions *core stables* malgré les changements dynamiques du contexte des agents (tâches et voisinages). *S-NRB* associe : - les lois de probabilités pour prédire les changements sur les tâches et sur la disponibilité des agents, et - les méthodes de la théorie des jeux pour trouver l'ensemble des agents qui peuvent former des coalitions *core stables* en profitant des dépendances et de la structure en réseau des agents. Le contexte applicatif est un environnement de sinistre où les agents sont déployés dans des nœuds ad-hoc (e.g. PDAs, capteurs, Smart-phones) qui ont des disponibilités aléatoires, où il est impossible de prédire a priori les interdépendances des agents et les stratégies de collaboration à établir avant le début de l'exécution des tâches stochastiques. Nous considérons aussi que, le contexte est ouvert (un nœud ad-hoc peut aléatoirement quitter ou rejoindre l'environnement de déploiement). Les interdépendances peuvent provenir du fait que des intervenants de spécialités différentes (e.g. pompiers, organismes de secours, volontaires, etc.) qui hébergent des agents dans leur kit de secours, doivent prendre des décisions en fonction de leur groupe d'appartenance et des informations contextuelles. Il est clair que dans un tel contexte, ces composants ad-hoc intégrés dans les kits d'assistances ne peuvent pas avoir des ressources énergétiques et une puissance de calcul illimitée. De plus, ils doivent tenir compte de la portée limitée de leur signal et de leur indisponibilité temporaire ou définitive (e.g. pannes, dysfonctionnement électronique, etc.). Par conséquent, il n'est pas abusif de considérer que ces composants électroniques et ces intervenants n'ont pas de connaissances a priori sur les tâches à exécuter ou sur les coalitions (équipes) à former.

2. Concepts de notre modèle

Nous définissons une tâche stochastique comme un ensemble d'actions non-déterministes à exécuter pour résoudre un problème qui peut changer d'états d'un instant à l'autre, par exemple, les actions à effectuer pour éteindre un feu qui peut s'amplifier, rester

stationnaire ou régresser. Le but d'un agent est d'exécuter un ensemble d'actions non-déterministes pour maximiser son utilité. Formellement, soit A=$\{a_1, ..., a_n\}$ un ensemble des agents de l'environnement multi-agents où chaque $a_i \in A : i \in [1, |A|]$ est un agent égoïste déployé dans un unique composant ad-hoc. Par C, nous définissons une coalition telle que C=$\{A_c, G_c, T_c, V_c\}$, où $A_c \subset A$=$\{a_1, a_2, ..., a_k\}$ est un ensemble d'agents qui partagent la coalition C et qui ont comme ensemble de buts $G_c \subseteq \{G_{a_i} : i \in N, a_i \in A_c\}$. T_c définit une tâche stochastique à exécuter par la coalition C et V_c est la récompense associée à la coalition C après l'exécution de T_c. En raison du contexte incertain et dynamique considéré, les paramètres essentiels que nous prenons en compte pour un agent a_i s'expriment à un instant t sous la forme : $\{R_{a_i}, Aut^t_{a_i}, Hs_{a_i}, \vartheta^t_{a_i}, U_{a_i}, L^{Net}_{a_i}\}$. Pour l'agent a_i, l'information R_{a_i} décrit sa ressource et $Aut^t_{a_i}$ son autonomie énergétique à l'instant t. Hs_{a_i} représente l'*historique* de ses interactions avec les autres agents. $\vartheta^t_{a_i}$ est la vue de l'agent a_i à l'instant t. U_{a_i} est la fonction d'utilité de a_i et $L^{Net}_{a_i}$ définit, à un instant t, le niveau de dépendance de l'agent a_i par rapport aux autres agents de son voisinage $\vartheta^t_{a_i}$ dans une structure réseau notée Net.

Définition 1

La fonction d'utilité U_{a_i} d'un agent a_i s'exprime sous la forme U_{a_i}=$\sum_{c=1}^{\infty} u_c^{a_i}$ où $u_c^{a_i}$ est son utilité suite à sa participation à la coalition C (cf. équation 1) . $U^t_{a_i}$ est la valeur de la fonction d'utilité de l'agent a_i à l'instant t.

La récompense d'un agent a_i dans une coalition C est notée v_{a_i}, celle de la coalition C s'exprime par V_c=$\sum_{a_i \in c} v_{a_i}$.

Définition 2

La fonction d'utilité qu'un agent a_i cherche à maximiser en participant dans la coalition C s'écrit :

$$u_c^{a_i} = v_{a_i} - Cost_c^{a_i} \tag{1}$$

Cette fonction d'utilité est une information privée vue que la fonction de coût $Cost_c^{a_i}$ est aussi privée. La seule connaissance partagée par les agents est la récompense V_c.

Définition 3

Soit $L^{Net}_{a_i}$ les dépendances de a_i avec les agents de sa vue $\vartheta^t_{a_i}$ qui forment une structure réseau Net telle que $\vartheta^t_{a_i} \subseteq Net$. $\forall a_i \in Net, \exists\, L^{Net}_{a_i}$ telle que $L^{Net}_{a_i} = \{\gamma^{Net}_{a_i}, H^{Net}_{a_i}, S^{Net}_{a_i}\}$. $\gamma^{Net}_{a_i}$ est l'ensemble des agents dont dépend a_i. $S^{Net}_{a_i}$ est l'ensemble des agents qui dépendent de a_i. $H^{Net}_{a_i}$=$(\vartheta^t_{a_i} \setminus \gamma^{Net}_{a_i}) \setminus S^{Net}_{a_i}$ est l'ensemble des agents qui ont la même importance que a_i dans l'environnement de déploiement.

Cette formulation des dépendances possibles pour un agent s'explique par le fait que, dans un environnement sinistré, des intervenants peuvent être sous les ordres d'un ensemble de leaders ou peuvent avoir des accords antérieurs auxquels ils doivent tenir compte.

3. Mécanisme de formation de coalitions : S-NRB

S-NRB fournit un ensemble de stratégies d'interactions que les agents utilisent afin de trouver le ou les meilleures coalitions A-cores dans un contexte incertain.

Définition 4

Soit une coalition C et E_C un ensemble de coalitions. Si $C \in E_C$ est une coalition A-core, alors :

- Il n'existe aucun agent ou groupe d'agents qui a la possibilité d'augmenter son utilité en changeant de coalition, c'est-à-dire, $\forall\, a_i \in C$ et $\forall\, C, C' \in E_C, u_c^{a_i} \geq u_{c'}^{a_i}$.

- Tout agent de C a une alliance non-dominée avec au moins un agent de C, c'est-à-dire, $\forall\, a_i \in C, \exists Al_{a_i,aj} : a_j \in C$.

- Les agents de C forment un ensemble connexe, ($\forall\ a_i \in C, \exists a_j \in \vartheta^t_{a_i} : a_j \in C$.)
À cause des incertitudes et de la dynamique du contexte considéré, ***S-NRB*** se fonde sur un principe de sondage parallèle et asynchrone et sur des offres composées.
Définition 5
Une offre composée pour la formation d'une coalition C d'un agent a_i est notée $O^{a_i}_c$. Formellement, $O^{a_i}_c$=$\{\{B_i, B_j, B_k, ...\}, \{pl_{a_i}\}\} : B_i$=$(Ri, \Delta i, Ui)$. B_i est un but à atteindre par l'offre qui spécifie, la ressource Ri requise pour T_c, les contraintes Δi de B_i et la récompense Ui associée à B_i. pl_{a_i}=$\{a_i, a_j, a_k, ...\} : a_i \in A$ désigne une liste de sondage qui permet de savoir dynamiquement, les agents qui acceptent de s'associer pour un but de l'offre de a_i.
Définition 6
Deux agents a_j et a_k sont en conflit pour une offre composée $O^{a_i}_c$, lorsque a_j et a_k ont la même contribution marginale (X_{a_j}=X_{a_k}), le même but $B_x \in O^{a_i}_c$ et veulent tous les deux participer à la même coalition C.
Notations
$\equiv$ L'équivalence entre deux paramètres. $\not\equiv$ La non équivalence entre deux paramètres.
$\succsim_P$ Dominance au sens de Pareto. $\succsim_L$ Dominance au sens de Lorenz.

3.1. Les étapes de notre mécanisme S-NRB

Le mécanisme ***S-NRB*** (cf. algorithme 1), se résume en ces quatre étapes :
Étape 1 : Stratégies de diffusion des offres de formation de coalition C.
Tout agent a_j qui reçoit une offre $O^{a_i}_c$ pour former C doit envoyer au maximum un message avec un TTL=1 (cf. équation 6) dans son voisinage $\vartheta^t_{a_j}$. Ce message doit indiquer sa réponse à l'agent par qui il a reçu cette offre. Nous imposons $TTL = 1$ pour permettre aux agents de privilégier les agents de leur vue et pour simplifier les négociations.
- Un agent a_j ne diffuse le message de sondage contenant l'offre $O^{a_i}_c$ dans $\vartheta^t_{a_j}$ que s'il accepte de participer à C (cf. ligne 1 - 4 algorithme 1). Cet envoi du message est effectué après avoir vérifié que son but $G_{a_j} \equiv B_i : B_i \in O^{a_i}_c$, que son utilité $U^t_{a_j}$ à cet instant t est supérieure à son utilité $U^{t-1}_{a_j}$ à l'instant $t-1$ et après avoir inscrit son identifiant dans la liste de sondage $pl_{a_i} \subset O^{a_i}_c$.
À noter que, si un agent accepte une offre, il supprime de cette offre l'information sur son but $G_{a_j} \equiv B_i : B_i \in O^{a_i}_c$ afin de n'envoyer dans son voisinage que les buts qui restent à atteindre et pour réduire le nombre de conflits de buts à satisfaire, par la coalition à former, qui peuvent apparaître entre les agents. De plus, cette suppression des informations sur les buts déjà acquis permet dans une certaine mesure, d'accélérer la convergence du mécanisme et de favoriser les agents proches de l'agent qui a initié l'offre de coalition.
- Un agent peut répondre à l'offre par un rejet ($ResetProbe(O^{a_i}_c)$, cf. ligne 26 - 28 algorithme 1) ou par une demande de modification de l'offre (*counter($O^{a_i}_c$)*, cf. ligne 23 - 25 algorithme 1). C'est seulement dans le cas d'une contre-proposition, qu'un agent peut entrer en négociation, si par exemple, il n'y a pas suffisamment d'agents qui ont accepté l'offre ou si la contre-proposition n'engendre pas d'incohérence sur l'offre $O^{a_i}_c$ proposée.
Étape 2 : Stratégies d'optimisation des messages envoyés.
$\forall\ a_j$ tel que, $\forall\ a_k \in \vartheta^t_{a_j}, a_k \in pl_{a_i}$ (c'est-à-dire, un agent avec des voisins déjà dans la liste de sondage) ou $\vartheta^t_{a_j} = \emptyset$ (c'est-à-dire, un agent sans voisin), a_k doit rendre sa décision (acceptation ou contre-proposition ou rejet) pour mettre fin au sondage sur le chemin de recherche auquel il appartient.
À cause des contraintes des équipements (e.g. portée du signal, débit des liens de communications) où les agents sont hébergés, parfois, un agent ne peut pas contacter au même

moment tous les autres agents de l'environnement de déploiement. Comme conséquence, plusieurs listes de sondage pl_{a_i}, pl'_{a_i}, $pl"_{a_i}$, ..., peuvent coexister, d'où la nécessité de la détection dynamique des listes de sondage redondantes.

Étape 3 : Assurer qu'une offre n'a qu'une liste de sondage.

Tout agent qui a reçu plus d'une liste de sondage et qui veut participer à C doit effectuer une fusion des différentes listes de sondage afin de limiter les conflits entre les agents et de diffuser une seule liste de sondage dans sa vue (cf. ligne 11 - 14 algorithme 1). Soit $pl^1_{a_i}, pl^2_{a_i}, pl^3_{a_i}$ un ensemble de 3 listes de sondage pour une même offre $O^{a_i}_c$. La fusion signifie la production d'une seule liste de sondage pl_{a_i} résultante qui ne contient pas de conflits entre les agents. Formellement, pl_{a_i}=$(pl^1_{a_i} \cup pl^2_{a_i} \cup pl^3_{a_i}) : (pl^1_{a_i} \cap pl^2_{a_i} \cap pl^3_{a_i}) = \emptyset$ et $\forall a_x, a_y \in pl_{a_i}, a_x \neq a_y$, alors $G_{a_x} \not\equiv G_{a_y}$.

Étape 4 : Prise d'une décision collective sur l'état du sondage.

Si $(\cup Bi \subset O^{a_i}_c) \neq (\bigcup_{a_j \in pl_{a_i}} G_{a_j})$, alors les agents qui sont dans la liste cherchent à prédire s'ils sont capables de former une coalition stable et dans quelle mesure y parvenir (cf. ligne 20 - 22 algorithme 1). Pour ce faire, ils utilisent la loi de probabilité hypergéométrique [3], [4] . Cette loi de probabilité est la loi a priori pour la prédiction de l'issue probable d'un sondage. Ainsi, $\forall\ a_j \in pl_{a_i}$, *échantillon*=$n$=$Card(pl_{a_i})$, *population*=$N$=$Card(\cup B_i) : B_i \in O^{a_i}_c$. Soit $P_{counter}$ la probabilité d'avoir un agent qui accepte de rejoindre la liste de sondage après avoir renégocié l'offre. Soit P_{reject} la probabilité d'avoir un agent qui rejette l'offre. $\forall\ a_j \in pl_{a_i} : a_x \in \vartheta^t_{a_j}, a_x \notin pl_{a_i}$, il calcule $Q_{counter}$=$1 - (P_{counter} + P_{reject})$ c'est-à-dire la probabilité de contacter un agent qui accepte sans renégocier l'offre. Soit k le nombre de fois qu'il est encore nécessaire de contacter des agents pour former C, alors la probabilité de X=k est donnée par :

$$P[X = k] = \frac{C^k_{NQ_{counter}} C^{n-k}_{NP_{counter}}}{C^n_N} \tag{2}$$

L'espérance de cette probabilité est :

$$E(X) = nQ_{counter} \tag{3}$$

La décision de poursuivre leurs recherches dépend des valeurs de ces indicateurs. En effet, si $P[X = k] \neq 0$, l'agent responsable de la poursuite du sondage est celui de la liste de sondage qui domine au sens de Lorenz les autres agents de cette liste si on considère les paramètres $P[X = k]$ et $E(X)$. La dominance de Lorenz à été proposée par [5].

Considérons deux agents a_j et a_k, tel que :

pour a_j, notons la probabilité $P[X = k]$ par $P_j[X = k]$ et l'espérance $E(X)$ par $E_j(X)$.

pour a_k, notons la probabilité $P[X = k]$ par $P_k[X = k]$ et l'espérance $E(X)$ par $E_k(X)$.

La dominance de Lorenz pour faire un choix entre ces deux agents a_j et a_k, en utilisant $P[X = k]$ et $E(X)$, revient à poser comme vecteur de Lorenz généralisé associé :

pour a_j, $L(x)$=$(x_1, x_1 + x_2)$=$(P_j[X = k], P_j[X = k] + E_j(X))$ et

pour a_k, $L(y)$=$(y_1, y_1 + y_2)$=$(P_k[X = k], P_j[X = k] + E_k(X))$.

Si $\sum^2_{i=1} L_i(x) \geq \sum^2_{i=1} L_i(y)$, alors $x \succsim_L y \Leftrightarrow L(x) \succsim_P L(y)$. Cela, signifie que x domine y au sens de Lorenz et donc a_j est responsable de la poursuite du sondage.

Notons que le premier identifiant d'une liste de sondage (e.g. pl_{a_i}) est celui de l'agent a_i qui a initié l'offre $O^{a_i}_c$. Cela, permet de différencier les offres.

Mécanisme de résolution des conflits.

Ce mécanisme que nous notons S-NRB-CRP (Sequential Non-Return Broadcast Conflit

Resolution Protocol) est fondé sur :
Étape 1 : Matching(comparaison) et mesure de dissemblances.
L'indice et la distance de Jaccard [6] sont deux métriques utilisées pour évaluer la similarité et la diversité entre des ensembles. Soit n ensembles S_1, S_2, ..., S_n, l'indice de Jaccard est :

$$J(S_1, S_2, ..., S_n) = \frac{Card(S_1 \cap S_2 \cap ... \cap S_n)}{Card(S_1 \cup S_2 \cup ... \cup S_n)} \quad (4)$$

La distance de Jaccard mesure la dissemblance entre les ensembles. Elle consiste à soustraire l'indice de Jaccard à 1, c'est-à-dire, $J_\delta(S_1, S_2, ..., S_n) = 1 - J(S_1, S_2, ..., S_n)$.

$$J_\delta(S_1, S_2, ..., S_n) = \frac{Card(S_1 \cup S_2 \cup ... \cup S_n) - Card(S_1 \cap S_2 \cap ... \cap S_n)}{Card(S_1 \cup S_2 \cup ... \cup S_n)} \quad (5)$$

À la ligne 4 de l'algorithme 2, $Jacc(a_j)$ et $Jacc(a_k)$ signifie que a_j et a_k doivent donner respectivement leur liste $L_{a_j}^{Jacc}$ et $L_{a_k}^{Jacc}$ qui contiennent respectivement l'ensemble des agents des listes de sondage pl_{a_i} de a_j et pl'_{a_i} de a_k avec qui, ils ont des accords (alliances, dépendances) pour participer à C. Si nous notons par $AllianceSet^{a_x}$ l'ensemble des agents alliés d'un agent a_x, alors : $L_{a_j}^{Jacc}$=$(\gamma_{a_j}^{Net} \cup S_{a_j}^{Net} \cup AllianceSet^{a_j}) \cap pl_{a_i}$.
$L_{a_k}^{Jacc}$=$(\gamma_{a_k}^{Net} \cup S_{a_k}^{Net} \cup AllianceSet^{a_k}) \cap pl_{a_i}$.
$Jaccard(a_j, a_k)$ est l'indice de Jaccard entre les deux ensembles $L_{a_j}^{Jacc}$ et $L_{a_k}^{Jacc}$.
Étape 2 : Prise de décision en se fondant sur la dominance au sens de Lorenz.
Il utilise les ensembles $L_{a_j}^{Jacc}$ et $L_{a_k}^{Jacc}$, sur les fiabilités ρ_{a_j} et ρ_{a_k} et sur les probabilités de stabilité $P_s^{a_j}$ et $P_s^{a_j}$. Si $L(x)$=$(x_1, x_1+x_2, x_1+x_2+x_3)$=$(Card(L_{a_j}^{Jacc}), Card(L_{a_j}^{Jacc})+\rho_{a_j}, Card(L_{a_j}^{Jacc})+\rho_{a_j}+P_s^{a_j})$ et $L(y)$=$(y_1, y_1+y_2, y_1+y_2+y_3)$=$(Card(L_{a_k}^{Jacc}), Card(L_{a_k}^{Jacc})+\rho_{a_k}, Card(L_{a_k}^{Jacc}) + \rho_{a_k} + P_s^{a_k})$. Alors, si $\sum_{i=1}^{3} L_i(x) \geq \sum_{i=1}^{3} L_i(y)$, alors $x \succsim_L y \Leftrightarrow L(x) \succsim_P L(y)$ signifie que x domine y au sens de Lorenz. Alors, la participation de a_j est préférée à celle de a_k.

3.2. Analyse de S-NRB

Nous donnons un ensemble de propriétés de S-NRB qui mène à l'existence de coalitions *A-core*.

Lemme 1 *Qu'une possibilité de former une coalition core stable existe ou non, S-NRB termine toujours sans blocage.*

Lemme 2 *S'il est possible de former une coalition core stable, S-NRB permet aux agents de converger vers celle-ci.*

Théorème 1 *S-NRB permet toujours la converge vers une coalition core stable si celle-ci existe.*

Lemme 3 *Le processus de sondage de notre mécanisme S-NRB fait émerger une structure connexe d'agents pl_{a_i} où chaque agent a_j de cette structure a au moins un agent voisin a_k dans pl_{a_i} avec lequel il est accepte de partager la coalition C.*

Lemme 4 *Les agents les plus proches et les plus fiables du voisinage de l'agent qui a initié le mécanisme de formation de coalitions S-NRB ont toujours une plus grande probabilité d'être choisi pour former la coalition.*

Théorème 2 *S-NRB permet une auto-stabilisation des coalitions core stables.*

4. Évaluations de performances de notre mécanisme S-NRB

Nous effectuons ces simulations sur notre plateforme basée sur JADE (JAVA Agent DEvelopment Framework). Les indisponibilités, les préférences et les dépendances des agents et les états des tâches suivent des distributions de probabilités de loi Uniforme ou Normale. Ces deux lois sont les lois a priori à chaque fois qu'on ignore l'ensemble des situations qui peuvent justifier l'apparition de certains évènements. Aléatoirement, un agent autonome et égoïste avec des dépendances natives ou non peut initier un processus de formation de coalitions, être libre de toute dépendance ou devenir indisponible. La dynamique de la localisation, les propriétés et les états des tâches sont générées aléatoirement de même que pour les ressources nécessaires pour l'exécution d'une tâche à un instant t donné. Le nombre d'agents dans l'environnement est choisi entre 10 et 100 et le nombre de tâches simultanées est compris entre 3 et 8 tâches. Les simulations sont effectuées dans une machine Intel i7 (4 GHZ) avec 4GB de RAM. À cause de la dynamique des ressources disponibles ou requises pour une tâche, nous faisons nos simulations en considérant les *ratios*. Par exemple, s'il y a 12 agents de même caractéristiques. Si l'objectif est de former deux coalitions où la première requiert 9 agents et la seconde 6 agents, alors nous avons, *ratio*=$\frac{12}{9+6}$=$\frac{12}{15}$=0.8. Pour obtenir des valeurs moyennes, nous répétons 10 fois chaque simulation. Nous étudions ici, la capacité des coalitions *core stables* formées

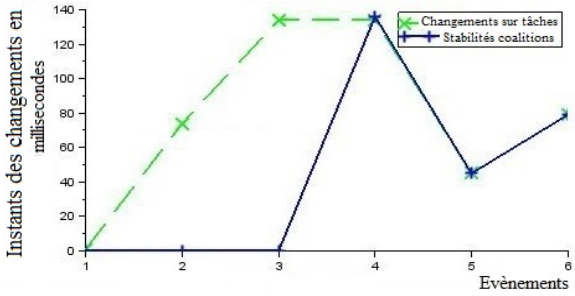


Figure 1. *Auto-stabilisation des coalitions si les changements d'états des tâches et la disponibilité des agents suivent une distribution de probabilités Uniforme* $U(a, b)$ *où* a=0 *et* b=10. *Évolutions des* $ratios$ *: 1.42, 1.66, 2, 3.33, 1.66 et 2.*

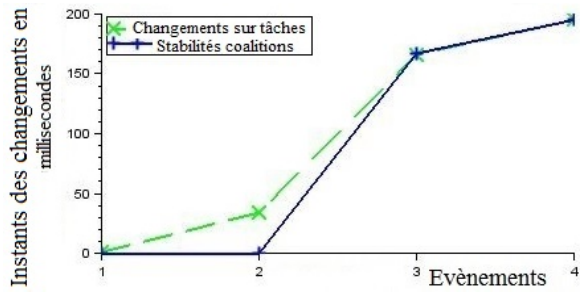


Figure 2. *Auto-stabilisation des coalitions si les changements d'états des tâches et la disponibilité des agents suivent une distribution de probabilités Normale* $N(\mu, \sigma^2)$ *où* μ=0.5 *et* σ=0.1. *Évolutions des* $ratios$ *: 0.90, 3.33, 3.33, 2.50.*

à s'auto-stabiliser. L'axe des abscisses (*Evènements*) répertorie l'ensemble des changements sur les $ratios$ et sur les tâches. Les résultats des simulations illustrés par ces figures montrent que, si une coalition est *core stable*, elle le reste même s'il y a des changements dynamiques sur les $ratios$ et les tâches.

5. Conclusion

Ce travail aborde la mise en œuvre d'un mécanisme de formation de coalitions *core stables* adaptatif, asynchrone et décentralisé qui permet une auto-stabilisation des coalitions formées d'agents égoïstes (coalitions *A-cores*). Pour ce faire, nous avons proposé le mécanisme S-NRB. Nous avons, principalement, prouvé de manière théorique et expérimentale, sa convergence vers des coalitions *A-cores* en cas de disparition d'un ensemble d'agents ou de changements sur les contraintes des tâches et en présence d'incertitudes sur les dépendances, les conflits, les préférences et la disponibilité des agents. À cause des incertitudes et des contraintes, nous ne supposons aucune connaissance a priori sur les coalitions stables à former pour les tâches stochastiques à exécuter et qu'il n'est pas possible de calculer de manière centralisée les coalitions à former et qu'il est impossible de prédire la stabilité des coalitions avant le début de l'exécution des tâches.

6. Bibliographie

[1] B. HORLING, R. VINCENT, R. MAILLER, J. SHEN, R. BECKER , K. RAWLINS , V. LESSER, « Distributed sensor network for real time tracking », *AGENTS'01 Autonomous Agents,ACM Press*, 417-424, 2001.

[2] O. SHEHORY, S. KRAUS, « Methods for Task Allocation via Agent Coalitions Formation », *AI Journal*,May, 165-200, 1998.

[3] R. D. YATES, D. J. GOODMAN, « Probability and Stochastic Processes : A Friendly Introduction for Electrical and Computer Engineers », *John Wiley and Sons, INC*, 2005.

[4] B. BAYNAT, « Théorie Des Files D'attente - Des Chaînes De Markov Aux Réseaux À Forme Produit », *Hermes Science Publications - Lavoisier*, June, 1-328, 2000.

[5] A. W. MARSHALL, I OLKIN, « Inequalities : Theory of Majorization and Its Applications », *The Mathematics in Science and Engineering Series*, n° 143, 1-569, 1979.

[6] K. JAHANBAKHSH, V. KING, G. C. SHOJA, « Predicting missing contacts in mobile social networks », *Pervasive and Mobile Computing, Elsevier*, n° 8, issue 5, 698-716, October, 2012.

[7] P.F. FAYE, S. AKNINE, O. SHEHORY, M. SÈNE, « Stabilizing Agent's Interactions in Dynamic Contexts », *AINA 2014*, May, 925-932, 2014.

[8] P.F. FAYE, S. AKNINE, M. SÈNE, O. SHEHORY, « Dynamic Coalitions formation in Dynamic Uncertain Environments », *IAT 2015*, December, 273-276, 2015.

[9] M. A. KHAN, D. TURGUTL. BÖLÖNI, « Optimizing coalition formation for tasks with dynamically evolving rewards and nondeterministic action effects », *Autonomous Agents and Multi-Agent Systems*, May, 415-438, 2011.

7. Annexe A

Preuve du lemme 1
Soit $U^t_{a_j}$ l'utilité de a_j à l'instant t. $\forall Probe(O^{a_i}_c)$ d'un agent a_i, un agent $a_j \neq a_i$ ne diffuse ce message de sondage que si $U^{t-1}_{a_j} \leq U^t_{a_j}$ et $\vartheta^t_{a_j} \neq \emptyset$, ce qui est une manière de gérer les coûts des communications et d'éviter les négociations en boucles. Tout conflit entre a_i et a_j est géré par les autres agents appartenant à leur liste de sondage grâce au calcul de la dominance de Lorenz entre a_i et a_j. Cette méthode évite le cas où a_i et a_j sont dans une impasse au moment où les agents de leur liste attendent la validation de la liste de sondage en une coalition. Ainsi, les blocages sont évités grâce à cette gestion des communications et la résolution décentralisée des conflits. Ce qui prouve notre lemme.□
Preuve du lemme 2
Soit $U^t_{a_i}$ l'utilité de a_i à l'instant t et $U^t_{a_j}$ l'utilité de a_j à l'instant t. $\forall B_i \in G_c$ et $B_i \in O^{a_i}_c$, si $\exists a_i, a_j \in A : R_{a_i} \in B_i, Aut^t_{ai} \neq 0, U^{t-1}_{a_i} \leq U^t_{a_i}$ et $R_{a_j} \in B_i, Aut^t_{aj} \neq 0, U^{t-1}_{a_j} \leq U^t_{a_j}$, alors, a_i (respectivement a_j) va atteindre un accord en utilisant ses dépendances $L^{Net}_{a_i}$ (respectivement $L^{Net}_{a_j}$). Cela se vérifie à la ligne 1 de l'algorithme 2. L'utilisation de l'index de Jaccard afin d'évaluer les différences entre les agents a_i et a_j, par rapport à leurs alliances et à leurs fiabilités, ainsi que la dominance de Lorenz permettent de faire un choix entre les deux agents. Si la liste de sondage pl_{a_i} est validée comme une coalition C, cela signifie que :
(1) $\forall a_i \in pl_{a_i}, \exists a_j \in pl_{a_i} : a_j \in \gamma^{Net}_{a_i}$ ou $a_j \in H^{Net}_{a_i}$ ou $a_j \in S^{Net}_{a_i}$;
(2) $\forall a_i \in pl_{a_i}, U^{t-1}_{a_i} \leq U^t_{a_i}$;
(3) $\forall a_i \in pl_{a_i}, \exists a_j \in pl_{a_i} : \exists Al_{a_i,a_j}$.
De là, C sera *core stable* car aucun agent $a_i \in pl_{a_i}$ ne sera motivé à quitter C pour une autre coalition C' et risquer ainsi de perdre ses alliance(s), de réduire son utilité et être considéré comme non fiable par les autres agents. Ce qui prouve notre lemme.□
Preuve du théorème 1
Le lemme 1 montre que S-NRB termine son exécution et le lemme 2 prouve que S-NRB mène à la formation de coalitions *core stables* si ces dernières peuvent exister entre les agents. Ainsi, nous pouvons affirmer que S-NRB cherche toujours à déterminer une coalition stable qui maximise le bien-être social des agents quels que soient les états initiaux des agents. Ce qui prouve notre théorème.□
Preuve du lemme 3
$\forall a_j \in A$, s'il reçoit une offre $O^{a_i}_c$, cela signifie que, $a_j \in \vartheta^t_{a_j}$ ou $\exists \{a_x, a_y, ...\} \in A : a_j \in L^{Net}_{a_i}$ et $a_i \in L^{Net}_{a_j}$. Ainsi, si a_i et $a_j \in pl_{a_i}$ alors, $a_j \in \vartheta^t_{a_i}$ ou $\exists \{a_x, a_y, ...\} \in pl_{a_i}$ car dans notre approche a_x ne diffuse pl_{a_i} que s'il appartient à pl_{a_i} (cf. lignes 3 - 5 algorithme 1) ou s'il existe un message de mise à jour de pl_{a_i} ($UpdateProbe(O^{a_i}_c)$). Dans ce dernier cas, si un agent a_j souhaite atteindre un but $B_i \in O^{a_i}_c$ il doit appartenir à $\vartheta^t_{a_i}$ ou à $\vartheta^t_{a_x} : a_x \in pl_{a_i}$. Ce qui prouve notre lemme.□
Preuve du lemme 4
Considérons les agents a_i, a_j et a_k tels que $a_j \in \vartheta^t_{a_i}$ et $a_k \in \vartheta^t_{a_j}$. Si $G_{a_j} \equiv G_{a_k} \equiv B_i \in O^{a_i}_c$ et $U^{t-1}_{a_j} \leq U^t_{a_j}$ et $U^{t-1}_{a_k} \leq U^t_{a_k}$ alors, a_j et a_k ne seront pas en conflits car si a_j reçoit l'offre $O^{a_i}_c$, il ajoute son identifiant dans la liste de sondage pl_{a_i} et supprime l'information concernant $G_{a_j} \equiv B_i \in O^{a_i}_c$ avant d'envoyer l'offre $O^{a_i}_c$ à $a_k \in \vartheta^t_{aj}$.
Ainsi, si a_k reçoit $O^{a_i}_c$ alors l'information $G_{a_k} \equiv B_i \notin O^{a_i}_c$ et donc aucun conflit n'arrive entre a_j et a_k. Ce qui prouve qu'un agent plus proche est toujours privilégié s'il faut former une coalition. Maintenant supposons que, $a_j, a_k \in \vartheta^t_{a_i}$ alors, les deux agents vont recevoir l'offre $O^{a_i}_c$ en même temps.

Si $G_{a_j} \equiv G_{a_k} \equiv B_i \in O_c^{a_i}$ et $U_{a_j}^{t-1} \leq U_{a_j}^t$ et $U_{a_k}^{t-1} \leq U_{a_k}^t$ alors, a_i, a_j et a_k vont procéder à la résolution du conflit en utilisant l'algorithme 2. Cela mène à la détermination de l'agent le plus fiable entre a_j et a_k. Ce qui donne la preuve de notre lemme.□

Preuve du théorème 2

Le lemme 2 prouve la convergence de S-NRB vers des coalitions *core stables* où aucun agent n'est motivé à quitter sa coalition. Le lemme 3 montre que, chaque évènement qui impacte dynamiquement l'état d'une (des) tâche(s) ou la disponibilité d'un ensemble d'agents sera détecté par au moins un agent de la liste de sondage ou de la coalition. Les lemmes 1 montrent qu'après une instabilité d'une coalition, cette dernière redevient stable après un nombre fini d'étapes. De plus, pour tout agent ajouté dans une coalition *core stable*, les conflits éventuels sont gérés de manière décentralisée et la décision d'ajouter un ensemble d'agents dans une coalition se fait en respectant les préférences des agents qui sont déjà dans la coalition. Avec ces contraintes de participation dans une coalition, notre mécanisme S-NRB évite le cas où un ensemble d'agents est motivé à quitter sa coalition après une instabilité causée par un ensemble de conflits ou à cause d'une décroissance de leur utilité. Ce qui prouve notre théorème.□

8. Annexe B

Règle 1 : offre et participation valide. Un agent a_i peut proposer une offre $O_c^{a_i}$ si et seulement si $a_i \in C$ et chaque but $B_i \in O_c^{a_i}$ est requis pour la formation et la stabilité de C. a_i peut conseiller (cf. règle 2) un autre agent sur les modifications à apporter sur son offre pour avoir des participants potentiels à la coalition C. Un agent a_j peut accepter $B_i \equiv G_{a_j} \in O_c^{a_i}$ si et seulement si $Ri \in B_i : Ri \in R_{a_j}$. De plus, a_j ne peut participer à C que si $\forall a_i \in C, u_{c'}^{a_i} \geq u_c^{a_i} : C'$=$C \cup a_j$.■

Règle 2 : Conseiller un agent. Les conseils de modification des offres ont pour objectif de permettre aux agents d'améliorer leurs connaissances sur les autres agents et sur les offres pouvant les intéresser. a_j peut conseiller a_i ($Counsel(a_i)$) sur des modifications à apporter sur $O_c^{a_i}$, si Al_{a_i,a_j} existe, si $G_{a_i} \in O_c^{a_i}$ et si cela peut éventuellement intéresser un agent appartenant à $L_{a_j}^{Net} \neq \emptyset$. Ainsi :

(1) Un conseil doit assurer que la modification de l'offre pour intéresser un agent a_j à C ne va pas diminuer l'utilité d'un agent qui a déjà donné son accord pour participer à C, c'est-à-dire, $\forall a_i \in C\ u_{c'}^{a_i} \geq u_c^{a_i} : C'$=$C \cup a_j$,

(2) a_j peut conseiller a_i s'il existe une *alliance* Al_{a_i,a_j}, sinon a_i ignore le conseil de a_j.

(3) L'expression formelle d'un conseil est sous la forme : $Counsel(a_i)$=$(O_c^{a_i}, B_i)$, qui signifie que le conseil de modification de l'offre $O_c^{a_i}$ de a_i concerne le but $B_i \in O_c^{a_i}$.■

Règle 3 : Acceptation d'une offre. a_j peut accepter une offre si $B_i \equiv G_{a_j} \in O_c^{a_i}$ et $Ri \in B_i : Ri \in R_{a_j}$. a_j peut participer dans C si $\forall a_i \in C, u_{c'}^{a_i} \geq u_c^{a_i} : C'$=$C \cup a_j$.

Une fois que a_j accepte l'offre de a_i, $B_i \equiv G_{a_j} \in O_c^{a_i} : a_i \in C$ et $a_j \in A \setminus C$, alors les deux agents mettent à jour leurs connaissances sur la(les) tâche(s) (état, localisation, etc.) et la stabilité de leur coalition C si elle est validé.

Si a_j formule une contre-proposition de l'offre $O_c^{a_i}$ telle que $B_i \equiv G_{a_j} \in O_c^{a_i} : a_i \in C$, alors a_j est automatiquement désengagé de tout accord précédent et doit refaire une négociation avec les autres agents ou avec a_i. Cependant, si la coalition a été validée, la fiabilité de a_j est d'abord réduite par tous les autres agents avant de vérifier si une autre négociation peut être entreprise avec a_j. Un agent peut se désengager d'un accord s'il doit pénaliser un autre agent de C ou s'il devient indisponible.■

Algorithme 1 : : S-NRB

Require: $Probe(O_c^{a_i}) : \vartheta_{a_i}^t \neq \emptyset, O_c^{a_i}$=$\{\{B_i : i \in [1, |T_c|]\}, \{pl_{a_i} : i \in [1, |A|]\}\}$
RESULT : Coalition C

1: **if** $G_{a_j} \equiv B_i \in O_c^{a_i}$ et $Aut_{aj}^t \neq 0$ et $U_{a_j}^{t-1} \leq U_{a_j}^t$ **then**
2: **if** $\vartheta_{a_j}^t \neq \emptyset$ et $a_j \in pl_{a_i}$ **then**
3: Envoyer $Probe(O_c^{a_i}) : pl_{a_i}$=$\{a_i, a_j\}\ \forall\ a_k \in \vartheta_{a_j}^t$
4: **end if**
5: **if** $\vartheta_{a_j}^t$==$\emptyset$ et $a_j \in pl_{a_i}$ **then**
6: Envoyer $CommitProbe(O_c^{a_i}) : pl_{a_i}$=$\{a_i, a_j\}\ \forall\ a_k \in \vartheta_{a_j}^t$
7: **end if**
8: **if** $\vartheta_{a_j}^t \neq \emptyset$ et $a_j \in pl_{a_i}$ et $\exists\ CommitProbe(O_c^{a_i})$ **then**
9: Envoyer $Commit(pl_{a_i})$
10: **end if**
11: **if** ($\exists\ a_j \in pl_{a_i}$) ou ($\exists\ a_k \in pl'_{a_i}$) et ($a_k \notin pl_{a_i}$) **then**
12: $UpdateProbe(O_c^{a_i})$ *//pour mettre à jour* pl_{a_i} *après fusion de* pl_{a_i} *et* pl'_{a_i}
13: **end if**
14: **if** $\exists$ conflit entre a_j et a_k **then**
15: $Matching(pl_{a_i}, pl'_{a_i})$ *//pour initier l'exécution de S-NRB-CRP() (algorithme 2)*
16: Envoyer $UpdateProbe(O_c^{a_i})$ *//pour mettre à jour* pl_{a_i}
17: **end if**
18: **if** $\cup Bi \subset O_c^{a_i}$==$\cup_{a_j \in pl_{a_i}} G_{a_j}$ **then**
19: Valider C=$\{\forall\ a_j \in pl_{a_i}\}$ et G_c=$\cup_{a_j \in pl_{a_i}} G_{a_j}$
20: **else**
21: $\forall\ a_j \in pl_{a_i}$ trouver plus de ressources en utilisant $Al_{a_j,a_k} : a_k \in A$, $L_{a_j}^{Net}$ et $Counsel(a_i)$ *//Le* $Counsel(a_i)$ *est pour guider les recherches de* a_i *(cf. règle 2 section règles générales).*
22: **end if**
23: **else**
24: **if** $G_{a_j} \equiv B_i \in O_c^{a_i}$ et $Aut_{aj}^t \neq 0$ et $U_{a_j}^{t-1} > U_{a_j}^t$ **then**
25: Envoyer $Counter(O_c^{a_i})$
26: **else**
27: Envoyer $ResetProbe(O_c^{a_i})$
28: **end if**
29: **end if**

Règle 4 : Schéma de communication. Chaque agent peut formuler une offre, faire une contre-proposition ou conseiller la modification d'une offre à tout moment car il n'y a pas d'ordonnanceur. Cependant, pour : - assurer une convergence des négociations, - gérer la consommation de ressources, et - les offres en boucle, (1) toutes les communications se font en mode *non-return broadcast* et (2) chaque message à un *Time To Live (TTL)*. Le mode *non-return broadcast* signifie que lorsqu'une information de mise à jour (disparition d'agents de C, évolution sur une tâche, nouvelle offre, ...) qui arrive d'un agent a_k, ses agents voisins $\vartheta_{a_j}^t$ ne peuvent pas lui retourner la même information de mise à jour. Le *TTL* permet de définir le nombre de sauts autorisés pour un message.

$$1 \leq TTL \leq |\frac{Sz}{2 * \Upsilon}| \tag{6}$$

où Sz est la taille de l'environnement couvert par la tâche, Υ la portée du signal du composant qui héberge l'agent qui a initié le message (e.g. Bluetooth 100 mWatt, $\Upsilon \leq$100 mètres). Ainsi, le *TTL* permet de confiner les messages car un message n'est envoyé que si son *TTL* n'est pas épuisé. Cependant, même si un agent reçoit un message, il ne répond

ou ne le diffuse en mode *non-return broadcast* que selon ses propres stratégies qui visent à augmenter son efficacité et à ménager ses ressources.■

Algorithme 2 : : S-NRB-CRP()

Require: $\exists\ (a_j \in pl_{a_i})$ et $(a_k \notin pl'_{a_i})$ et G_{a_j}==G_{a_k} et X_{a_j}==X_{a_k}.
RESULT : $pl_{a_i} \cup pl'_{a_i} : pl_{a_i} \cap pl'_{a_i} = \emptyset$

if $a_j \in S_{a_k}^{Net}$ **then**
  Notifier à a_j que sa demande de participation est annulée
else
  $L_{a_j}^{Jacc}$=$Jacc(a_j)$ et $L_{a_k}^{Jacc}$=$Jacc(a_k)$
  $Jaccard(a_j, a_k)$=$|\frac{L_{a_j}^{jac} \cap L_{a_k}^{jac}}{L_{a_j}^{jac} \cup L_{a_k}^{jac}}|$
  if $1 - Jaccard(a_j, a_k) \neq 0$ **then**
    if $Card(L_{a_j}^{Jacc}) < Card(L_{a_k}^{Jacc})$ **then**
      Notifier à a_j que sa demande de participation est annulée
    else
      Notifier à a_k que sa demande de participation est annulée
    end if
  else
    $L(x)$=$(x_1, x_1 + x_2, x_1 + x_2 + x_3)$=$(Card(L_{a_j}^{Jacc}), Card(L_{a_j}^{Jacc}) + \rho_{a_j}, Card(L_{a_j}^{Jacc}) + \rho_{a_j} + P_s^{a_j})$
    $L(y)$=$(y_1, y_1 + y_2, y_1 + y_2 + y_3)$=$(Card(L_{a_k}^{Jacc}), Card(L_{a_k}^{Jacc}) + \rho_{a_k}, Card(L_{a_k}^{Jacc}) + \rho_{a_k} + P_s^{a_k})$
    if $x \succsim_L y$ **then**
      Notifier à a_k que sa demande de participation est annulée
    end if
    if $y \succsim_L x$ **then**
      Notifier à a_j que sa demande de participation est annulée
    end if
  end if
  if $\exists\ (a_j \in pl_{a_i})$ et $(a_k \notin pl'_{a_i})$ et G_{a_j}==G_{a_k} et X_{a_j}==X_{a_k} **then**
    Sélectionner un agent tel que, $Max(\rho_{a_j}, \rho_{a_k})$, $Max(Aut_{a_j}^t, Aut_{a_k}^t)$ et $Max(P_s^{a_j}, P_s^{a_k})$.
  end if
end if

Règle 5 : Terminaison d'une négociation. Une fois qu'un agent accepte ou reçoit une acceptation d'une offre, il doit interrompre la négociation en envoyant un message d'information à son voisinage pour faire part de sa décision ou du résultat de la négociation. Tout agent qui voit son offre refusée par un autre doit interrompre sa négociation avec cet agent et se tourner vers ceux susceptibles de l'accepter. Si un ensemble d'agents rivalisent pour la même offre, les agents retenus pour la coalition sont confirmés par un message de validation de participation tandis que les autres reçoivent un message de rejet. Ces messages mettent fin aux négociations. Ainsi, pour un agent dont la participation est rejetée, il doit interrompre la négociation afin d'avoir la possibilité de négocier une autre offre. Toute négociation a une durée au-delà de laquelle tout agent qui n'a pas un accord de participation pour son offre doit la modifier ou utiliser ses dépendances pour améliorer son offre. Cette offre est supprimée dans l'impossibilité de trouver des accords de participation à une coalition pour l'offre.■

Intelligent prognostic

Towards an intelligent prognostic approach based on data mining and knowledge management

Safa Ben Salah, Imtiez Fliss and Moncef Tagina

{safa.bensalah, imtiez.fliss, moncef.tagina}@ensi-uma.tn
COSMOS Laboratory, National School of Computer Sciences, Manouba University, Tunisia

ABSTRACT. Due to the complexity of increasingly growing industrial processes, many of undetected faults can lead to catastrophic consequences for the entire system functioning. It is then crucial to detect and better more to anticipate the detection of faults. In this context, this paper presents an intelligent prognostic approach to anticipate the detection of faults that can affect a complex system. The proposed approach consists in proposing a multi-agent system using data mining and knowledge management techniques. It finally displays a list of faults that may appear to inform the human operator of the possible state of the system and help him to take the necessary preventive measures.

KEYWORDS : Fault prognostic, Complex Systems, Data Mining, Knowledge Management, Multi-Agent Systems.

1. Introduction

During their life cycle, industrial systems are prone to faults, which can cause a great damage or even disasters. The need to improve the availability, reliability and thus security lead to change the way of maintenance : Passing from corrective maintenance to predictive maintenance (called in literature Condition-Based Maintenance : CBM) [1, 2]. In this context, prognostic has become a crucial strategy to avoid "catastrophic" fault results. The term prognostic founds its origin in the Greek word "prognôstikos", which means "to know in advance" and it is defined as the estimation of the time to fault of a component (or a system) and the existence of risk or subsequent occurrence of one or more fault modes.

In order to predict different types of complex systems(continuous, discrete, hybrid, centralized and distributed) and guarantee reuse and better performance of our solution, we opted for the development of a multi-agent system[9]. Different approaches to perform fault prognostic have been developed, these methods may be associated with one of the three main categories according to [2, 3] ; model-based prognostic, data-driven based prognostic and experienced-based prognostic. In model based prognostic, the physical components or system and its degradation phenomenon are represented by a set of mathematical lows. Whereas, the data-driven approach aims at transforming the raw monitoring data into relevant behavior models of the system including its degradation. Finally, the approach of prognostic based on experience take into account the data and the knowledge accumulated by experience.

To ensure performance, computation cost, convenience and accuracy of prognostic, we propose to combine the use of data mining and knowledge management. In fact, in several cases, it is difficult to obtain a model that translates accurately the system. On the other hand the proper use of expert feedback and the historical data can lead to significant gains.

A list of faults that may appear are displayed to inform the user of the possible state of the system and help preventive actions. To validate the proposed approach, we rely on a simulation of a complex industrial system : Aircraft Elevator Control System. This paper is organized as follows : the second section is dealing with the prognosis approaches to justify our choice. The third section is devoted to present the new intelligent prognostic approach we propose. The forth section is dedicated to the validation of this approach. Finally, some concluding remarks will be made.

2. Prognostic approaches

The main goal of our work is to propose a prognostic approach in order to assist human operator to properly and timely manage faults. Fault prognostic consists of estimating the time before fault of a component (or a system) and the existence of risk or subsequent occurrence of one or more fault modes. Several methods are used to produce powerful solutions to anticipate detection of faults in complex systems [17, 16, 14, 12]. These methods may be associated with one of the three main categories according to [2, 3]namely : model-based prognostic, experience-based prognostic and data-driven based prognostic. The derived model of each approach is then used to predict the future evolution of the degradation of industrial system. To choose the best approach to use, we made a study of the three approaches.

2.1. Model based prognostic

This approach consists of representing the physical components or system and its degradation phenomenon by a set of mathematical low. There are several works using this approach for instance [11] and [12].

2.2. Data-driven prognostic

This approach aims at transforming historical data into relevant behavior models of the system including its degradation. The historical surveillance data is often the fastest and most reliable source of information to understand the degradation phenomena. Indeed, some previously experienced situations can breed and therefore the prognosis system will recognize it such as the deterioration of a parameter, system transition to monitor in a state of fault, the malfunction of a component, etc. Several prognostic works are based on data-driven [4, 5, 6], etc.

2.3. Experience-based prognostic

The approach of prognostic based on experience take into account the knowledge accumulated by experience during the whole exploitation period of the complex system. In fact, the activity of supervision and control of complex systems is a very complex task that requires a great experience. This experience is gained by experts over the years.

2.4. The chosen approach

To choice the appropriate approach in our study case, we have done a comparison between the three approaches(summed up in table 1).

	Advantages	**disadvantages**
Data-Driven appoach	-easy to implement -Performance enhances over time. -Low implementation cost	-Need a lot of data -Abscence of physical implementation
Model-based approach	-Physical approach :quantification of the degradation -Precise	-reduced applicability -High implementation cost
Experience-based approach	-No physical model is required. -Simple to develop and easy to understand	-Domain expert with strong experiential knowledge -Domain expert required to develop rules.

Tableau 1. *Comparative study of three prognostic approach*

It is obvious that none of approaches has only advantages as shown in table1. Therefore, it would be interesting to combine prognostic approaches to improve their prognosis result. The integration of various characteristics is a way to develop new hybrid approaches to overcome the limitations of individual strategies of each method. As it is difficult and expensive, in several cases, to obtain a model that translates accurately the system, we proposed a hybrid approach based on combining the data-driven prognostic and experience-based prognostic.

3. A hybrid intelligent approach for prognostic based on data mining and knowledge management

We recall that the objective of our work is to define a reliable prognostic approach for monitoring complex systems and predict the faults that may possibly appear. This a complex task. Moreover, we aim to predict faults in different types of complex systems : continuous, discrete, hybrid, centralized and even distributed ones. To guarantee reuse and better performance of our solution, it will be very interesting to exploit the Multi-agent paradigm [9]. Indeed, the contribution of Multi-agent systems in this perspective is to distribute intelligence across multiple entities which can cooperate in the resolution of the prognosis procedure combing the data mining and knowledge management. Each agent in our system is specialized and has a defined role and is able to communicate with others. The used agents are : the user interface agent, data mining agent, Knowledge management agent, the simulator agent and the predictor agent. The proposed approach is based on the definition of reactive and intelligent agents that can participate in the construction of a comprehensive prognosis solution.

3.1. Reactive agents

The user interface agent and the agent simulator are both reactive agents that perform their functions without intelligence.

3.1.1. User Interface agent

The User Interface agent handles everything regarding the communication of the system with the external environment. It provides user with friendly graphical interface through which the prognosis procedure is initiated or stopped. Furthermore, this agent has interactions with different agents. Indeed, user interface agent takes care of sending the data and knowledge for the prognosis procedure and receipt of the final results.

3.1.2. Simulator agent

The simulator will be used to simulate the system model to send the current state to data mining agent and knowledge management agent.

3.2. Cognitive agents

The thee other agents are cognitive : data mining agent, Knowledge management agent and the predictor agent.

3.2.1. Data mining agent

The data provided by the User Interface agent will serve as input for the prognosis based on data mining. Data mining is a process of discovering unknown, hidden information from a large volumes of data, extracting valuable information, and then using the information to make critical business decisions. We have used very simple and easy but very powerful data mining technique for predicting the upcoming faults :decision tree [18, 20]. The decision tree is applicable to any type of data whether quantitative, qualitative or a combination of both. It allows the graphic representation of a classification procedure and it has an immediate translation in terms of decision rules. We have used C4. 5 algorithm developed by Quinlan[7] as part of our prognostic approach based on data mining. In our study, our data mining approach extracts information from the stored data by building a decision tree from which we can get decision rules as shown in figure 1.

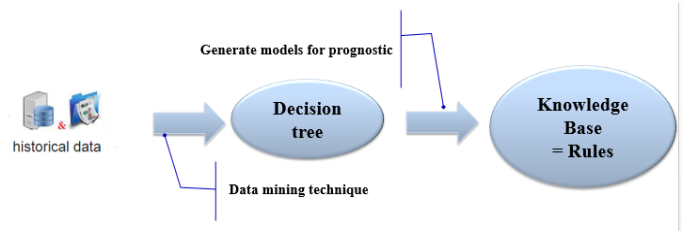


Figure 1. *Data mining technique*

3.2.2. Knowledge management agent

Expert knowledge will be used for the prognosis based on knowledge management[8]. Thus, the rules extracted from the expertise will determine the future state of components of the complex system.
Knowledge handled and implemented in our project were acquired by the study of books, articles and reference documents concerning the complex system and also from expert interviews : we directly asked questions to the expert, which helped us to understand how the system works and define the functioning rules. In fact, in order to properly use expert knowledge, these knowledge are expressed as logical rules like "If Condition then Conclusion" as shown in figure 2.

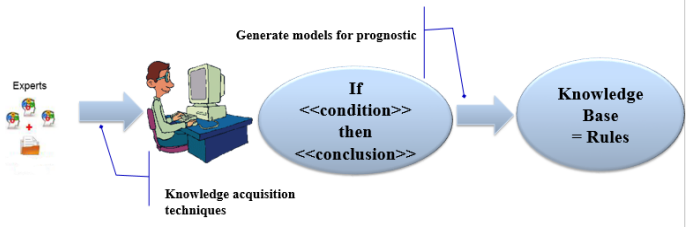


Figure 2. *Knowledge management technique*

3.2.3. predictor agent

The Predictor Agent has as role to identify the overall future state of the system. It receives the result of prognosis based on data mining and the result of the prognosis based on knowledge management and subsequently sending the list of faults that can affect the system.

3.3. Communication between agents

A fundamental charactistic of multi-agent systems is that individual agents communicate and interact. This is accomplished through the exchange of messages. The figure 4 (in annex) presents a sequence diagram showing messages exchanged between the five agents.

4. Validation of the proposed approach

To validate our multi-agent fault prognostic system, we are based on the simulation of a complex industrial system : Aircraft Elevator Control System[13].

4.1. System description

The aircraft elevator control system consists of two elevators, the control surfaces. Each of these are controlled by one of two hydraulic actuators while the other one is operating as a passive load. The four actuators take their power from three hydraulic subsystems as depicted in figure 7 (in annex). Two primary flight control units are available to compute actuator control signals and modes.

4.2. Data description

The historical data used in our simulations are the data of 8 variables (7 independent variables+ 1 dependent variable). The independent variables are the measures that we have extracted from the simulated system which are data of system components show in figure7(C1 : The right inner actuator, C2 : the right outer actuator, C3 : the left inner actuator, C4 : the left outer actuator to the, H1 : the hydraulic circuit 1, H2 : the hydraulic2 circuit H3 : hydraulic3 the circuit). The dependent variable represents the state of the system (faulty or in not). All independent variables are digital and the dependent variable is nominal. In our study, more than 10 000 values for each variable were recorded every 0. 1 seconds (as simulation lasts 100 seconds).

4.3. Knowledge Description

After studying the reference materials and the discussions with experts, the following rules are generated :

– If the aircraft is flying perfectly level, then the actuator position should maintain a constant value.

– If the position of an actuator increases or decreases by 10 cm from this zero point, then the fault detection system registers a fault in that actuator.

– The fault detection system also registers a fault if the change in actuator position is very rapid (i. e. , the position changes at least 20 cm in 0.01 seconds).

– the fault detection system registers a fault in one of the hydraulic circuits if the pressure is out-of-bounds or if the pressure changes very rapidly.

– the fault detection system checks that the pressure in the hydraulic circuit is between 500 kPA and 2 MPa, and that the pressure changes no more than 100 kPa in 0.01 seconds.

4.4. Results and discussion

To assess our fault prognosis multi-agent system (developed using the Jade environment [15]), we have made 127 tests. Whenever we launch system simulation using Matlab Simulink library [10] (the agent simulator handles the connection between our tool and the model simulation in Simulink), then we run the prediction of the data mining agent and the knowledge management agent simultaneously. Data mining agent uses the Weka environment[19] to treat the stored data based and generate the decision tree. The resulted decision tree is given in annex (figure5). The generated rules are given in annex (figure6). The knowledge management agent uses the rules previously presented (section 4.3). Finally, the predictor agent combines the results generated by data mining agent (which use the data mining technique) and knowledge management agent (which will operate knowledge management technique) by attributing to each of them a 50% probability. The result is transmitted to the user interface system to help the user to make the appropriate decision as shown in figure 3. In this example, a message appears informing the user that there will

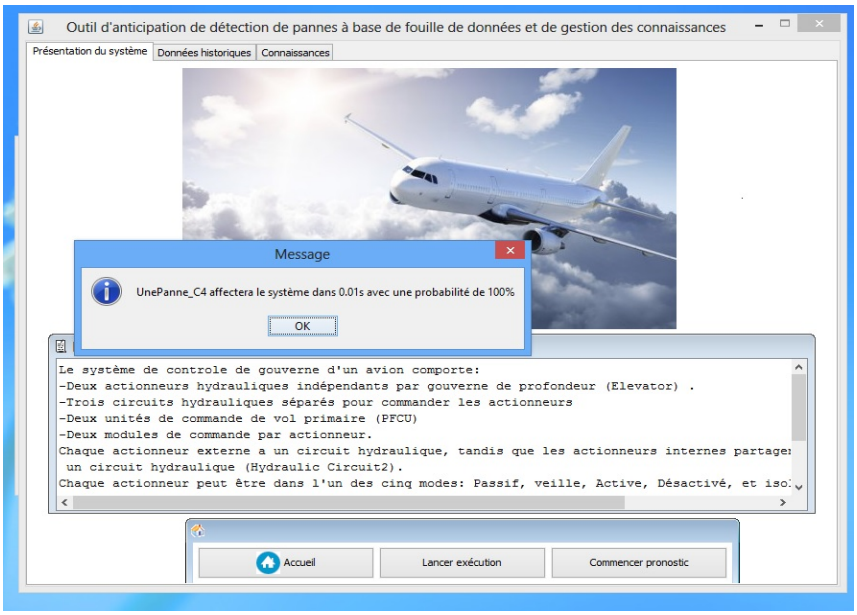


Figure 3. *Example of prognostic result*

be a breakdown in 0.01s with a probability of 100% (50% from prognostic result based on data mining and 50% prognostic result based on knowledge management). After the various tests of our multi-agent system, the results are very encouraging. Indeed, we have obtained the correct decision in 100% of tests with accuracy the expected timing and likelihood of occurrence of such faults is case of the example shown in the following figure. This can be explained by the combination of the results of data mining and knowledge management. Indeed, in our study the number of recorded data forms a very good basis for learning and the effectiveness of the used data mining technique. The step of acquiring knowledge also forms another crucial as a basis step for decision making.

5. Conclusion

The area of Intelligent Decision Support Systems is very interesting as it assist the decision maker to take the most appropriate decisions at the right time. In this context, we are particularly interested in intelligent prognosis which offers support to decision-maker in case of preventive maintenance. This paper proposed a multi-agent approach to predict faults that may appear in complex systems. This approach is based on the combination of data mining and knowledge management techniques. The simulation results of this approach for the case of the aircraft elevator control system are very encouraging. Future works aim to highlight the potential of such approach in real systems cases.

6. Bibliographie

[1] Jardine, A. K. , Lin, D. , & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical systems and signal proces-

sing, 20(7), p. 1483-1510.

[2] M. Lebold, M. Thurston (2001). Open standards for condition-based maintenance and prognostic systems, in :Maintenance and Reliability Conference (MARCON).

[3] D. A. Tobon-Mejia, K. Medjaher, N. Zerhouni, G. Tripot (2010). A mixture of Gaussians hidden Markov model for failure diagnostic and prognostic, in :IEEE Conference on Automation Science and Engineering, CASE10.

[4] P. Wang, G. Vachtsevanos(2001). Fault prognostics using dynamic wavelet neural networks, AI EDAM-Artificial Intelligence for Engineering Design Analysis and Manufacturing 15, p. 349-365.

[5] HU, Jinqiu, ZHANG, Laibin, MA, Lin, et al. (2011). An integrated safety prognosis model for complex system based on dynamic Bayesian network and ant colony algorithm. Expert Systems with Applications, vol. 38, no 3, p. 1431-1446.

[6] WIDODO, Achmad et YANG, Bo-Suk(2011). Machine health prognostics using survival probability and support vector machine. Expert Systems with Applications, vol. 38, no 7, p. 8430-8437.

[7] Quinlan, J. R. (2014). C4. 5 : programs for machine learning. Elsevier.

[8] Qureshi, S. , V. Hlupic, et R. O. Briggs (2004). On the convergence of knowledge management and groupware. In Groupware : Design, Implementation, and Use, p. 25-33. Springer.

[9] Ferber, J. et J. -F. Perrot (1995). Les systèmes multi-agents : vers une intelligence collective. InterEditions.

[10] The Mathworks, I. , Simulink. Simulation and Model-Based Design. http ://www. mathworks. com/products/simulink/.

[11] Jianhui Luo, Krishna R Pattipati, Liu Qiao et Shunsuke Chigusa(2008). Modelbased prognostic techniques applied to a suspension system. Systems, Man and Cybernetics, Part A : Systems and Humans, IEEE Transactions on, vol. 38, no. 5, p. 11561168.

[12] Douglas E Adams et Madhura Nataraju(2002). A nonlinear dynamical systems framework for structural diagnosis and prognosis. International Journal of Engineering Science, vol. 40, no. 17, p. 19191941.

[13] Otter, Martin(2002), Simulation for Analysis of Aircraft Elevator Feedback and Redundancy Control, dynamics (typically not more than three continuous state variables), vol. 5, no. 13, 27.

[14] Kan, Man Shan and Tan, Andy CC and Mathew, Joseph (2015), A review on prognostic techniques for non-stationary and non-linear rotating systems, Mechanical Systems and Signal Processing, vol. 62, p. 1-20, Elsevier.

[15] Bellifemine, Fabio Luigi and Caire, Giovanni and Greenwood, Dominic(2007), Developing multi-agent systems with JADE, vol. 7, John Wiley & Sons.

[16] El-Koujok, Mohamed and Gouriveau, Rafael and Zerhouni, Noureddine(2014), Development of a prognostic tool to perform reliability analysis, Proc. of the ESREL-17th SRA-Europe Conf. , Valencia, Spain, sept. 22, vol. 25, p. 191-199.

[17] Widodo, Achmad and Caesarendra, Wahyu(2014), SUMMARY OF THE RECENT DEVELOPED TECHNIQUES FOR MACHINE HEALTH PROGNOSTICS, ROTASI, vol. 16, no. 1, p. 21-27.

[18] Quinlan, J. Ross(1986), Induction of decision trees, Machine learning, vol. 1, no. 1, p. 81-106, Springer.

[19] Hall, Mark and Frank, Eibe and Holmes, Geoffrey and Pfahringer, Bernhard and Reutemann, Peter and Witten, Ian H (2009), The WEKA data mining software : an update, ACM SIGKDD explorations newsletter, vol. 11, no. 1, p. 10-18, ACM.

[20] Kothari, RAVI and Dong, MING(2001), Decision trees for classification : A review and some new results, Pattern Recognit, vol. 171, p. 169-184, World Scientific.

7. Annex

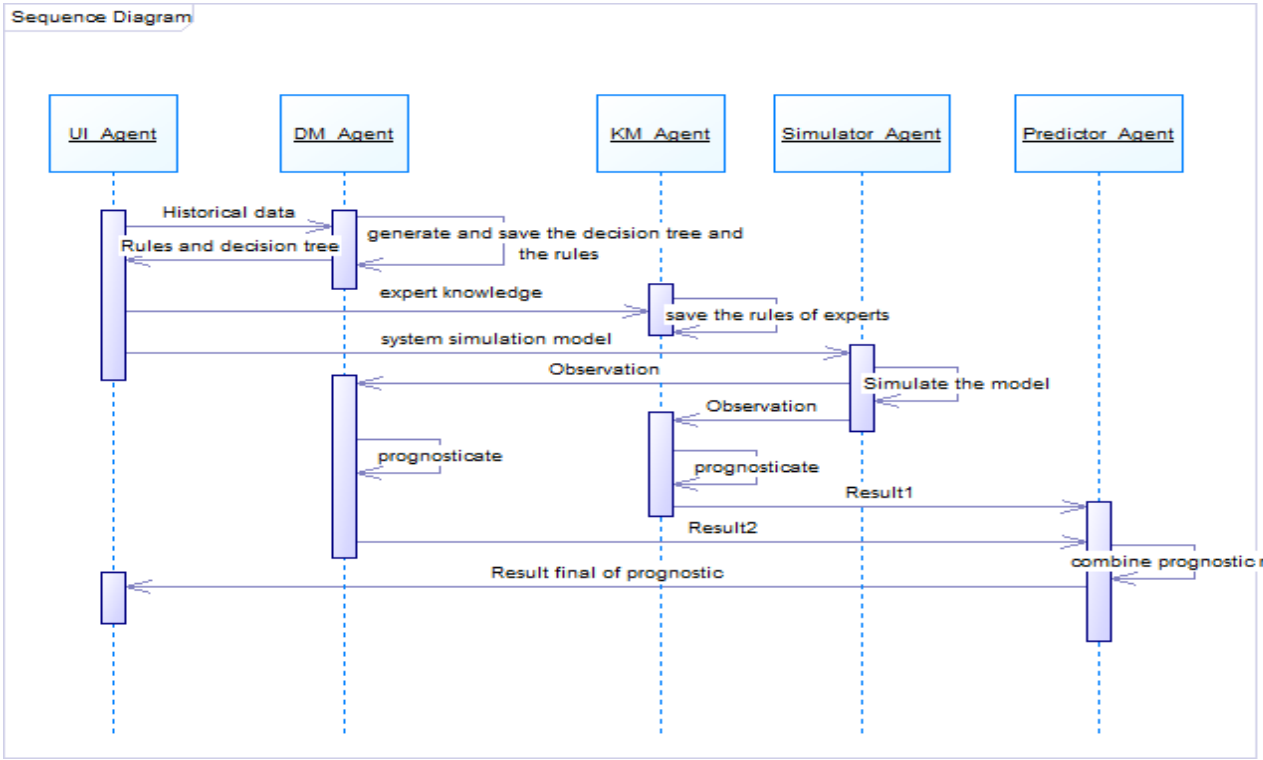


Figure 4. *sequence diagram*

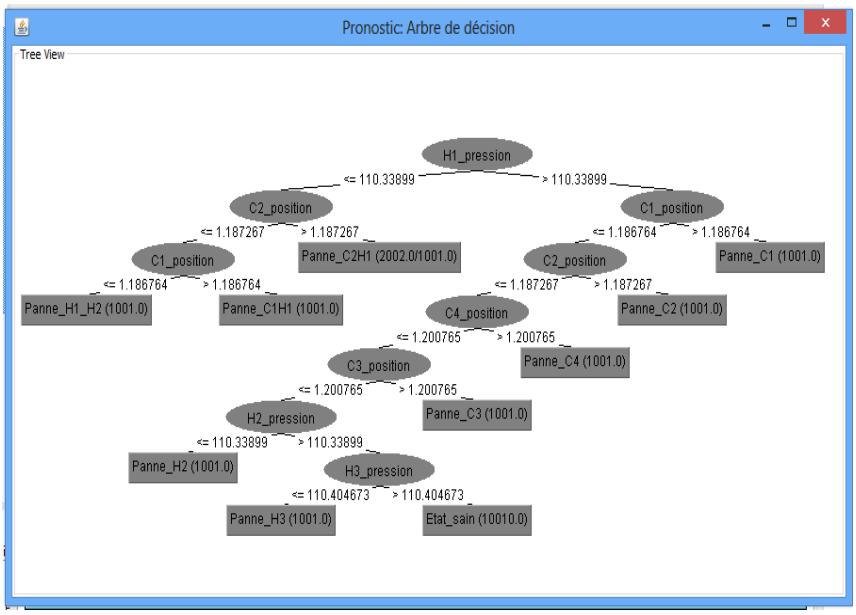


Figure 5. *Decision tree*

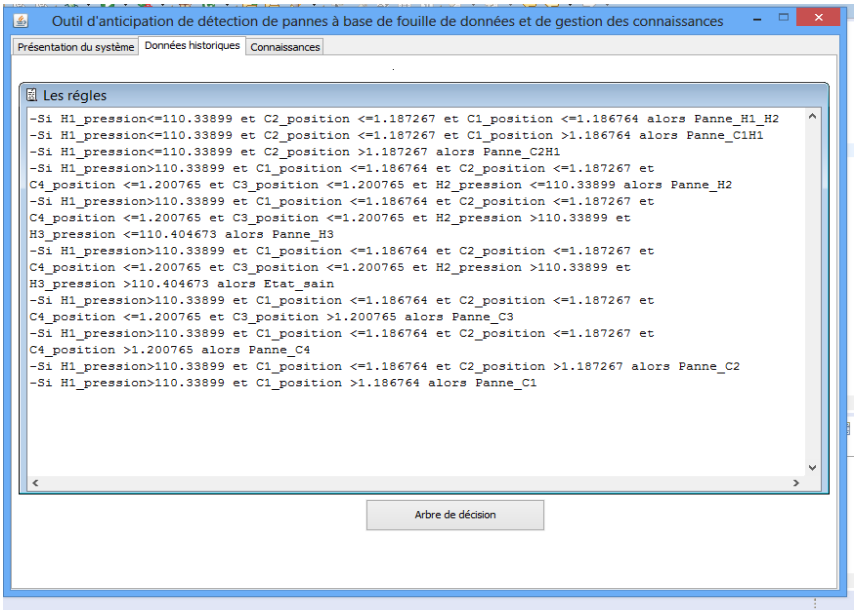


Figure 6. *Decision rules*

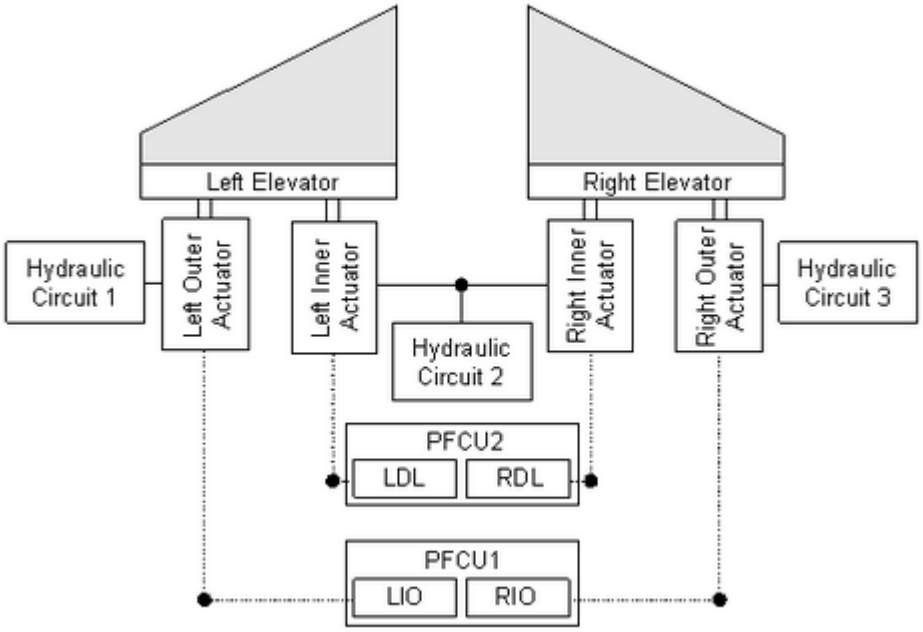


Figure 7. *The aircraft elevator control system[13]*

Amélioration de la visite de classe de l'enseignement technique : intégration d'un dispositif de médiation

Frédéric T. Ouédraogo[1], Daouda Sawadogo[2], Solange Traoré[3] et Olivier Tindano[1]

frederic.ouedraogo@univ-ouaga.bf , daouda.sawadogo@univ-lr.fr, tsolange54@yahoo.fr, tindanolivier83@gmail.com

[1] Ecole Normale Supérieure, Université de Koudougou Burkina Faso.

[2] Laboratoire L3i, Université de La Rochelle, France.

[3] Inspection de l'enseignement technique de Ouagadougou, Burkina Faso.

RÉSUMÉ. L'amélioration des pratiques de l'enseignant est portée par la visite de classe effectuée par le conseil pédagogique. Cet article présente l'intégration d'un système de partage de document comme un dispositif de médiation dans la visite de classe. Ce système de partage permet d'améliorer les interactions entre l'enseignant et le conseil pédagogique sur l'élaboration des principaux documents pédagogiques relatif à la visite de classe. L'expérimentation a été réalisée avec le service de partage de documents Google Drive sur une population constituée d'une quinzaine d'enseignants et de trois encadreurs pédagogiques de l'enseignement technique. Les résultats ont montré une amélioration de la qualité pédagogique de la visite de classe et une amélioration de la communication entre l'enseignant et l'encadreur pédagogique. Il ressort de cette étude que l'enseignant devient moins stressé et il sollicite beaucoup plus le conseil pédagogique que lorsqu'il n'utilise pas ce dispositif.

ABSTRACT. The improvement of teacher practice is supported by the class visit. This paper presents the integration of a sharing system of documents as a mediation system tool in the class visit. This sharing system enhances the interaction between the teacher and the pedagogical supervisor when they work on the main pedagogical documents of the class visit. We made an experiment with the service Google Drive with fifteen teachers and three pedagogical supervisors of the technical education. The results of this experiment have shown an improvement of the educational quality of the class visit and better communication between teacher and pedagogical supervisor. It appears from this study that the teacher becomes less stressed and he requests much more help of the pedagogical supervisor than usual.

MOTS-CLÉS : visite de classe, Encadrement pédagogique, Médiation, Enseignement, Google Drive.

KEYWORDS: class visit, pedagogical supervision, mediation, Teaching, Google Drive.

1. Introduction

L'éducation tient un rôle fondamental dans toutes les sociétés. Elle représente l'ensemble des activités visant à développer les potentialités physiques, intellectuelles, morales, psychologiques et sociales d'un humain. Elle assure sa socialisation, son autonomie, son épanouissement et sa participation au développement économique, social et culturel[1] de sa société. Elle constitue une priorité nationale pour le Burkina Faso. Elle se révèle être nécessaire pour l'individu et pour sa communauté. En dépit de son importance, la qualité de l'enseignement, aussi bien au niveau de l'éducation de base qu'au niveau de l'éducation post-primaire et secondaire, constitue un véritable problème pour l'ensemble des acteurs et des partenaires du système éducatif burkinabé. Cependant, cette recherche de la qualité de l'enseignement passe nécessairement par la qualité professionnelle des enseignants [1]–[4]. D'après l'auteur dans [5], la qualité de l'apprentissage est un facteur important pour la réussite du processus d'apprentissage. C'est pourquoi au cours de ces dernières décennies, les mouvements de réforme du système éducatif se sont appuyés sur la formation des enseignants en tant qu'élément essentiel pour modifier les pratiques pédagogiques et le développement d'outils d'aide à l'amélioration de la qualité de l'enseignement.

Cette préoccupation se traduit sur le terrain par plusieurs actions. Nous pouvons citer entre autres l'organisation des conférences annuelles sur l'enseignement [6] qui ont pour thème de réflexion, la recherche des meilleures stratégies et actions pour améliorer la qualité de l'enseignement. Cette initiative favorise une formation continue des enseignants sur le terrain par un encadrement pédagogique de proximité qui se traduit par des visites de classe. Ces visites ont pour objectif de donner un appui pédagogique adapté aux besoins de chaque enseignant afin d'améliorer sa pratique de classe.

Cependant, les visites de classe telles qu'elles sont organisées connaissent un certain nombre de difficultés. C'est pour ces raisons qu'elles n'arrivent toujours pas à développer la compétence escomptée chez les enseignants [6], [7]. Au regard des insuffisances constatées après plusieurs années de pratique de la visite de classe et avec l'avènement des Technologies de l'Information et de la Communication (TIC), nous nous proposons d'explorer une piste d'amélioration. Les Technologies de l'Information et de la Communication tant renommées dans diverses activités pourront-elles rehausser l'efficacité de la visite de classe ? Nous avons choisi d'expérimenter la plate-forme Google Drive dans la visite de classe en enseignement technique. Cette expérience a montré un apport andragogique dans la formation continue des enseignants.

[1] Article 2 de la LOI N°013-2007 / AN portant loi d'orientation de l'éducation au Burkina Faso.

Cet article est structuré de la manière suivante : le contexte et la problématique que nous abordons sont présentés dans le premier point ; le second est consacré à notre méthodologie pour améliorer la visite de classe ; le troisième présente le cadre expérimental et les résultats ; enfin le quatrième point présente une discussion sur notre contribution par rapport aux travaux antérieurs.

2. Contexte

Les encadreurs pédagogiques en service dans les inspections sont chargés principalement de l'encadrement des enseignants. Nous nous intéressons à l'encadrement pédagogique dont bénéficient les enseignants. Cet encadrement pédagogique a pour but de permettre aux enseignants d'améliorer leur pratique de classe. La visite de classe est une activité de l'encadrement pédagogique. Cette visite comporte deux volets : l'analyse du dossier d'intervention pédagogique et l'observation de la conduite de classe de l'enseignant suivie d'un entretien.

L'encadreur pédagogique apporte un appui qui répond aux besoins spécifiques de l'enseignant. L'enseignant devrait pouvoir profiter de la visite de classe comme moyen de renforcement de ses compétences. Mais nous constatons que l'organisation de la visite de classe dans l'enseignement technique ne permet pas à l'encadreur pédagogique de percevoir efficacement les compétences et les insuffisances de l'enseignant afin d'apporter une solution appropriée.

En général, l'encadreur pédagogique prend connaissance du dossier d'intervention pédagogique pendant la visite de classe, au moment où l'enseignant débute sa prestation, voir Figure 1. Pourtant, ce document a un volume qui varie entre vingt et quarante pages qu'il sera difficile à l'encadreur d'exploiter et de suivre la prestation de l'enseignant. Face à cette difficulté, l'encadreur est obligé de faire un choix cornélien :

- soit il privilégie la prestation de l'enseignant au détriment du dossier d'intervention pédagogique. Conséquence, l'entretien sera focalisé sur la prestation au détriment du document ;
- soit il se concentre sur le dossier d'intervention pédagogique. Conséquence, il sera moins attentif à l'observation de la prestation de l'enseignant. Alors, l'entretien sera focalisé sur le dossier d'intervention pédagogique ;
- soit il essaie de suivre les deux volets à la fois. Conséquence, aucun des volets ne sera traité efficacement.

Pourtant, c'est la synthèse de l'évaluation de ces deux parties majeures qui devrait lui permettre de mieux juger l'enseignant, afin de lui proposer des solutions appropriées. De ce qui précède, il ressort que l'enseignant ne profite pas suffisamment de l'expertise de l'encadreur pédagogique.

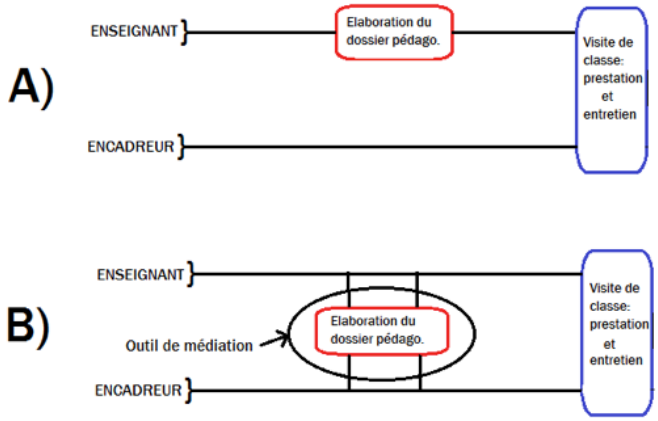


Figure 1 : Processus de la visite de classe avec et sans un dispositif de médiation

De plus, l'insuffisance du nombre d'encadreurs pédagogiques vient aggraver ce problème. Ceux-ci sont, non seulement fortement sollicités mais, interviennent même au-delà de leur spécialité.

Lors d'une visite de classe, l'encadreur pédagogique joue deux rôles. Il est conseiller et évaluateur. Dans le milieu des enseignants, ce rôle d'évaluateur est mal perçu, de sorte qu'il s'est développé le stéréotype de l'encadreur « gendarme ». Cette situation constitue un blocage à la collaboration enseignant-encadreur et ne favorise pas une interaction constructive dans la formation continue des enseignants.

Dans le processus actuel de la visite de classe, il n'existe pas un cadre de rencontre préparatoire où l'enseignant peut solliciter l'encadreur, comme le montre la Figure 1(a). Un tel cadre permettrait à l'enseignant de bénéficier de l'expertise de l'encadreur pour préparer son dossier d'intervention pédagogique avant la visite de classe. Dans le même temps, l'encadreur serait au même niveau que l'enseignant, car il aurait pris connaissance de son dossier d'intervention pédagogique. L'encadreur aurait alors suffisamment de temps pour se consacrer au second volet de la visite de classe. Ce volet concerne la prestation de l'enseignant. Par ailleurs, l'existence d'un tel cadre de médiation aurait le bénéfice d'améliorer la collaboration entre enseignants et encadreurs et ainsi briser les préjugés de l'encadreur « gendarme » entretenus au milieu des enseignants.

En résumé, les problèmes qui entravent l'efficacité de la visite de classe peuvent être classés à deux niveaux : le processus actuel de la visite de classe et le stéréotype sur l'encadreur pédagogique. Dans ce contexte, l'introduction d'un dispositif de

médiation dans le processus de la visite de classe peut être une solution pour améliorer celle-ci. Dans le présent article, nous proposons une solution axée sur les technologies de l'information et de la communication qui consiste à introduire un cadre de médiation dans le processus de la visite de classe.

3. Méthodologie pour améliorer la visite de classe

Dans cette partie de notre étude, nous présentons premièrement les fondements de l'encadrement pédagogique et de la visite de classe puis nous montrons notre proposition pour l'amélioration de la visite de classe.

3.1. Cadre conceptuel : encadrement pédagogique

Selon le dictionnaire actuel de l'éducation [8], l'encadrement regroupe les activités qui visent à fournir une aide aux apprenants pris individuellement ou en groupe. Cela vise à favoriser la prise en charge par chacun de sa propre formation. Encadrer une personne, c'est assurer auprès d'elle un rôle de directeur de formation. Le terme de direction sous-tend l'action de guider, d'exercer une influence, d'animer et surveiller, alors que celui de formation revêt le sens de développement, d'apprentissage. La pédagogie, selon l'APPAC[2], est l'art d'enseigner ou les méthodes d'enseignement propres à une discipline, à une matière, à un ordre d'enseignement, à un établissement d'enseignement ou à une philosophie de l'éducation.

L'encadrement pédagogique représente donc l'ensemble des actions posées par les encadreurs pédagogiques, qui ont pour but de conduire l'enseignant à acquérir des aptitudes pour bien enseigner en utilisant les meilleures méthodes d'enseignement et les contenus adaptés.

3.2. Cadre conceptuel : la visite de classe

La visite de classe se déroule en deux phases, comme les auteurs BOUVIER et OBIN la décrivent [9]. La première phase qui est l'observation et l'analyse d'une conduite de classe d'un enseignant par un encadreur pédagogique suivi de la seconde phase qui est un entretien entre l'encadreur pédagogique et l'enseignant.

Dans le contexte du Burkina Faso, la visite de classe est considérée comme une forme de formation continue adressée régulièrement aux enseignants en poste [7]. Elle est une pratique d'encadrement, de suivi et de contrôle en vue d'aider les enseignants à

[2] APPAC : Association Professionnelle des Professeurs et Professeures d'Administration au Collège.

améliorer leurs prestations pédagogiques. Lorsque de nouvelles pratiques sont adoptées, les visites de classe interviennent comme une formation de recyclage des enseignants.

3.3. Outil de médiation : Google Drive

Google Drive est un service de stockage et de partage de fichiers dans le *Cloud* qui permet le travail collaboratif. Ce service a été lancé par la société Google en avril 2012 [10]. La Figure 2 montre l'interface principale de Drive qui présente de manière analogue à Microsoft Office, une suite bureautique permettant d'éditer plusieurs formats de fichiers de type texte, tableur, présentation, etc. Il est également possible d'avoir une organisation hiérarchique des fichiers en dossiers ou d'importer des fichiers sur Drive.

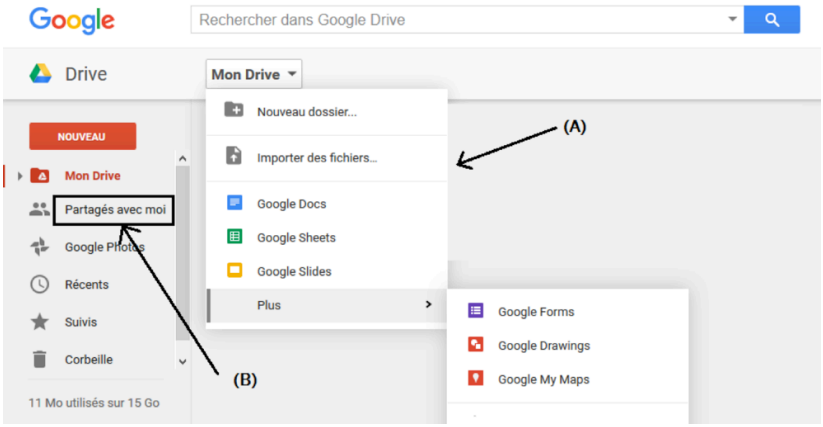


Figure 2 : Interface du service Drive de Google. A) Edition de plusieurs types de fichiers; B) Accès aux fichiers partagés avec moi.

Avec l'option partage, les utilisateurs peuvent partager avec possibilité de modification ou non. La possibilité de modification permettra à plusieurs utilisateurs d'éditer un même fichier. Les fichiers partagés peuvent être recherchés sur Google Drive par le biais de moteurs de recherche Web.

4. Cadre expérimental

Nous avons opté pour une approche mixte conjuguant l'approche qualitative et l'approche quantitative. Cela se justifie parce que notre étude nécessite plusieurs variables à la fois qualitatives et quantitatives. L'approche qualitative va nous permettre

de recueillir les avis de chaque membre de l'échantillon sur l'introduction d'un dispositif de médiation dans la visite de classe. De l'approche quantitative, on va considérer des résultats issus des données de l'échantillon dont on aura effectué un traitement statistique. Ces données nous permettent d'avoir l'appréciation de l'enseignant et de l'encadreur sur ce cadre de médiation dans la visite de classe.

4.1. Caractéristiques et choix du dispositif

Le dispositif de médiation de la visite de classe doit satisfaire un minium de caractéristiques qui répondent aux besoins suscités. Ces caractéristiques sont : partage de ressources ; communication ; accessibilité sur internet.

Les technologies de l'information et la communication offrent une gamme variée d'outils dans presque tous les domaines d'activité de l'homme. Il nous paraissait alors évident que le choix de l'outil pour le cadre de médiation devait tenir compte des potentialités offertes par les TIC pour être pertinent et aussi pour ne pas être en déphasage avec la tendance générale. Cependant, plusieurs outils TIC peuvent être utilisés comme outil de médiation dans la visite de classe, mais ce choix doit tenir compte d'autres facteurs tels que le coût financier, le niveau de complexité, le temps de mise en œuvre, etc. Il faut donc choisir un outil à moindre coût, facile à maîtriser par les différents acteurs et rapide à mettre en œuvre.

Nous avons fait une recherche comparative des différents outils TIC pouvant être utilisés et nous avons trouvé que les systèmes de partage de fichiers Drive de Google et OneDrive de Microsoft sont adaptés pour être un dispositif de médiation dans le processus de visite de classe. Nous avons donc choisi d'utiliser Google Drive au lieu de OneDrive car les services de Google sont plus connus dans l'environnement de l'expérimentation.

4.2. Cadre expérimental

Les différents documents du dossier d'intervention pédagogique est composé du dossier technique et du dossier pédagogique. Le dossier technique comporte le résumé du cours, la documentation technique permettant de réaliser les travaux pratiques et le mode opératoire. Le dossier pédagogique est constitué de cinq fiches : identification pédagogique, exploitation pédagogique, progression pédagogique, déroulement et évaluation du domaine cognitif et du psychomoteur. Chaque enseignant initie l'élaboration de son dossier d'intervention pédagogique sur Google Drive et le partage avec l'encadreur pédagogique ou le conseil pédagogique, quand il y a plusieurs encadreurs qui interviennent dans la visite de classe.

Les fiches sont remplies au fur et à mesure par l'enseignant, supervisé par l'encadreur, de telle sorte qu'à la fin, les différents acteurs ont connaissance de leurs

contenus. Dans le cadre de la présente étude, chaque enseignant nous donnera accès (en lecture seule) à son dossier d'intervention pédagogique. Ainsi, nous avons accès aux échanges entre enseignant et encadreur, afin de suivre le processus de préparation de la visite sur Google Drive. Quand un enseignant finit d'élaborer ces fiches sur la plateforme, il se prépare pour la visite en poursuivant les échanges avec son encadreur pédagogique. La visite de classe donnera l'occasion à l'encadreur pédagogique d'observer la pratique de l'enseignant.

4.3. Echantillon

L'expérimentation s'est effectuée sur la Région du Centre au Burkina Faso, plus précisément dans les communes de la capitale Ouagadougou. Elle a concerné des établissements publics et privés. Nous avons choisi cette région parce qu'elle regroupe beaucoup d'établissements techniques avec une offre variée dans les spécialités enseignées. D'après les statistiques de l'année 2012 du Ministère des enseignements, cette région possède la majorité des enseignants et encadreurs pédagogiques de l'enseignement technique du Burkina. Nous allons donc constituer notre échantillon dans cette région. Le Tableau 1 montre que la majorité des enseignants, qui constituent l'échantillon, n'est pas sans expérience. En effet, ils cumulent plus de dix ans d'ancienneté dans l'enseignement.

5. Évaluation et résultats

Une présentation de l'environnement Drive de Google a été faite aux différents participants de l'étude, que sont les enseignants et les encadreurs. La mise en œuvre de la visite de classe est faite selon un protocole qui régit les différentes interactions et les supports utilisés. La chronologie de l'expérimentation est comme suit :

1. l'enseignant prépare le dossier d'intervention pédagogique et le partage avec son encadreur pédagogique. Ce dossier est composé de fiches pédagogiques et de fiches techniques ;
2. l'encadreur pédagogique consulte le dossier pédagogique de l'enseignant et apporte des corrections et commentaires ;
3. l'enseignant intègre les corrections et continue les échanges avec son encadreur jusqu'à satisfaction de ce dernier ;
4. la visite de classe est effectuée et l'encadreur analyse la prestation de l'enseignant selon une grille d'observation. L'encadreur fait part à l'enseignant de son appréciation de la visite.

L'expérimentation a été réalisée avec quinze enseignants et trois encadreurs pédagogiques. Elle s'est déroulée de novembre 2013 à mars 2014, soit une période de cinq mois.

Expérience par tranche d'année	Inférieure à 5 ans	De 5 à 10 ans	De 10 à 15 ans	Supérieure à 15 ans
Nombre d'enseignants	1	6	5	3

Tableau 1 : Ancienneté professionnelle des enseignants.

La plupart des enseignants possède un diplôme professionnel de l'enseignement. Les autres enseignants exercent uniquement sur la base du diplôme universitaire. Ces derniers sont surtout dans les établissements privés d'enseignements.

L'expérimentation a connu une participation satisfaisante des enseignants. En effet, la fréquentation de la plate-forme témoigne de l'activité des enseignants pendant l'expérimentation. Plus de la moitié des enseignants se connectaient au moins deux fois par jour pendant la préparation des fiches pédagogiques et techniques de la visite de classe.

L'on constate que les enseignants n'ont pas l'habitude de solliciter l'aide d'encadreur pédagogique pour élaborer leurs fiches pédagogiques lors de la préparation de la visite de classe. Par exemple, aucun enseignant participant n'avait eu à solliciter l'aide d'un encadreur pédagogique avant l'expérimentation. Pourtant, ce n'est pas parce que le besoin n'existe pas. Pour preuve, presque tous les enseignants ont fait recours à l'aide d'un encadreur pendant notre étude, comme le montre le Tableau 2. Les encadreurs pédagogiques ont été sollicités par les enseignants pour un appui-conseil.

Nombre de sollicitations	1	2	3	4	5	6
Nombre d'enseignants	0	2	4	5	3	1

Tableau 2 : Répartition des enseignants par nombre de sollicitations de l'encadreur pédagogique dans l'élaboration du dossier d'intervention pédagogique.

Les participants, enseignants comme encadreurs, sont unanimes que Google Drive représente un cadre d'échanges approprié pour l'élaboration d'un dossier d'intervention pédagogique. Les services offerts sont faciles d'utilisation. Enfin, ils trouvent que cet outil favorise une bonne collaboration entre encadreurs pédagogiques et enseignants.

Les enseignants reconnaissent avoir entretenu des échanges plus relaxes et plus ouverts avec leurs encadreurs pédagogiques. En outre, pour l'ensemble des enseignants participants, l'environnement Google Drive maintient un bon climat de travail entre enseignants et encadreurs pédagogiques. Cela facilite le recours à un encadreur pédagogique pour un appui-conseil.

Notons cependant que tous les enseignants ne sont pas satisfaits des réponses données à leurs sollicitations. Mais ce fait est d'un autre ordre et n'est pas lié à l'usage de l'outil. Il est important de savoir que ce *feedback* de l'enseignant sur une sollicitation de l'encadreur n'est pas envisageable en présentiel, car les rapports entre ces deux corps de l'enseignement sont toujours empreints de l'autorité de l'encadreur. Tous ces éléments nous permettent de dire que les participants sont ouverts aux innovations dans le cadre de la visite de classe.

Pour les enseignants, l'utilisation du dispositif de médiation encourage le recours à l'encadreur pédagogique en cas de difficulté. Les encadreurs pédagogiques, pour leur part, ont une bonne appréciation du dispositif. Selon eux, le dispositif leur a permis de de mieux cerner les difficultés de l'enseignant dans l'exercice de son métier.

Nous considérons que l'augmentation des sollicitations avec l'introduction du dispositif de médiation montre que cet outil est efficace à plusieurs niveaux. Il est non seulement approprié à la visite de classe, mais aussi contribue à une meilleure communication entre les acteurs.

6. Discussions

Notre étude montre que l'utilisation d'un système de partage de documents comme dispositif de médiation entre un enseignant et son encadreur peut améliorer le processus de visite de classe. Ainsi les TIC peuvent-elles constituer un accélérateur vers un enseignement de qualité [11]. Les enseignants et les encadreurs pédagogiques ont favorablement apprécié l'intégration du dispositif de médiation dans le processus de la visite de classe. L'avis des encadreurs est motivé par le fait que le dispositif permet d'avoir un avis d'un autre encadreur sur le dossier technique avant la visite de classe. Rappelons qu'il est courant dans l'enseignement technique que l'encadreur pédagogique assure les visites de classe d'autres disciplines autres que la sienne.

Notons cependant que les enseignants ont soulevé des difficultés concernant les conditions de l'expérimentation qui n'était pas de notre ressort. Il s'agit de la connexion Internet qui était souvent médiocre et ne permettait pas l'utilisation du dispositif. Les enseignants et les encadreurs n'ont pas trouvé nécessaires une formation sur l'utilisation du dispositif de médiation. Cela montre que le choix de Google Drive pour sa simplicité est approprié.

L'amélioration de la pratique de l'enseignant a été abordée dans [2], [4], [12], [13] et [14]. Lebrun, dans son article [2] s'interroge sur les modèles actuels et propose l'élaboration de nouveaux usages et méthodes de formation des enseignants. Il aborde de façon générale l'amélioration de l'enseignement par la formation des enseignants aux TIC. Il aborde très peu l'importance des outils de médiation pour les enseignants. Ainsi notre approche vise-t-elle à offrir aux enseignants des outils de médiation de leurs activités. Nous avons appliqué cette approche dans le cas de la visite de classe. La visite de classe est un aspect très important dans l'amélioration de la qualité de la formation des enseignants.

Les travaux effectués par Lerouge [15] sur la visite de classe proposent une approche praxéologique afin d'améliorer la pratique de l'enseignant en se basant sur l'analyse de sa pratique spontanée lors des visites de classes. Cette approche permet d'améliorer la formation professionnelle de l'enseignant. Contrairement à notre proposition, cette approche ne tient pas compte des outils technologiques. Pourtant, les TIC constituent un facteur indéniable dans le succès de la formation des enseignants.

Une des particularités de notre approche est qu'elle propose un outil simple et facilement utilisable par tous les enseignants. Ainsi la simplicité du dispositif améliore-t-elle son appropriation par ses utilisateurs. En effet, la complexité d'un outil technologique peut entraver son appropriation.

En résumé, notre approche introduit une innovation multidimensionnelle. Elle améliore l'approche pédagogique des enseignants, contribue à une communication efficiente et consolide le rapport socio-professionnel entre enseignant et encadreur pédagogique.

7. Conclusion

Notre travail de recherche a permis de montrer que la pratique actuelle de la visite de classe a besoin d'être reformée. Les outils offerts par les TIC représentent une opportunité pour la réalisation de cette réforme. C'est dans ce contexte que nous avons proposé l'utilisation d'un dispositif technologique de médiation dans le processus de la visite de classe. Nous avons décrit les caractéristiques que doit avoir un tel dispositif, comme le partage de documents dans le processus de la visite de classe. Ainsi lors de notre expérimentation, nous avons proposé un dispositif basé sur le service Drive de Google comme outil de médiation dans la visite de classe.

L'analyse des résultats obtenus lors de l'expérimentation a montré une amélioration de plusieurs aspects de la visite de classe. Par exemple, l'encadreur a une meilleure appréciation des difficultés pédagogiques de l'enseignant. Elle a permis également d'améliorer le rapport socio-professionnel entre enseignant et encadreur. La facilité d'adoption de l'outil permet d'affirmer qu'un dispositif de médiation dans la visite de

classe est important. Ce travail pourrait relancer la revalorisation de la visite de classe dont les enseignants sont souvent très sceptiques.

Nous envisageons dans nos futurs travaux, améliorer les résultats de cette étude avec un échantillon plus important.

7. Références

[1] T. Karsenti, C. Raby, and S. Villeneuve, "Quelles compétences technopédagogiques pour les futurs enseignants du Québec," *Formation et pratiques d'enseignement en questions*, vol. 7, pp. 117–136, 2008.

[2] M. Lebrun, "La formation des enseignants universitaires aux TIC: allier pédagogie et innovation.," *Revue internationale des technologies en pédagogie universitaire*, vol. 1, no. 1, pp. 11–21, 2004.

[3] G. Sang, M. Valcke, J. van Braak, and J. Tondeur, "Student teachers' thinking processes and ICT integration: Predictors of prospective teaching behaviors with educational technology," *Computers & Education*, vol. 54, no. 1, pp. 103–112, 2010.

[4] L. Talbot, "Les recherches sur les pratiques enseignantes efficaces," *Questions Vives. Recherches en éducation*, vol. 6, no. 18, pp. 11–13, 2012.

[5] C. M. Stracke, "The future of learning innovations and learning quality. How do they fit together?," in *Proceedings of the European Conference LINQ 2012*, 2012.

[6] MESSRS, "Conférence annuelle de l'enseignement secondaire (CAES)," Ouagadougou, 2010.

[7] INSPECTION du Burkina Faso, "Actes de la 2e Conférence Annuelle des Inspecteurs de l'Enseignement Secondaire," Bobo-dioulasso, 2001.

[8] R. Legendre, "Dictionnaire actuel de l'éducation," 1988.

[9] A. Bouvier, L. M. Bélair, and J.-P. Obin, *La formation des enseignants sur le terrain*. Hachette éducation, 1998.

[10] Wikipédia, "Google Drive - Wikipédia," 2016. [Online]. Available: https://fr.wikipedia.org/wiki/Google_Drive. [Accessed: 09-Feb-2016].

[11] D. Peraya, J. Viens, and T. Karsenti, "Introduction: Formation des enseignants à l'intégration pédagogique des TIC: Esquisse historique des fondements, des recherches et des pratiques," *Revue des sciences de l'éducation*, vol. 28, no. 2, pp. 243–264, 2002.

[12] K. Assagaye, Agaissa; Achille, "Les Technologies de l'information et de la communication dans la formation continue des enseignants d'un lycée au Niger," *Frantice*, vol. 9, 2014.

[13] M. Altet, "L'analyse de pratiques une démarche de formation professionnalisante: Formes et dispositifs de la professionnalisation," *Recherche et formation*, no. 35, pp. 25–41, 2000.

[14] C. Depover, T. Karsenti, and V. Komis, *Enseigner avec les technologies: favoriser les apprentissages, développer des compétences*. PUQ, 2007.

[15] A. Lerouge, "Un dispositif innovant de conseil pédagogique: la visite de classe formative," *Tréma*, no. 20–21, pp. 55–78, 2003.

Efficient high order schemes for stiff ODEs in cardiac electrophysiology

C. Douanla Lontsi* , Y. Coudière, C. Pierre

INRIA Bordeaux Sud Ouest
Université de Bordeaux
* charlie.douanla-lontsi@inria.fr

ABSTRACT. In this work, we introduce some exponential high order schemes for stiff $ODEs$ coming from the models used in cardiac electrophysiology. We show in this context that despite the stiffness of the equations, the use of high order (order 3) stabilized schemes is beneficial in terms of cost with the possibility to use large time-steps as for implicit schemes. We describe a methodology and introduce some tools allowing to compare the numerical schemes used to solve ODEs. These tools and methodology are then used through the Beeler Reuter ionic model (BR) [2] to compare the exponential schemes with the classical explicit and implicit schemes at various orders. It comes from this comparisons a good alternative in terms of cost, accuracy and ease of implementation.

RÉSUMÉ. Dans ce travail, nous définissons des schémas d'ordre élevé de type exponentiels appliqués aux $EDOs$ à caractères raides provenant des modèles utilisés en électrophysiologie cardiaque. Nous montrons dans ce contexte que malgré la raideur des équations, le recours à l'ordre élevé (ordre 3) des schémas stabilisés est profitable en terme de coût avec une possibilité d'utiliser des pas de temps aussi grands que lors de l'utilisation des schémas implicites. Nous décrivons des outils de comparaisons et une méthodologie permettant de comparer les schémas numériques pour la résolution des EDOs. Cette méthodologie est ensuite utilisée à travers le modèle ionique Beeler Reuter (BR) [2] pour comparer en terme de coût et de précision les schémas exponentiels aux schémas classiques implicites et explicites à différents ordres. Il en ressort de cette comparaison l'apport d'une bonne alternative à la fois en terme de coût, de précision et de facilité d'implémentation.

KEYWORDS : Exponential schemes, stiff ordinary differential equations, high order schemes, cardiac electrophysiology

MOTS-CLÉS : Electrophysiologie cardiaque, schémas d'ordre élevé, équations aux dérivées ordinaires raides, schémas exponetiels

1. Introduction

The numerical resolution of stiff ordinary differential equations $ODEs$ is an issue encountered in many field of applied sciences. In cardiac electrophysiology, the model describing the electrical activity of the heart is a system of parabolic partial differential equations coupled with a system of $ODEs$ called *ionic models* that is highly nonlinear and exhibiting a stiff behavior, making their numerical resolution very challenging. The classical schemes have serious drawbacks to solve such $ODEs$. On one hand classical stable methods are implicit and lead to high computational cost associated with large time-steps due to nonlinear solvers, on the other hand explicit solvers require very small time steps also leading to high computational costs. Meanwhile current solvers in the field are usually based on order 1 or 2 schemes. In this paper we investigate the resort to a class of both explicit and stable schemes referred as «exponential methods » of high order as an alternative to solve cardiac electrophysiological problems. Namely we will consider the exponential Adams Bashforth (EAB) and the Rush Larsen (RL) techniques.
Let us consider the general initial value problem,

$$\frac{dy}{dt} = F(t,y) \;\; t \in (0,T], \quad y(0) = y_0 \in \mathbb{R}^N. \tag{1}$$

EAB and RL schemes take advantage of a splitting of the model function F into some linear part a and a nonlinear part b,

$$\frac{dy}{dt} = a(t,y)y + b(t,y), \quad y(0) = y_0 \in \mathbb{R}^N. \tag{2}$$

Notice that in (2), a is not the exact linear part of F (its differential) but, an approximation or a guess thereof. The EAB and RL are built from a transformation of (2) on each time discretization interval $[t_n, t_{n+1}]$ in the following form,

$$\frac{dy}{dt} = \alpha_n y + c_n(t,y), \quad y(0) = y_0 \in \mathbb{R}^N. \tag{3}$$

Where $\alpha_n \in \mathbb{R}^N$ is a stabilizer set at every time step and $c_n(t,y) = (a(t,y) - a_n)y + b(t,y)$. With formulation (3), the exact solution satisfies the variation of the constant formula,

$$y(t_{n+1}) = \mathrm{e}^{\alpha_n h}\left(y(t_n) + \int_{t_n}^{t_{n+1}} \mathrm{e}^{-\alpha_n(\tau - t_n)} c_n(\tau, y(\tau))d\tau\right). \tag{4}$$

The aim of this paper is to study the efficiency of EAB and RL methods of order 1 up to 4. This efficiency is analyzed both in terms of accuracy and of cost. The comparison is made using a realistic test case and is completed by including a benchmark with several classical methods either of implicit or explicit type, the Crank-Nicolson (CN), the Runge Kutta (RK_4), the Adams Bashforth (AB_k), and the backward differentiation (BDF_k) (see [3]).

The paper is organized as follows. In section 2 are presented the stabilized schemes. A brief description of the transmembrane action potential and ionic model is given in the section 3. The methodology used to compared the methods are developed in the section 4. The comparison of the methods follows in the section 5 where the methodology defined is used to compare the numerical schemes applied to the Beeler Reuter (BR) ionic model [2].

2. EAB_k and RL_k scheme statements

When the function $c_n(t,y)$ in (3) is a polynomial $P_n = \sum_{j=0}^{k-1} p_j(t-t_n)^j$, the relation (4) becomes $y_{n+1} = \mathrm{e}^{\alpha_n h} y_n + h \sum_{j=0}^{k-1} p_j j! h^j \varphi_{j+1}(\alpha_n h)$, with $\varphi_0(z) = \mathrm{e}^z$ and $\varphi_j(0) = \frac{1}{j!},\ j \geq 0$. The schemes introduced in the sequel are multi-steps. We will use notation $a_n = a(t_n, y_n)$, $b_n = b(t_n, y_n)$.

– EAB_k : On one hand we set $\alpha_n = a_n$, on other hand the function c_n in (3) is approximated by its Lagrange interpolation polynomial $\tilde{c}_n$ of degree $k-1$ at the time instants $t_n, \ldots, t_{n-k+1}$. This polynomial satisfies $\tilde{c}_n(t_{n-j}) = c_n(t_{n-j}, y_{n-j})$ for $j = 0, \ldots, k-1$. The values $c_n(t_{n-j}, y_{n-j})$ for $j = 0, \ldots, k-1$ are given by $c_n^{n-j} = b_{n-j} + (a_{n-j} - a_n) y_{n-j}$. If we write $\tilde{c}_n(t) = \sum_{j=0}^{k-1} \frac{\gamma_{nj}}{j!} \left(\frac{t-t_n}{h}\right)^j$, the definition of the EAB_k scheme is deduced from the formula (4) by

$$y_{n+1} = \mathrm{e}^{a_n h} y_n + h \sum_{j=0}^{k-1} \gamma_{nj} \varphi_{j+1}(a_n h), \tag{5}$$

where the coefficients γ_{nj} are given in the table bellow.

k	1	2	3	4
γ_{n0}	c_n^n	c_n^n	c_n^n	c_n^n
γ_{n1}	–	$c_n^n - c_n^{n-1}$	$\frac{3}{2}c_n^n - 2c_n^{n-1} + \frac{1}{2}c_n^{n-2}$	$\frac{11}{6}c_n^n - 3c_n^{n-1} + \frac{3}{2}c_n^{n-2} - \frac{1}{3}c_n^{n-3}$
γ_{n2}	–	–	$c_n^n - 2c_n^{n-1} + c_n^{n-2}$	$2c_n^n - 5c_n^{n-1} + 4c_n^{n-2} - c_n^{n-3}$
γ_{n3}	–	–	–	$c_n^n - 3c_n^{n-1} + 3c_n^{n-2} - c_n^{n-3}$

Table 1. *Coefficients γ_{nj} for the EAB_k schemes.*

– RL_k : In the case the function $c_n(t,y)$ in (4) is a constant $c_n = \beta_n \in \mathbb{R}$ then we have the following simple scheme definition,

$$y_{n+1} = y_n + h\varphi_1(\alpha_n h)(\alpha_n y_n + \beta_n), \tag{6}$$

that we refer as Rush Larsen schemes as in the continuity of [1]. The following choices for defining α_n and β_n ensure the convergence at order k of the scheme (6) and thus are named Rush Larsen schemes of order k (RL_k).

- $k = 1 : \alpha_n = a_n \quad , \quad \beta_n = b_n$.
- $k = 2 : \alpha_n = \frac{3}{2}a_n - \frac{1}{2}a_{n-1}$ and $\beta_n = \frac{3}{2}b_n - \frac{1}{2}b_{n-1}$.
- $k = 3 : \alpha_n = \frac{1}{12}(23a_n - 16a_{n-1} + 5a_{n-2})$,
  $\beta_n = \frac{1}{12}(23b_n - 16b_{n-1} + 5b_{n-2}) + \frac{h}{12}(a_n b_{n-1} - a_{n-1}b_n)$.
- $k = 4 : \alpha_n = \frac{1}{24}(55a_n - 59a_{n-1} + 37a_{n-2} - 9a_{n-3})$,

$\beta_n = \frac{1}{24}(55b_n - 59b_{n-1} + 37b_{n-2} - 9b_{n-3}) + \frac{h}{12}(a_n(3b_{n-1} - b_{n-2}) - (3a_{n-1} - a_{n-2})b_n)$.

Notice that the EAB_1 scheme is the same with RL_1 scheme and also the exponential Euler scheme.

The previous descriptions of the EAB_k and RL_k schemes here have been given very briefly but, more details for the EAB_k schemes can be found in [5] (for general $ODEs$) and in [4] for cardiac electrophysiology application, where the two methods are shown to be stable under perturbations and convergent at the order k.

3. Modeling in cellular cardiac electrophysiology

3.1. The action potential

The phenomenon studied here is the so called *cellular action potential*, that we briefly present here. A potential difference is observed between the inside and outside of the cell, said membrane potential and denoted V. This potential caused by the differences in ion concentrations between the inside and outside of the cells is dynamic in time, as well as these ionic concentrations. The potential V can abruptly switch from a *resting* state (during which $V = V_r \simeq -100mV$) to an *excited* state (where V is in the range of 10 mV) in which it is maintained during a few tenth of seconds before returning to its resting state (see figure 1).

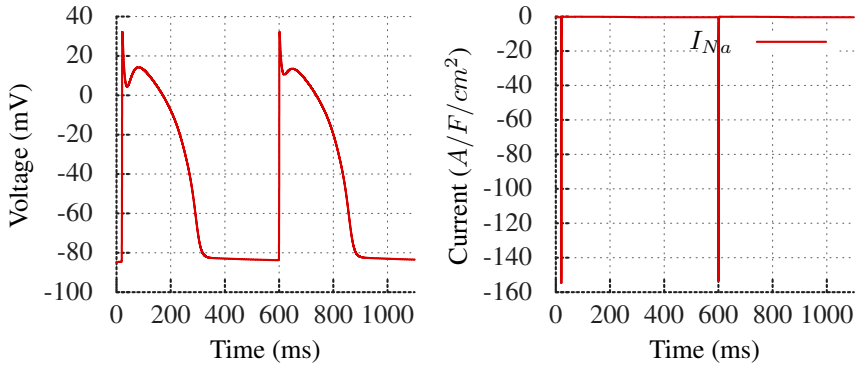


Figure 1. *BR Model [2] illustration. Left, two cellular action potentials : starting at a negative resting value, the transmembrane voltage $V(t)$ has a stiff depolarization followed by a plateau and repolarizing to the resting value. Right : each depolarization is induced by an ionic sodium current $I_{Na}(t)$*

3.2. Ionic Models

The variations of the ionic concentrations are described by ionic models and are systems of ODE. The innovated ones are consisting of the following variables with their associated ordinary differential equation.

– **The membrane potential:** V in mV. The equation on the potential is written,

$$\frac{dV}{dt} = -I_{ion}(y(t), I_{st}(t)) + I_{st}(t), \tag{7}$$

where I_{ion} (reaction term) is the total ionic current crossing the membrane cell and I_{st} is the stimulation current, it is a source term.

– **The gating variables:** they are parameters between 0 and 1 expressing the variability and the permeability of the membrane cell for the specific ionic species. One denote by $W \in \mathbb{R}^P$ the vector of gating variables. The equations on W are, for $i = 1 \dots P$,

$$\frac{dW_i}{dt} = \frac{W_{\infty,i}(y) - W_i}{\tau_i(y)}, \tag{8}$$

where $W_{\infty,i}(y) \in \mathbb{R}$, $\tau_i(y) \in \mathbb{R}$ are scalar functions given by the model. In these equations the linear and nonlinear parts are encoded in the model and are equal to $-1/\tau_i(y)$ and $W_{\infty,i}(y)/\tau_i(y)$ respectively.

– **Ionic concentrations:** One denote by $C \in \mathbb{R}^{N-P-1}$ the vector of concentrations. All these previous variables can be collected in a vector $y \in \mathbb{R}^N$ as follows ,

$$y = \begin{bmatrix} W \\ X \end{bmatrix}, \quad X = \begin{bmatrix} C \\ V \end{bmatrix}, \quad W \in \mathbb{R}^P, \quad C \in \mathbb{R}^{N-P-1}, \quad V = y_N \in \mathbb{R},$$

The sub-vectors W and X correspond to the lines of (1) for which the linear and nonlinear part is given or null respectively. The associated ODE written in the form (2) is then defined by,

$$a(t,y) = \begin{bmatrix} A_1(t,y) & 0 \\ 0 & 0 \end{bmatrix}, \quad b(t,y) = \begin{bmatrix} B_1(t,y) \\ B_2(t,y) \end{bmatrix},$$

where the matrix $A_1(t,y) \in \mathbb{R}^P \times \mathbb{R}^P$ is diagonal, $A_1(t,y) = \text{Diag}(-1/\tau_i(y))$, and $B_1(t,y) = \{W_{\infty,i}(y)/\tau_i(y), i = 1 \dots P\} \in \mathbb{R}^P$.

4. Scheme analysis methods

– **Test case :** The evaluation and comparisons between different ODE solvers is done with a test case. Specifically, the Beeler Reuter [2] model is considered and written in the form (2) as described in the section 3.2. We denote by $y(t)$ the solution of the associated ODE (2) in $(0,T]$ with $T = 396\ ms.$ this solution is uniquely defined once the initial condition y_0 and the stimulation current I_{st} in (7) are fixed. y_0 is the resting state as described by the model. The function $I_{st}(t)$ is positive, null outside the interval $(t_s - 1, t_s + 1)$, t_s=20 ms and with integral $\int_0^T I_{st}(t)dt = I_{stim}$, a typical current of stimulation fixed by the models, in the range of 50 mA. We also impose to I_{st} a C^4 regularity in oder to observe the convergence orders of schemes up to 4.

– **Numerical solution:** Let $m \geq 1$ be an integer for which one associated the time-step $h = T/m$ and the regular mesh $\mathcal{T}_m = \{t_n = jh,\ j = 0 \dots m\}$ of the interval $(0,T]$. The numerical solution (y^n) is the element of the space E_m, $E_m = \{(y^n)_{0 \leq n \leq m},\ y^n \in \mathbb{R}^N\}$. The space E_m of the numerical solutions is simply $(\mathbb{R}^N)^m$ but to $(y^n) \in E_m$ is implicitly associated a time-step dt and a mesh $\mathcal{T}_m$, such that each value y^n, $0 \leq n \leq m$ of $(y^n) \in E_m$ is supposed to be an approximation of $y(t_n)$.

– **Reference solution:** For a test case given, we cannot access to the exact solution $y(t)$ of the associated ODE. So for a numerical solution $(y^n) \in E_m$, we set $m' = 2^r m$ with $r \geq 0$ an integer and define the reference solution associated to (y^n) (or m) as the numerical solution $y_{ref} \in E_{m'}$ for the problem (1), computed by the RK_4 scheme with the time-step $h_{ref} = T/m' = h/2^r$. The reference solution y_{ref} is then not unique and depend on r. In practice r is chosen *large enough* such that the error between the exact solution y and y_{ref} is negligible compared to the error between the numerical solution (y^n) and y_{ref}.

– **Interpolation of the solution:** To compare the numerical solution with the reference solution and to be able to compute the numerical error in terms of function norm, we define an interpolator $\pi_{m,i} : E_m \longrightarrow C^0(0,T]$, transforming the component i of the numerical solution $(y^n) \in E_m$ in $C^0(0,T]$, the set of the continuous functions on $(0,T]$. Otherwise, we require to the interpolate $\pi_{m,i} y^n$ to be a polynomial piecewise function

of degree 3, this constraint is necessary to observe the convergence order up to 4. We assume that m is a multiple of 3 and fix $(y^n) \in E_m$. We decompose the interval $[0,T]$ in a sequence of 3 intervals packages $P_s = [t_{3s}, t_{3s+1}] \cup [t_{3s+1}, t_{3s+2}] \cup [t_{3s+2}, t_{3(s+1)}]$, for $s = 0 \dots m/3$. The interpolated $f := \pi_{m,i} y^n$ is the unique polynomial of degree 3 on each P_s, continuous on $[0,T]$, such that $f(t_n) = y_i^n$ for all $n = 0 \dots m$. This interpolator is not Canonical: an H^3-Hermite interpolation on each interval (t_n, t_{n+1}) is an alternative. The emphasis will be here on the membrane potential $V(t) = y_N(t)$ and for more simplicity we note $\pi_m = \pi_{m,N}$ and $\pi = \pi_{m,N}$ in confusion absence.

– **Accuracy:** Let (y^n) be a numerical solution and y_{ref} a reference solution associated. We denote $\pi y^n = \widehat{V}$ and $\pi Y_{ref} = \widehat{V}_{ref}$ the membrane potential interpolating associated. The accuracy of each method is evaluated through a relative error between the reference solution and the numerical solution. We define the errors in norm L^∞ by :

$$e_\infty = \frac{\max |\widehat{V} - \widehat{V}_{ref}|}{\max |\widehat{V}_{ref}|}. \tag{9}$$

Notice that the choice of the membrane potential V is arbitrary and that any other component of (y^n) could have been considered. The accuracy notion will be central here and it is convenient to identify several aspects.

– **Cost:** The accuracy takes all its meaning when one associate it a cost. Here it is a *computational* cost and is evaluated with the CPU time during a simulation. It is evaluated by the fortran90 software for each simulation setting by a time-step h. The CPU times depend on the computer used to perform the solutions. This is balanced by using the ratio between them for comparisons.

5. Numerical results

5.1. Accuracy

The relative error $e(h)$ is computed for various time-steps h and collected in the table 2 where it can be observed that all the methods exhibit the expected order of convergence. A general view of the table 2 shows that the RL_k is always more accurate than EAB_k and unlike the classical explicit schemes, the stabilized schemes allows the use of large time-steps as the implicit except at the order four where it is not possible for $h = 0.2$.

The table 2(a) shows that the CN is the most accurate among the methods of order 2 with a factor in the range of 10. The table 2(b) shows that the BDF_3 method is better than the stabilized schemes for $h \geq 0.0125$ with a coefficient 10 for $h = 0.2$ while for $h < 6.25 \times 10^{-3}$ the RL_3 is more accurate. The table 2(c) shows that the RK_4 method is the most accurate among the methods of order 4 for $h \leq 0.025$ while for $h > 0.025$ the BDF_4 is more accurate than the stabilized schemes.

5.2. Cost

A general observation of the figure 2 on the top shows that for the error between 1% and 10% the gain in terms of CPU time is high (with a factor in the range of 10) when moving from the order 1 to order 2 schemes. This gain remains important (with a factor in the range of 5) when moving from the order 2 to order 3 schemes while for the errors between 1% and 10% there is no gain when moving from the order 3 to the order 4 schemes. However the order 4 becomes advantageous for the errors less than 0.1%.

(a) AB_2, RL_2, EAB_2 and CN

h	AB_2	RL_2	EAB_2	CN
0.2	–	0.251	0.284	4.11×10^{-2}
0.1	–	0.107	9.26×10^{-2}	1.13×10^{-2}
0.05	–	3.35×10^{-2}	2.31×10^{-2}	2.65×10^{-3}
0.025	–	8.88×10^{-3}	5.39×10^{-3}	6.66×10^{-3}
0.0125	–	2.23×10^{-3}	1.29×10^{-3}	1.68×10^{-4}
6.25×10^{-3}	2.07×10^{-4}	5.6×10^{-4}	3.17×10^{-4}	4.25×10^{-5}

(b) AB_3, RL_3, EAB_3 and BDF_3

h	AB_3	RL_3	EAB_3	BDF_3
0.2	–	0.148	0.516	4.09×10^{-2}
0.1	–	4.07×10^{-2}	9.17×10^{-2}	1.04×10^{-2}
0.05	–	6.34×10^{-3}	1.09×10^{-2}	2.29×10^{-3}
0.025	–	7.57×10^{-4}	1.17×10^{-3}	3.84×10^{-4}
0.0125	–	9.07×10^{-5}	1.4×10^{-4}	5.25×10^{-5}
6.25×10^{-3}	1.13×10^{-5}	8.23×10^{-6}	1.72×10^{-5}	2.01×10^{-5}

(c) RK_4, RL_4, EAB_4 and BDF_4

h	RK_4	RL_4	EAB_4	BDF_4
0.2	–	–	–	4.98×10^{-2}
0.1	–	5.86×10^{-2}	0.119	1.27×10^{-2}
0.05	–	4.58×10^{-3}	8.96×10^{-3}	2.02×10^{-3}
0.025	4.65×10^{-5}	2.61×10^{-4}	4.33×10^{-4}	1.93×10^{-4}
0.0125	2.67×10^{-6}	1.62×10^{-5}	2.67×10^{-5}	3.52×10^{-5}
6.25×10^{-3}	1.65×10^{-7}	9.94×10^{-7}	1.73×10^{-6}	2.01×10^{-5}

Table 2. *Accuracy for the BR model for various classical and stabilized methods.*

The figure 2 on the bottom shows that the RL_3 and the RL_4 are less costly than the EAB_3 and EAB_4 respectively. The factor is not so high but in terms of implementation, the RL is easier than the EAB schemes.

The figure 2 on the bottom left shows that when using high order stabilized schemes instead of implicit schemes, the gain in time CPU is very high with a coefficient greater than 10. This is due to the fact that the nonlinear solver is very expensive and its cost become very high for large time-steps.

The figure 2 on the bottom right shows that the order 4 stabilized schemes are less costly than the classical explicit schemes but it is much more better to use the RL_4 scheme instead of the EAB_4 scheme. Because of they stability properties the explicit schemes require the use of small time-steps that make them sometimes useless. For instance the RK_4 is very accurate but its use require to take a small time step. This small time steps produces a very small error that might be not needed and then its use will induce an additional cost.

6. Conclusion

Two families of explicit high order stabilized methods (EAB_k, RL_k) have been introduced in this work. Excepted the order four, both have been shown to be as stable as the classical implicit methods for the test case we have chosen. Otherwise, it has also been

demonstrated that the use of high order (3 or 4) of the stabilized methods instead of the classical high order implicit methods allows to decrease the cost almost 50 times.

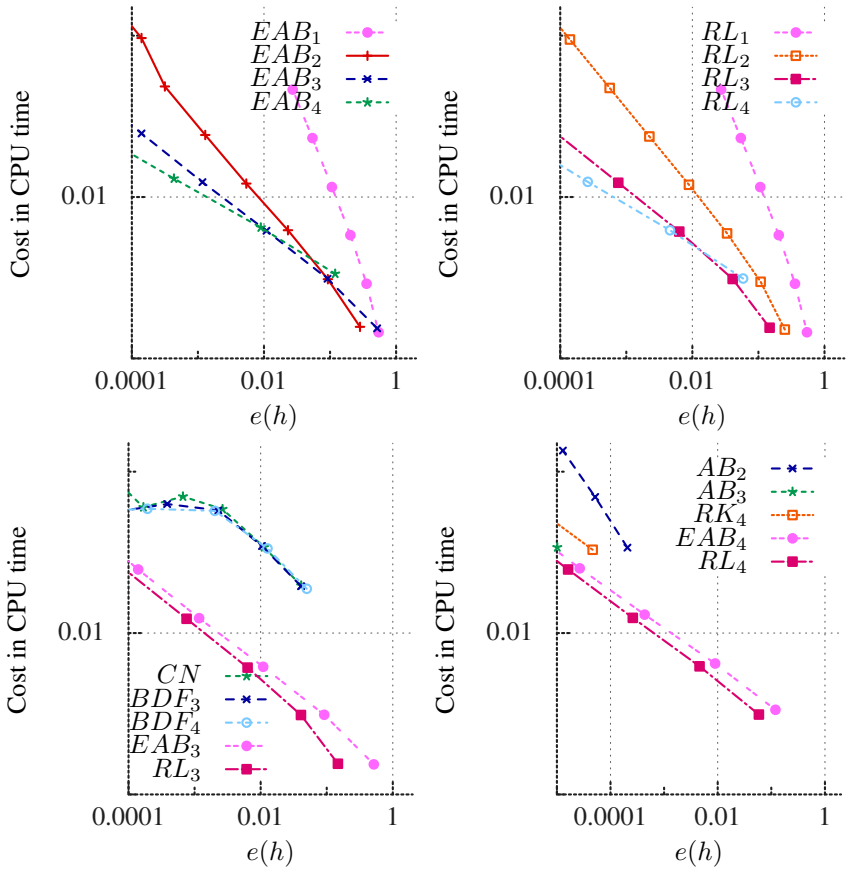


Figure 2. *The CPU time plotted in Log/Log scale against the error for various schemes*

7. References

[1] M. PEREGO, A.VENEZIANI, "An efficent generalization of the Rush-Larsen method for solving electro-physiology membrane equations", *ETNA*, vol. 35, 2009.

[2] G.W. BEELLER, H. REUTER, "Reconstruction of the Action Potential of Ventricular Myocardial Fibres", *J. Physiol*, vol. 268, 1977.

[3] E. HAIRER, S.P. NORSETT, G. WANNER, "Solving ordinary differential equations I", *Springer-Verlag, Berlin*, vol. 8, 1993.

[4] Y. COUDIÉRE , C. DOUANLA LONTSI, C. PIERRE, "High order Rush Larsen solver for stiff ODEs", *Hal*, 2016.

[5] M. HOCHBRUCK, A. OSTERMANN, "Exponential multistep methods of Adams-type", *BIT*, vol. 51, 2011.

ARIMA

A model of flocculation in the chemostat

R. Fekih-Salem [a,c,*] — T. Sari [b,d]

[a] Université de Tunis El Manar, École Nationale d'Ingénieurs de Tunis, LAMSIN,
B.P. 37, Le Belvédère, 1002 Tunis, Tunisie.
(E-mail: radhouene.fekihsalem@isima.rnu.tn)

[b] IRSTEA, UMR Itap,
361 rue Jean-François Breton, 34196 Montpellier, France.
(E-mail: tewfik.sari@irstea.fr)

[c] Université de Monastir, ISIMa,
BP 49, Av Habib Bourguiba, 5111 Mahdia, Tunisie.

[d] Université de Haute Alsace, LMIA,
4 rue des frères Lumière, 68093 Mulhouse, France.

[*] Corresponding author.

ABSTRACT. In this work, we study a flocculation model with a single resource and a single species which is present in two forms: isolated bacteria and attached bacteria. With monotonic growth rates and distinct removal rates, we show that this model presents a rich behavior with multiplicity of positive equilibria and bi-stability. Whereas, this bi-stability could occur in the classical chemostat model only with a non-monotonic growth rate.

RÉSUMÉ. Dans ce travail, nous étudions un modèle de floculation avec une seule ressource et une seule espèce qui se présente sous deux formes: des bactéries isolées et en flocs. Avec des taux de croissance monotones et des taux de prélèvement distincts, nous montrons que ce modèle présente un comportement très riche avec multiplicité des équilibres positifs et bi-stabilité. Cependant, dans le modèle classique du chémostat, cette bi-stabilité ne peut se produire qu'avec un taux de croissance non monotone.

KEYWORDS : Bifurcation, Bi-stability, Chemostat, Flocculation

MOTS-CLÉS : Bifurcation, Bi-stabilité, Chémostat, Floculation

1. Introduction

Flocculation is a process wherein microorganisms isolated or planktonic bacteria cluster together to form a flocs and reversibly this flocs can split and liberate isolated bacteria [10]. The attachment of planktonic bacteria could be also on a wall as biofilms [1]. This flocculation mechanism can explain the coexistence between species when the most competitive species inhibits its growth by the formation of flocs [3, 6]. Indeed, the flocs consume less substrate than isolated bacteria since they have less access to the substrate, given that this access to the substrate is proportional to the outside surface of the floc.

In order to understand and predict these flocculation phenomena, several extensions of the well-known chemostat model [9] have been proposed and studied in the literature by considering two compartments of isolated and attached biomass for each species [3]. For instance, Pilyugin and Waltman [8] have treated a model of wall growth where the attachment and detachment rates are constant, and the population on the wall does not wash out of the chemostat. The Freter model [5] describes a microbial population constituted of planktonic cells in the fluid and adherent cells on the surface. Their model was studied by Jones et al. [7], assuming that the planktonic bacteria are attracted to the wall at a rate proportional to planktonic cell density and the fraction of unoccupied colonization sites on the wall. More recently, the competition model of two species for a single substrate has been studied by Haegeman and Rapaport [6], assuming that only the most competitive species inhibits its growth by the formation of flocs. The study of [6] has been extended by [2] and [4].

In this paper, we study the model of flocculation considered in [3], where the isolated bacteria can stick with isolated bacteria or flocs to form new flocs. We do not assume as in [3] that attachment and detachment dynamics are fast compared to the growth of bacteria. The paper is organized as follows. In Section 2, we present the model of flocculation proposed in [3]. In Section 3, we study the existence and the local stability of the equilibria of system (1) for non-negative attachment and detachment rates. In Section 4, numerical simulations are presented with realistic growth functions (of Monod type) and the conclusion is drawn in the last Section 5. Most of the proofs are reported in the Appendix A.

2. Mathematical model

In this paper, we consider the model of flocculation proposed in [3]

$$\left\{\begin{array}{rcl} \dot{S} &=& D(S_{in}-S)-\mu_u(S)u-\mu_v(S)v \\ \dot{u} &=& [\mu_u(S)-D_u]u-a(u+v)u+bv \\ \dot{v} &=& [\mu_v(S)-D_v]v+a(u+v)u-bv \end{array}\right. \tag{1}$$

where $S(t)$ denotes the concentration of the substrate at time t; $u(t)$ and $v(t)$ denote, respectively, the concentration of planktonic and attached bacteria at time t; $\mu_u(S)$ and $\mu_v(S)$ represent, respectively, the per-capita growth rates of planktonic and attached bacteria; S_{in} and D denote, respectively, the concentration of the substrate in the feed device and the dilution rate of the chemostat; D_u and D_v represent, respectively, the removal rates of planktonic and attached bacteria such that $D_v \leqslant D_u \leqslant D$.

We assume that the planktonic bacteria can stick with the isolated bacteria or the flocs to form a new flocs, with rate $a(u+v)$, where a is a non-negative constant, and that

the flocs can split and liberate an isolated bacteria, with rate b, where b is a non-negative constant. We add the following assumptions on the growth rates:

H1: The functions $\mu_u(\cdot)$ and $\mu_v(\cdot)$ are increasing for all $S > 0$ and satisfy $\mu_u(0) = \mu_v(0) = 0$.

Since the bacteria in flocs have less access to the substrate, given that this access to the substrate is proportional to the outside surface of the floc, we assume that the bacteria in flocs consume less substrate than isolated bacteria:

H2: $\mu_u(S) > \mu_v(S)$ for all $S > 0$.

Let $\varphi_u(S)$ and $\varphi_v(S)$ be the functions defined by

$$\varphi_u(S) = \mu_u(S) - D_u \quad \text{and} \quad \varphi_v(S) = \mu_v(S) - D_v.$$

When equations $\mu_u(S) = D_u$, $\mu_v(S) = D_v$ and $\varphi_v(S) = b$ have solutions, they are unique and we define the usual *break-even concentrations*

$$\lambda_u = \mu_u^{-1}(D_u), \quad \lambda_v = \mu_v^{-1}(D_v) \quad \text{and} \quad \lambda_b = \varphi_v^{-1}(b).$$

Otherwise, we put $\lambda_u = +\infty$ or $\lambda_v = +\infty$ or $\lambda_b = +\infty$. We have the following result:

Proposition 2.1 *For any non-negative initial condition, the solutions of system (1) remain non-negative and bounded for all $t \geqslant 0$. The set*

$$\Omega = \left\{ (S, u, v) \in \mathbb{R}^3_+ : S + u + v \leqslant \frac{D}{D_v} S_{in} \right\}$$

is positively invariant and is a global attractor for (1).

3. Analysis of the model

3.1. Existence of equilibria

In the following, we propose to study the existence of equilibria of (1). We use the following definitions

$$I = \begin{cases}]\lambda_u, \lambda_v[& \text{if} \quad \lambda_u < \lambda_v \\ ]\lambda_v, \min(\lambda_b, \lambda_u)[& \text{if} \quad \lambda_v < \lambda_u. \end{cases}$$

$$U(S) = \frac{D(S_{in} - S)\varphi_v(S)}{D_u\varphi_v(S) - D_v\varphi_u(S)} \quad \text{and} \quad V(S) = \frac{D(S_{in} - S)\varphi_u(S)}{D_v\varphi_u(S) - D_u\varphi_v(S)}.$$

$$H(S) = \frac{\varphi_u(S)(\varphi_v(S) - b)[D_u\varphi_v(S) - D_v\varphi_u(S)]}{a[\varphi_v(S) - \varphi_u(S)]\varphi_v(S)}. \tag{2}$$

Lemma 3.1 *The system (1), with $a > 0$ and $b > 0$, admits the following equilibria:*

– The washout, $E_0 = (S_{in}, 0, 0)$, which always exists.

– A positive equilibrium, $E^* = (S^*, u^*, v^*)$ for each S^* solution of the equation

$$D(S_{in} - S) = H(S), \tag{3}$$

$u^* = U(S^*)$ and $v^* = V(S^*)$, which exists if and only if $S^* \in I$.

The case $a = b = 0$ is simply the classical model of competition of two microbial species for which the competitive exclusion principle takes place [9].In this case, the system (1) has an equilibrium of extinction of v, $E_u = (\lambda_u, u^*, 0)$, which exists if and only if $\lambda_u < S_{in}$ and an equilibrium of extinction of u, $E_v = (\lambda_v, 0, v^*)$, which exists if and only if $\lambda_v < S_{in}$ with

$$u^* = U(\lambda_u) = \frac{D}{D_u}(S_{in} - \lambda_u) \quad \text{and} \quad v^* = V(\lambda_v) = \frac{D}{D_v}(S_{in} - \lambda_v).$$

In the following, we study the existence of positive equilibria of (1). Each solution of the equation (3) belonging to the interval I gives rise to a positive equilibrium of system. Note that

$$H'(S) = \frac{\mu_u'(\varphi_v - b)\varphi_v F + \mu_v'\varphi_u G}{a(\varphi_v - \varphi_u)^2\varphi_v^2} \tag{4}$$

where

$$\begin{array}{rcl} F & = & [D_u\varphi_v^2 - 2D_v\varphi_u\varphi_v + D_v\varphi_u^2] > 0 \\ G & = & [bD_u\varphi_v^2 + (D_v - D_u)\varphi_u\varphi_v^2 + bD_v(\varphi_u^2 - 2\varphi_u\varphi_v)] \end{array} \tag{5}$$

In the case $\lambda_u < \lambda_v$, the sign of $H'(S)$ can be positive or negative for $S \in I$ (see Fig. 1). In the case $\lambda_v < \lambda_u$, one has $H'(S) < 0$ on $I =]\lambda_v, \min(\lambda_b, \lambda_u)[$. Therefore, the function $H(\cdot)$ is decreasing on I, but equation (3) can have many solutions (see Figs. 2 and 3).

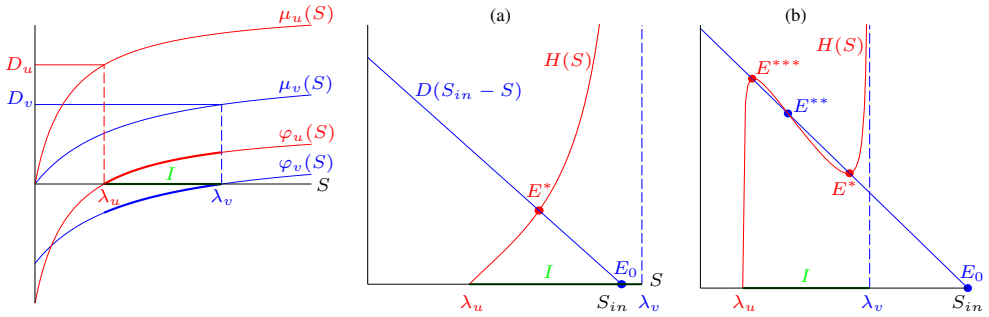


Figure 1. *The case $\lambda_u < \min(\lambda_v, S_{in})$: (a) Existence of unique positive equilibrium. (b) Existence of three positive equilibria.*

Therefore, the equation (3) may have several solutions whose number is generically odd in the case $\lambda_u < \lambda_v$ or $\lambda_v < \lambda_u < S_{in}$ (see Figs. 1 and 3 (b-c)) and even in the case $\lambda_v < S_{in} < \lambda_u$ (see Figs. 2 and 3 (a)). In all figures, we choose the color red to represent the locally exponentially stable equilibria and blue to represent the unstable equilibria. We will show the asymptotic behavior of the equilibria in section 3.2.

In the case $\lambda_u < \lambda_v$, the function $H(\cdot)$ is defined and positive on the interval $I =]\lambda_u, \lambda_v[$ since $\varphi_u(S) > 0$ and $\varphi_v(S) < 0$ for all $S \in]\lambda_u, \lambda_v[$. Moreover, it vanishes at λ_u and tends to infinity as S tends to λ_v. Hence, we have the following result:

Proposition 3.1 *If $\lambda_u < \min(\lambda_v, S_{in})$, then there exists at least one positive equilibrium. Generically, there is an odd number of positive equilibria. If $S_{in} \leqslant \lambda_u < \lambda_v$, then there is no positive equilibrium.*

In the second case $\lambda_v < \lambda_u$, the function $H(\cdot)$ is defined and positive on the interval $I =]\lambda_v, \min(\lambda_u, \lambda_b)[$ since $\varphi_u(S) < 0$ and $0 < \varphi_v(S) < b$ for all $S \in]\lambda_v, \min(\lambda_u, \lambda_b)[$. Moreover, it vanishes for $S = \min(\lambda_u, \lambda_b)$ and tends to infinity as S tends to λ_v. Hence, we have the following result:

Proposition 3.2 *If $\lambda_v < \min(\lambda_u, \lambda_b) < S_{in}$, then there exists at least one positive equilibrium. Generically, one has an odd number of positive equilibria. If $S_{in} < \min(\lambda_u, \lambda_b)$, then the system has generically no positive equilibrium or an even number of positive equilibria.*

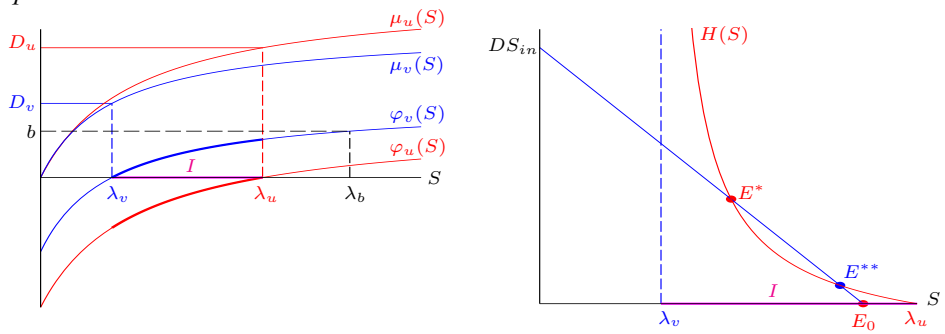


Figure 2. *The case $\lambda_v < \lambda_u < \lambda_b$: Existence of two positive equilibria and bistability for $S_{in} < \min(\lambda_u, \lambda_b)$.*

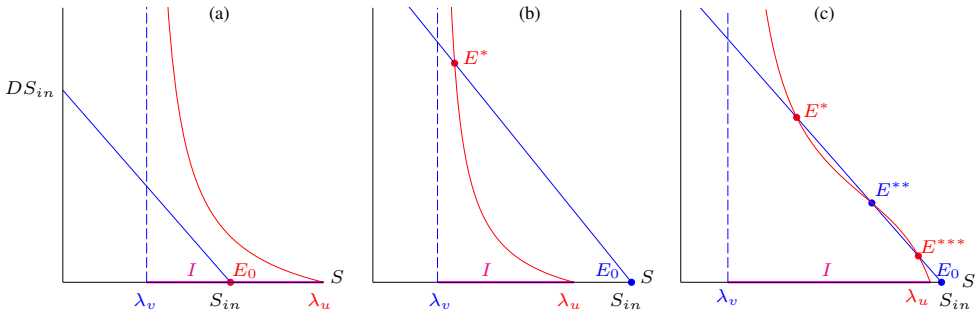


Figure 3. *The case $S_{in} < \min(\lambda_u, \lambda_b)$: There is no positive equilibria (a). The case $S_{in} > \min(\lambda_u, \lambda_b)$: Existence of one (b) or three (c) positive equilibria.*

Proposition 3.3 *Let $E^* = (S^*, u^*, v^*)$ and $E^{**} = (S^{**}, u^{**}, v^{**})$ be two positive equilibria of (1) such that $S^* < S^{**}$.*

1) If $\lambda_u < \lambda_v$, then $u^* > u^{**}$ and $v^* < v^{**}$, this means that the equilibrium E^* promotes isolated biomass u and E^{**} promotes biomass in flocs v.

2) If $\lambda_v < \lambda_u$, then $u^* > u^{**}$ and $v^* > v^{**}$, this means that the equilibrium E^* promotes simultaneously two biomass u and v.

Proposition 3.4 *The system (1) with $a = 0$, $b > 0$, admits at most three equilibria:*

– The washout, $E_0 = (S_{in}, 0, 0)$, which always exists.

– The equilibrium of extinction of v, $E_u = (\lambda_u, u^*, 0)$ with $u^* = U(\lambda_u)$, which exists if and only if $\lambda_u < S_{in}$.

– The positive equilibrium, $E^* = (S^*, u^*, v^*)$ with $S^* = \lambda_b$, $u^* = U(\lambda_b)$ and $v^* = V(\lambda_b)$, which exists if and only if $\lambda_v < \lambda_b < \lambda_u$ and $\lambda_b < S_{in}$.

Proposition 3.5 *The system (1), with $a > 0$ and $b = 0$, admits the following equilibria:*

– The washout, $E_0 = (S_{in}, 0, 0)$, which always exists.

– The equilibrium of extinction of u, $E_v = (\lambda_v, 0, v^*)$ with $v^* = V(\lambda_v)$, which exists if and only if $\lambda_v < S_{in}$.

– The positive equilibrium $E^* = (S^*, u^*, v^*)$, with S^* solution of the equation $D(S_{in} - S) = H(S)$, $u^* = U(S^*)$ and $v^* = V(S^*)$ which exists if and only if $\lambda_u < S^* < \lambda_v$ and $S^* < S_{in}$.

3.2. Stability of equilibria

We study in the following the local stability of the washout equilibrium of (1).

Proposition 3.6 *E_0 is locally exponentially stable if and only if $S_{in} < \lambda_u$ and $S_{in} < \lambda_b$.*

In the following, we study the local asymptotic behavior of the positive equilibria of (1).

Proposition 3.7 *Let $E^* = (S^*, u^*, v^*)$ be a positive equilibrium with $a > 0$ and $b > 0$.*

1) The case $\lambda_u < \lambda_v$: E^* is locally exponentially stable if $H'(S^*) > -D$ and is unstable if $H'(S^*) < -D$.

2) The case $\lambda_v < \lambda_u$: E^* is locally exponentially stable if $H'(S^*) < -D$ and is unstable if $H'(S^*) > -D$.

Table 1 summarizes the previous results:

Equilibria	Existence condition	Stability condition
E_0	Always exists	$S_{in} < \min(\lambda_u, \lambda_b)$
E^*	(3) has solution $S^* \in I$	Case $\lambda_u < \lambda_v$: $H'(S^*) > -D$ Case $\lambda_u > \lambda_v$: $H'(S^*) < -D$

Table 1. *Existence and local stability of equilibria in system (1).*

The proofs of the following results are given in [2].

Proposition 3.8 *In the case $a = 0$ and $b > 0$:*

– E_u is locally exponentially stable if and only if $\lambda_u < \lambda_b$.

– Whenever E^* exists, it is locally exponentially stable.

Similarly to proofs of Props. 3.7 and 3.8 (see [2]), we obtain the following results:

Proposition 3.9 *In the case $a > 0$ and $b = 0$:*

– E_v is locally exponentially stable if and only if $S_{in} > \lambda_v + \frac{1}{D}H(\lambda_v)$.

– The positive equilibrium $E^* = (S^*, u^*, v^*)$ is locally exponentially stable if $H'(S^*) > -D$ and is unstable if $H'(S^*) < -D$.

4. Simulations

In the case where the growth rates are Monod-type, the equation $D(S_{in} - S) = H(S)$ is equivalent to a polynomial equation of fifth degree. Therefore, there is at most five solutions of this equation. The positive equilibria correspond to solutions which are in the interval I. We succeeded in finding parameters sets with 3 solutions at most in this interval. The following Monod-type growth rates are considered where all parameter values used are given in Table 2.

$$\mu_u(S) = \frac{m_1 S}{K_1 + S} \quad \text{and} \quad \mu_v(S) = \frac{m_2 S}{K_2 + S}.$$

Fig. 4 illustrates the case $\lambda_u < \lambda_v < S_{in}$ with three positive equilibria

$$E^* \simeq (3.06, 12.11, 157.46),\ E^{**} \simeq (5.17, 8.53, 524.30),\ E^{***} \simeq (8.81, 2.64, 1086.32).$$

The numerical simulations show the bi-stability with two basins of attraction, one toward E^* and the other toward E^{***} which are stable nodes. These two basins are separated by the stable manifold of a saddle point E^{**}. As it was proved in Prop. 3.3, u is promoted at E^* and v is promoted at E^{***}.

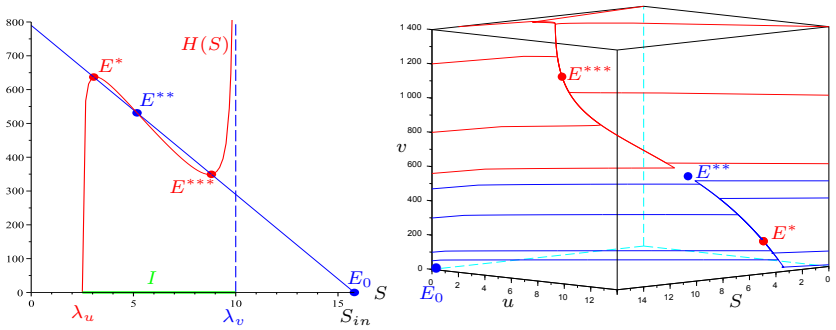


Figure 4. *The case $\lambda_u < \lambda_v < S_{in}$: Three positive equilibria and bi-stability.*

Fig. 5 illustrates the case $S_{in} > \lambda_u > \lambda_v$ with three positive equilibria

$$E^* \simeq (3.31, 2.23, 27.08),\ E^{**} \simeq (3.98, 1.67, 4.12),\ E^{***} \simeq (4.39, 0.63, 0.24).$$

The numerical simulations show the bi-stability of E^* and E^{***} which are stable nodes. The two basins of attraction are separated by the stable manifold of a saddle point E^{**}. As it was proved in Prop. 3.3, u and v are both promoted at E^*.

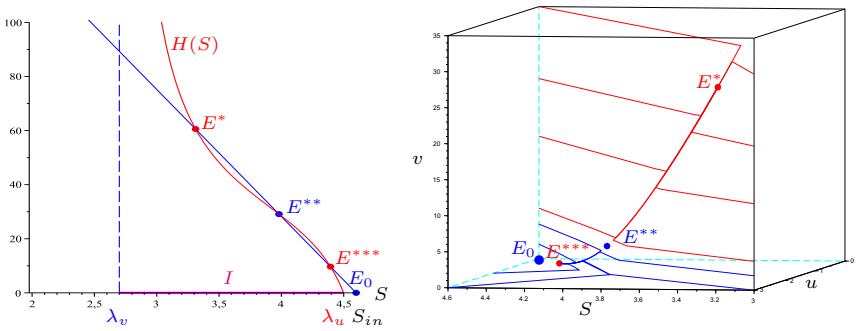


Figure 5. *The case $\lambda_v < \lambda_u < S_{in}$: Existence of three positive equilibria and bi-stability.*

5. Conclusion

In this work, we have considered a model of the chemostat with a single growth-limiting resource and one species that is present in two forms: isolated and aggregated bacteria. We have assumed that the growth rates are increasing and the dilution rates are distinct. Without assuming that attachment and detachment dynamics are faster than the

growth dynamics of planktonic and attached biomass, the qualitative behavior of three order model (1) is analyzed. We have shown the multiplicity of positive equilibria with the possibility of bi-stability of two positive equilibria which can promote the planktonic and/or aggregated biomass. Whereas, the bi-stability could occur in the classical chemostat model only when the growth rate is non monotonic. The simulations illustrate the mathematical results demonstrated.

A. Proofs

Proof of Prop. 2.1. One has

$$\begin{array}{l} S = 0 \;\Rightarrow\; \dot{S} = DS_{in} > 0\,, \\ v = 0 \;\Rightarrow\; \dot{v} = au^2 \geqslant 0\,. \end{array}$$

Hence $S(t) \geqslant 0$ and $v(t) \geqslant 0$ for all $t \geqslant 0$. One has also

$$u = 0 \;\Rightarrow\; \dot{u} = bv \geqslant 0\,,$$

and then $u(t) \geqslant 0$ for all $t \geqslant 0$. Denote $z = S + u + v$. The sum of the three equations of (1) gives

$$\dot{z}(t) \leqslant D_v \left[\frac{D}{D_v} S_{in} - z(t)\right].$$

Hence, one has

$$z(t) \leqslant \frac{D}{D_v} S_{in} + (z(0) - \frac{D}{D_v} S_{in}) e^{-D_v t} \quad \text{for all} \quad t \geqslant 0. \tag{6}$$

We deduce that

$$z(t) \leqslant \max\left(z(0), \frac{D}{D_v} S_{in}\right) \quad \text{for all} \quad t \geqslant 0.$$

Thus, the solution of (1) is positively bounded and is defined for all $t \geqslant 0$. From (6), it can be deduced that the set Ω is positively invariant and is a global attractor for (1). ■

Proof of Lemma 3.1. We must solve the system

$$\left\{\begin{array}{l} D(S_{in} - S) = \mu_u(S)u + \mu_v(S)v \\ 0 = [\mu_u(S) - D_u]u - a(u+v)u + bv \\ 0 = [\mu_v(S) - D_v]v + a(u+v)u - bv. \end{array}\right. \tag{7}$$

Making the sum of the second and the third equation of (7), we obtain

$$\varphi_u(S)u + \varphi_v(S)v = 0. \tag{8}$$

This equation admits positive solutions u and v if and only if $\varphi_u(S)$ and $\varphi_v(S)$ have opposite signs, i.e. S is between λ_u and λ_v. Therefore, we must seek solutions (S, u, v) of (7) such that S is between λ_u and λ_v. In this case, $\varphi_v(S) \neq 0$ and the equation (8) can be rewritten as

$$v = -\frac{\varphi_u(S)}{\varphi_v(S)} u. \tag{9}$$

If $u = 0$, then from the second equation of (7), we deduce $v = 0$. If $v = 0$, then from the last equation of (7), we deduce $u = 0$. Hence we cannot have an equilibria of extinction

only of u or only of v. Replacing v by its expression (9) in the second equation of (7), we obtain

$$u = U_1(S) \quad \text{with} \quad U_1(S) = \frac{\varphi_u(S)(\varphi_v(S) - b)}{a[\varphi_v(S) - \varphi_u(S)]}. \tag{10}$$

Note that u defined by (10) is positive if and only if $\lambda_u < S < \lambda_v$ or $\lambda_v < S < \min(\lambda_b, \lambda_u)$, that is to say, if and only if $S \in I$.

Therefore, we must seek the solutions of (7) such that $S \in I$. By replacing u by (10) in (9), we obtain

$$v = V_1(S) \quad \text{with} \quad V_1(S) = -\frac{\varphi_u^2(S)(\varphi_v(S) - b)}{a[\varphi_v(S) - \varphi_u(S)]\varphi_v(S)}. \tag{11}$$

Making the sum of three equations of (7) and replacing u and v by (10) and (11), it follows that S is solution of equation (3). Hence,

$$\frac{\varphi_u(S)(\varphi_v(S) - b)}{a[\varphi_v(S) - \varphi_u(S)]} = \frac{D(S_{in} - S)\varphi_v(S)}{D_u\varphi_v(S) - D_v\varphi_u(S)}.$$

Therefore, (10) and (11) can be rewritten as $u = U(S)$ and $v = V(S)$. ■

Proof of Prop. 3.3. We show that

1) If $\lambda_u < \lambda_v$, then $U(\cdot)$ is decreasing on $I \cap]0, S_{in}[$ and $V_1(\cdot)$ is increasing on I.

2) If $\lambda_v < \lambda_u$, then $U_1(\cdot)$, $V(\cdot)$ and $V_1(\cdot)$ are decreasing on $I \cap]0, S_{in}[$.

Indeed, we have

$$U'(S) = D\frac{-\varphi_v(D_u\varphi_v - D_v\varphi_u) - \mu_v' D_v\varphi_u(S_{in} - S) + \mu_u' D_v\varphi_v(S_{in} - S)}{(D_u\varphi_v - D_v\varphi_u)^2},$$

$$U_1'(S) = \frac{\mu_u'\varphi_v(\varphi_v - b) + \mu_v'\varphi_u(b - \varphi_u)}{a(\varphi_v - \varphi_u)^2}.$$

Therefore, if $\lambda_u < \lambda_v$, then $U'(S)$ is negative on $I \cap]0, S_{in}[$ and if $\lambda_v < \lambda_u$, then $U_1'(S)$ is negative on I. In addition, we have

$$V'(S) = D\frac{-\varphi_u(D_v\varphi_u - D_u\varphi_v) - \mu_u' D_u\varphi_v(S_{in} - S) + \mu_v' D_u\varphi_u(S_{in} - S)}{(D_v\varphi_u - D_u\varphi_v)^2},$$

$$V_1'(S) = \frac{-\mu_u'[\varphi_u\varphi_v(\varphi_v - b)](2\varphi_v - \varphi_u) + \mu_v'\varphi_u^2(\varphi_v - b)(2\varphi_v - \varphi_u)}{a(\varphi_v - \varphi_u)^2\varphi_v^2}.$$

If $\lambda_u < \lambda_v$, then $V_1'(S)$ is positive on I and if $\lambda_v < \lambda_u$, then $V'(S)$ is negative on $I \cap]0, S_{in}[$ and $V_1'(S)$ is negative on I. Therefore, if $\lambda_u < \lambda_v$, then

$$u^* = U(S^*) > u^{**} = U(S^{**}) \quad \text{and} \quad v^* = V_1(S^*) < v^{**} = V_1(S^{**}).$$

Furthermore, if $\lambda_v < \lambda_u$ then

$$u^* = U_1(S^*) > u^{**} = U_1(S^{**}) \quad \text{and} \quad v^* = V(S^*) > v^{**} = V(S^{**}).$$

■

Proof of Prop. 3.4. When $a = 0$, the system (7) is written as

$$\begin{cases} D(S_{in} - S) = \mu_u(S)u + \mu_v(S)v \\ 0 = [\mu_u(S) - D_u]u + bv \\ 0 = [\mu_v(S) - D_v]v - bv. \end{cases} \tag{12}$$

The third equation of (12) can be rewritten as

$$\varphi_v(S)v - bv = 0.$$

If $v = 0$, then from the second equation of (12), we deduce $u = 0$ or $S = \lambda_u$. If $u = v = 0$, then from the first equation, one has $S = S_{in}$. If $v = 0$ and $S = \lambda_u$, then from the first equation we deduce

$$u = \frac{D(S_{in} - \lambda_u)}{D_u} = U(\lambda_u)$$

which is positive if and only if $\lambda_u < S_{in}$. If v is nonzero and the equation $\varphi_v(S) = b$ has solution $S = \lambda_b$, then from the second equation of (12), we deduce u is nonzero and

$$\varphi_u(\lambda_b)u + bv = 0.$$

This equation admits positive solutions u and v if and only if $\lambda_b < \lambda_u$. Making the sum of the second and the third equation of (12), we obtain the equation (8) which admits positive solutions u and v if and only if λ_b is between λ_u and λ_v. Making the sum of the second and the third equation of (12), the first equation is rewritten as

$$D(S_{in} - \lambda_b) = D_u u + D_v v.$$

Replacing v by its expression (9), we obtain

$$D(S_{in} - \lambda_b) = D_u u - D_v \frac{\varphi_u(\lambda_b)}{\varphi_v(\lambda_b)} u = \frac{D_u\varphi_v(\lambda_b) - D_v\varphi_u(\lambda_b)}{\varphi_v(\lambda_b)} u.$$

Hence $u = U(\lambda_b)$ and from the equation (9), we deduce that $v = V(\lambda_b)$ which are positive if and only if $\lambda_v < \lambda_b < \lambda_u$ and $\lambda_b < S_{in}$. ■

Proof of Prop. 3.5. When $b = 0$, the system (7) is written as

$$\begin{cases} D(S_{in} - S) = \mu_u(S)u + \mu_v(S)v \\ 0 = (\mu_u(S) - D_u)u - a(u + v)u \\ 0 = (\mu_v(S) - D_v)v + a(u + v)u. \end{cases} \tag{13}$$

Note that in this case $b = 0$, the expression (2) of $H(S)$ is simplified and becomes

$$H(S) = \frac{\varphi_u(S)[D_u\varphi_v(S) - D_v\varphi_u(S)]}{a[\varphi_v(S) - \varphi_u(S)]}. \tag{14}$$

Moreover, $\lambda_b = \lambda_v$. Therefore, the interval I is empty in the case $\lambda_v < \lambda_u$. The second equation of (13) can be rewritten as

$$\varphi_u(S)u - a(u + v)u = 0.$$

If $u = 0$, from the last equation, we deduce $\varphi_v(S) = 0$, means that $S = \lambda_v$ and from the first equation $v = V(\lambda_v)$ which is positive if and only if $\lambda_v < S_{in}$. The previous calculation shows that if u is nonzero then

$$D(S_{in} - S) = D_u u + D_v v = D_u u - D_v \frac{\varphi_u}{\varphi_v} u.$$

Hence $u = U(S)$ and $v = V(S)$ which are positive if and only if $\lambda_u < S < \lambda_v$ with S solution of the equation $D(S_{in} - S) = H(S)$. ■

Proof of Prop. 3.6. The Jacobian matrix at washout $E_0 = (S_{in}, 0, 0)$, is given by

$$\mathbf{J_{E_0}} = \begin{bmatrix} -D & -\mu_u(S_{in}) & -\mu_v(S_{in}) \\ 0 & \varphi_u(S_{in}) & b \\ 0 & 0 & \varphi_v(S_{in}) - b \end{bmatrix}.$$

The eigenvalues are $-D, \varphi_u(S_{in})$ and $\varphi_v(S_{in}) - b$. ■

Proof of Prop. 3.7. The Jacobian matrix at a positive equilibrium $E^* = (S^*, u^*, v^*)$ is given by

$$\mathbf{J_{E^*}} = \begin{bmatrix} -m_{11} & -m_{12} & -m_{13} \\ m_{21} & -m_{22} & a_{23} \\ m_{31} & m_{32} & -m_{33} \end{bmatrix}$$

where $m_{11} = D + \mu_u'(S^*)u^* + \mu_v'(S^*)v^*,\ m_{12} = \mu_u(S^*),\ m_{13} = \mu_v(S^*),$

$$m_{21} = \mu_u'(S^*)u^*, \quad m_{22} = a(2u^* + v^*) - \varphi_u(S^*), \quad a_{23} = b - au^*,$$

$$m_{31} = \mu_v'(S^*)v^*, \quad m_{32} = a(2u^* + v^*) \quad \text{and} \quad m_{33} = b - au^* - \varphi_v(S^*).$$

From the second equation of (7), we have

$$\begin{aligned} \varphi_u(S^*)u^* - a(u^* + v^*)u^* + bv^* &= \varphi_u(S^*)u^* - a(2u^* + v^*)u^* + a(u^*)^2 + bv^* \\ &= -m_{22}u^* + a(u^*)^2 + bv^* = 0. \end{aligned}$$

Hence $m_{22} = au^* + bv^*/u^* > 0$. From the third equation of (7), we have

$$\varphi_v(S^*)v^* + a(u^* + v^*)u^* - bv^* = -m_{33}v^* + a(u^*)^2 = 0.$$

and therefore,

$$m_{33} = a\frac{(u^*)^2}{v^*} > 0.$$

Thus, all m_{ij} are positive for all $i, j = 1, \ldots, 3$ with $(i, j) \neq (2, 3)$. The characteristic polynomial is given by

$$P(\lambda) = |J_{E^*} - \lambda * I| = c_0\lambda^3 + c_1\lambda^2 + c_2\lambda + c_3,$$

where I is the 3×3 identity matrix, $c_0 = -1$, $c_1 = -(m_{11} + m_{22} + m_{33})$,

$$c_2 = -m_{12}m_{21} - m_{13}m_{31} + m_{32}a_{23} - (m_{11}m_{22} + m_{11}m_{33} + m_{22}m_{33}),$$

$$c_3 = -m_{11}(m_{22}m_{33} - m_{32}a_{23}) - m_{21}(m_{12}m_{33} + m_{32}m_{13}) - m_{31}(m_{12}a_{23} + m_{13}m_{22}).$$

It is clear that $c_0 = -1 < 0$ and, since $m_{ii} > 0, i = 1, \ldots, 3$, we have $c_1 < 0$. It can be shown by long and tedious calculations (see [2]) that

$$c_2 < 0 \text{ and } c_1c_2 - c_0c_3 > 0$$

and that we have the following properties

1) In the case where $\lambda_u < \lambda_v$, we have $c_3 < 0$ if and only if $H'(S^*) > -D$.

2) In the case where $\lambda_v < \lambda_u$, we have $c_3 < 0$ if and only if $H'(S^*) < -D$.
The result of stability follows from the Routh-Hurwitz criterion, which asserts that E^* is locally exponentially stable if and only if

$$\begin{cases} c_i < 0, & i = 0, \ldots, 3 \\ c_1 c_2 - c_0 c_3 > 0. \end{cases}$$

This completes the proof. ■

B. Parameters used in numerical simulations

Parameter	m_1	K_1	m_2	K_2	D	D_u	D_v	a	b	S_{in}	λ_u	λ_v
Fig. 4	60	0.5	0.6	20	50	50	0.2	0.01	0.01	15.8	2.5	10
Fig. 5	20	1.5	2	2.7	47	15	1	1.2	3	4.6	4.5	2.7

Table 2. *Parameter values and the corresponding λ_u and λ_v.*

Acknowledgments. The authors wish to thank the financial support of TREASURE euro-Mediterranean research network (https://project.inria.fr/treasure/). This work was partly done in the PhD thesis of the first author within the INRA/INRIA team MODEMIC, with the financial support of the Averroes program, the PHC UTIQUE project No. 13G1120 and the COADVISE project.

C. References

[1] J. COSTERTON, "Overview of microbial biofilms", *J. Indust. Microbiol.*, vol. 15, 1995, 137–140.

[2] R. FEKIH-SALEM, "Modèles mathématiques pour la compétition et la coexistence des espèces microbiennes dans un chémostat", *PhD thesis, UM2-UTM*, 2013.

[3] R. FEKIH-SALEM, J. HARMAND, C. LOBRY, A. RAPAPORT, T. SARI, "Extensions of the chemostat model with flocculation", *J. Math. Anal. Appl.*, vol. 397, 2013, 292–306.

[4] R. FEKIH-SALEM, T. SARI, A. RAPAPORT, "La floculation et la coexistence dans le chemostat", *Proceedings of the 5th conference on Trends in Applied Mathematics in Tunisia, Algeria, Morocco*, 2011, 477–483.

[5] R. FRETER, H. BRICKNER, S. TEMME, "An understanding of colonization resistance of the mammalian large intestine requires mathematical analysis", *Microecology and Therapy*, vol. 16, 1986, 147–155.

[6] B. HAEGEMAN, A. RAPAPORT, "How flocculation can explain coexistence in the chemostat", *J. Biol. Dyn.*, vol. 2, 2008, 1–13.

[7] D. JONES, H.V. KOJOUHAROV, D. LE, H.L. SMITH, "The Freter model: A simple model of biofilm formation", *J. Math. Biol.*, vol. 47, 2003, 137–152.

[8] S. PILYUGIN, P. WALTMAN, "The simple chemostat with wall growth", *SIAM J. Appl. Math.*, vol. 59, 1999, 1552–1572.

[9] H.L. SMITH, P. WALTMAN, "The Theory of the Chemostat: Dynamics of Microbial Competition", *Cambridge University Press*, 1995.

[10] D.N. THOMAS, S.J. JUDD, N. FAWCETT, "Flocculation modelling: a review", *Water Res.*, vol. 33, 1999, 1579–1592.

Rubrique

Modeling the dynamics of cell-sheet : From Fisher-KPP equation to bio-mechano-chemical systems

Fisher-KPP equation to study some predictions on the injured cell sheet

Mekki Ayadi [1] — Abderahmane Habbal [2] — Boutheina Yahyaoui [1]

[1] Tunis El Manar University, National Engineering School of Tunis
ENIT-LAMSIN BP 37, 1002 Tunis, LR 95–ES–20
Tunisia
mekki.ayadi@enis.rnu.tn
boutheinayahyaoui@hotmail.fr
[2] INRIA, 2004 route des lucioles-BP 93
06902 Sophia Antipolis Nice cedex
France
Abderrahmane.HABBAL@unice.fr

RÉSUMÉ. Dans le cadre de la cicatrisation d'un feuillet cellulaire, nous avons étudié la validité des modèles de réaction-diffusion de type Fisher-KPP pour la simulation de la migration de feuillets cellulaires. Afin d'étudier la validité de ce modèle, nous avons effectué des observations expérimentales sur les monocouches de cellules MDCK. Les videoscopies obtenues permettent, après segmentation et binarisation, d'obtenir avec précision les variations d'aire et de profils de fronts de cicatrice. Nous nous sommes intéressés à comparer les variations des fronts calculés à celles des fronts expérimentaux, après une étape de calage des paramètres.

ABSTRACT. This paper is devoted to study some predictions on the injured cell sheet mainly based on reaction-diffusion equations. In the context of healing of cell sheet, we investigated the validity of the reaction-diffusion model of Fisher-KPP type for simulation of cellular sheets migration. In order to study the validity of this model, we performed experimental observations on the MDCK cell monolayers. The obtained videoscopies allow to obtain, after segmentation and binarization, the variations of area and of scar fronts profiles with good accuracy. We were interested in comparing the calculated variations of fronts to those experimental fronts, after a step of calibration parameters.

MOTS-CLÉS : Cellules MDCK, Fisher-KPP, simulation 2D, dynamique cellulaire, coefficient de diffusion D, taux de prolifération r.

KEYWORDS : MDCK, Fisher-KPP, wound edge dynamics, diffusion coefficient D, proliferation rate r, activation, inhibition.

1. Introduction

"Medicine and mathematics, this may seem a bold rapprochement. And yet ... medical imaging provides a unique way to access the inaccessible, the shape and function of internal organs from living human body, without being invasive. Thanks to medical images, physicians and surgeons can see what remains invisible during an examination with the naked eye. What is also visible, it is the essential role of mathematics and computing not only in the formation of these images, but also in their use".

Grégoire Malandain [1].

Modeling cell dynamics, for studying the cell-sheet wound healing, is a subject of great importance and at the intersection of three major fields of science. In fact, biology for the part of experimental measurements and filtering data, mechanical for modeling the movement of the tissue and its effect on wound healing and mathematical and numerical modeling to quantify biological and mechanical quantities previously mentioned.

Modeling bio-mechano-chemical behavior coupled with complex biological systems, such as the formation of a pattern in embryogenesis, modeling of tumor growth and wound healing, is in mathematical term won by partial differential equations of reaction-diffusion type [2]. This family of equations is well suited to describe in time and in space changes that occur within the cell population and the production of migration and proliferation. Both mechanisms are most important during wound healing. Basically, the diffusion cells is related to their roving, while the reaction is related to proliferation. Reaction-diffusion equations coupled to mechanics with viscoelastic behavior, take into account haptotaxis andhaptokinesis of cell movement [3].

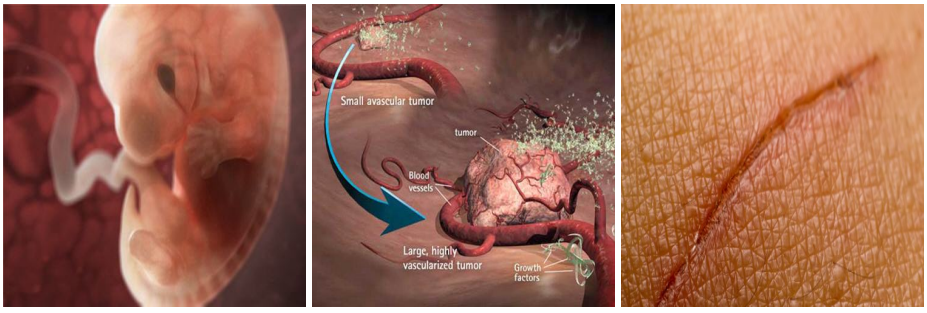


Figure 1. *Formation of a pattern in embryogenesis, tumor growth and wound healing, http ://www.linternaute.com/science/biologie/dossiers/07/cerveau-sexe/page4.jpg, http ://www.santevitalite.be/wp-content/uploads/2012/12/Croissance-tumorale.jpg, https ://www.simplyscience.ch/system/html/Croute-01b7ea33.jpg*

In this study, we consider a particular aspect of wound healing, namely that relative to the flow of monolayers of wounded epithelial cells of Madin-Darby Canine Kidney (MDCK) [4, 5]. The population of cells in epithelial monolayers, also called cellular sheet can be considered as a two-dimensional structure. After creating a wound, the cells begin to regrowth in order to fill the empty space. Although wound closure involves biochemical and biomechanical process, still far from being understood, which are distributed

throughout the monolayer, particular attention was paid to changes in the front. Moreover, the effects of migration activators of HGF (Hepatocyte Growth Factors) type [6] and the effects of inhibitors of PI3K (phosphoinositide 3-kinase) type were taken into account in an experimental test campaign.
To our knowledge, J.D. Murray published during the 2000s an interesting study describing the relationship between biology and mathematics in his book entitled *Mathematical Biology* in two parts [7], [8]. He proposes a vision of a mathematician to study reaction-diffusion models that describe the problems of interactions between biological, chemical and mechanical phenomena. Mathematical biology allows to pass from dynamic analysis of cells to a mechano-biochemical system governed by reaction-diffusion equations, coupled to mechanics equations with a visco-elastic behavior, as well as to explain the phenomena of chemotaxis and haptotaxis among other characteristics of cell movement [9, 10, 11].

In order to build a powerful mechanical model for modeling biological problems difficult to solve, we refer interested readers to the articles [12, 13, 14, 15] . The authors of these articles consider that healing is largely a mechanical process where the chemical effect simply acts to increase the overall behavior. Moreover, in the works of Maini, Olsen and Sheratt published in [3], [16], [17] presented a complete coupled model whose basic variables are cell density n, the density of ECM ρ and the displacement of tissue u, see the following equations.

$$\frac{\partial n}{\partial t} + div[n\frac{\partial u}{\partial t} + \chi(\rho)n\nabla\rho - D(\rho)\nabla n] = rn(1-n), \tag{1.1}$$

$$\frac{\partial \rho}{\partial t} + div[\rho\frac{\partial u}{\partial t}] = \varepsilon n(1-\rho), \tag{1.2}$$

$$div(\sigma) = \rho s u, \tag{1.3}$$

where
• $div(n\frac{\partial u}{\partial t})$ represents the passive convection, while, $div(\chi(\rho)n\nabla\rho)$ represents the haptotaxis phenomenon, $(-div(D(\rho)\nabla n)$ represents the haptokinesis phenomenon and $rn(1-n)$ represents the cell proliferation,
• $div(\rho\frac{\partial u}{\partial t})$ represents the passive convection, while $\varepsilon n(1-\rho)$ represents the ECM biosynthesis and the degradation of cells fibroplast,
• $\sigma = \sigma_{ECM} + \sigma_{cell}$ with $\sigma_{ECM} = \mu_1\frac{\partial \varepsilon}{\partial t} + \mu_2\frac{\partial \Theta}{\partial t}I + \frac{E}{1+\nu}(\varepsilon(u) + \frac{\nu}{1-2\nu}\Theta I)$ represents the viscous and elastic forces, while, $\sigma_{cell} = cnI$ represents the traction forces, $\varepsilon(u) = \frac{1}{2}(\nabla u + \nabla u^T)$ is the strain tensor and $\Theta = tr(\varepsilon)$ is the dilatation of the matrix material.

Many scientific articles describe in detail this model. These include for example that of Perelson and *all*, [18]. In this paper is shown how the above equations are found as well as their numerical implementation. In the article [19], Sherratt offers monodimensionel and another bidimensional model which include only biomechanical coupling to describe cellular dynamics during healing embryonic dermal wounds. In [20], Goto uses mechanochemical model, which is a simplified version of the full model mentioned above, for the formation of a somite to better understand the role played by the mechanical aspects of the cells and the extracellular matrix (ECM) in the somitogenesis.

2. Material and methods

2.1. Mathematical method

In what follows, the mechanical and chemical effects are neglected ; only the biological effect is considered. Hence, the full coupled model $(1.1) - (1.3)$ reduces to the Fisher-KPP equation.

$$n_t - D\Delta n + g(n) = 0, \quad (x, y) \in \Omega, \quad 0 < t \leq T, \tag{2.1}$$

with the initial condititon

$$n(x, y, 0) = n_0(x, y), \quad (x, y) \in \Omega, \tag{2.2}$$

and with the boundary conditions

$$n = 1, \quad \text{on} \quad \Gamma_D, \tag{2.3}$$

$$\frac{\partial n}{\partial y} = 0, \quad \text{on} \quad \Gamma_N, \tag{2.4}$$

where $g(n) = -rn(1 - n)$, Ω is a bounded rectangular open set of $\mathbb{R}^2$, Γ_D and Γ_N are respectively the vertical sides and the horizontal sides of Ω, see Figure 2 [21], D and r are a positive constants and stand for the cell diffusion coefficient and the cell proliferation rate respectively.

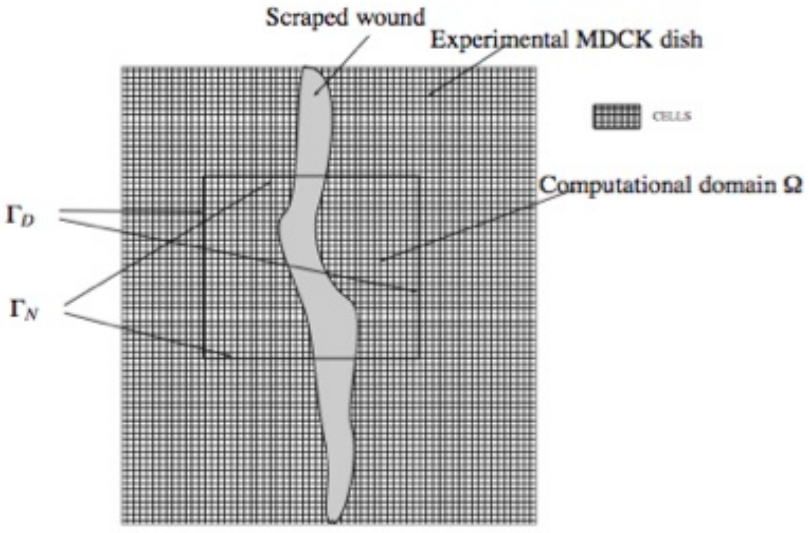


Figure 2. *A neighborhood Ω of the wound.*

2.2. Experimental method

The experimental method has been presented as following : we conducted five experiments for healing wound, which give five data sets, each composed of 360 images. From each set, we extract a series of 120 two-dimensional images of 1392×1040 pixels coded on 2 bytes, which corresponds to a step time of 6 minutes between two consecutive images. The tests are classified as follows :

• Assay I (Seq5) : considered as a control test or as a reference test (in which neither activator nor inhibitor migration was used).

• Assay II (Seq2) and Assay III (Seq4) : control test + HGF activator.

• Assay IV (Seq3) and Assay V(Seq6) : control test + inhibitor.

We recorded biological videos filming the various stages of wound closure. The videos were then segmented to obtain raw images, then they have been binarized to obtain images ready to deal with : the experimental density of cells, denoted n_{exp}, is provided. Using this density, we have successfully implemented experimental area of the wound :

$$W_{exp}(t) = \int_{\Omega} (1 - n_{exp}(x, y, t))dxdy. \tag{2.5}$$

These experimental results were compared to numerical results related to the numerical solution, denoted n_{num}, of the KPP-Fisher equation discretized in space using a finite difference scheme of order two and in time using the Crank Nicolson scheme with Splitting. It is more precisely to minimize with respect to parameters r and D the following two costs

$$J_U(r,\ D) = \int_{[T_0,T]} \int_{\Omega} |n_{num}(x, y, t) - n_{\text{exp}}(x, y, t)|dxdydt, \tag{2.6}$$

$$J_A(r,\ D) = \int_{[T_0,T]} |W_{num}(t) - W_{\text{exp}}(t)|dt. \tag{2.7}$$

The first cost is the norm of the error between the numerical solution and the experimental solution, while the second cost is the norm of the error between the numerical area and the experimental area. Figure (3) below shows the surfaces cost J_U and J_A in terms of parameters r and D.

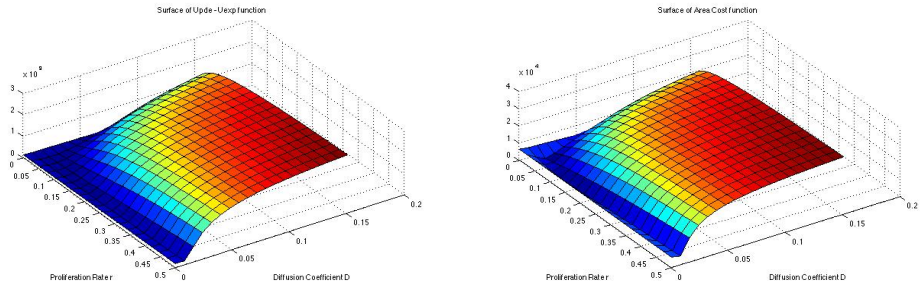


Figure 3. *Cost functions of surfaces J_U and J_A function of parameters (r,D)*

The numerical results on the area of the wound depending on time, obtained by Habbal et *al*, in the absence of activation and inhibition, published in [21], are shown in the following figures. The blue curve represents the experimental area variation of the wound

with respect to time, while, the red curve represents the numerical area of the wound with respect to time.

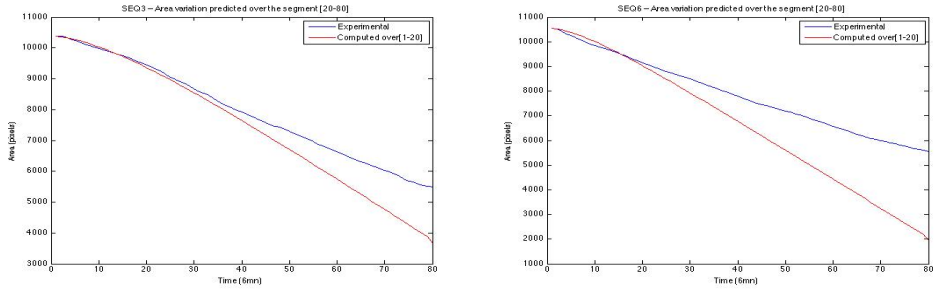


Figure 4. *The curves show the area of the wound depending on time for sequences 3 and 6 respectively.*

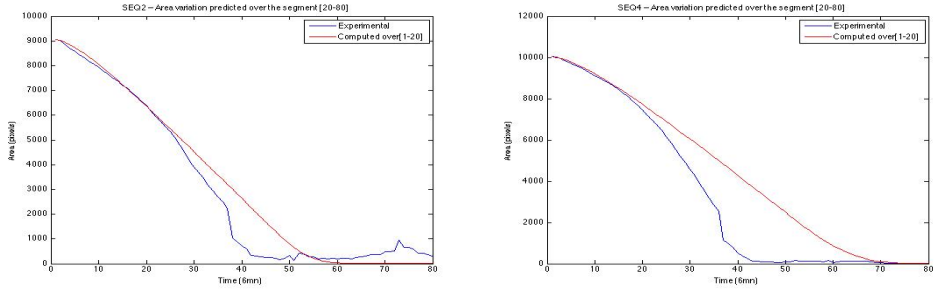


Figure 5. *The curves show the area of the wound depending on time for sequences 2 and 4 respectively.*

The above curves show that if the numerical area well approach the experimental area at the beginning of healing, this approximation is not at all satisfactory in the remaining time. To improve this approximation, we made recourse to activation and inhibition operations which are the novelty of this paper and the object of the following section.

From numerical point of view, activation and inhibition operations are taken into account by assuming that parameters D and r vary over time in a very precise manner. Charles Hansen suggested, after long studies on the choice of parameters including biological problems such as the prediction of cytotoxic drug interactions with DNA [22], a variation in sigmoid shape.

3. Results

3.1. Numerical Results

A sigmoid function, see [23, 24], is a S-shaped curve whose general expression is $\phi(t) = \frac{k}{1+\alpha exp(-\lambda t)}$. Its growth is slow at first, then accelerates strongly before slowing to end up not grow. In a first step, we have chosen to vary only parameter D over time :

$$D(t) = \frac{k}{1 + \alpha exp(-\lambda t)}, \qquad \lim_{t \to +\infty} D(t) = k$$

The numerical results are very satisfactory as shown in the following figures :

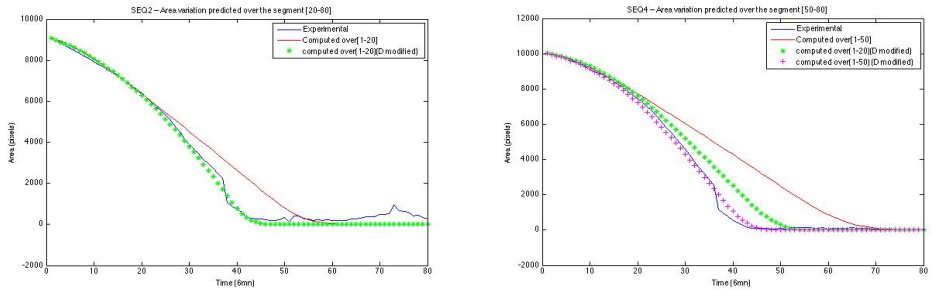


Figure 6. *The curves show the numerical areas obtained after activation with sequences 2 and 4, respectively*

These numerical results with the optimal settings : $k_2^* = 2.00e-02$, $\lambda_2^* = 2.49e-02$, $\alpha_2^* = 3.00e+01$, $r_2^* = 2.26e-01$ for the second sequence and $k_4^* = 2.00e-02$, $\lambda_4^* = 4.00e-02$, $\alpha_4^* = 6.00e+01$, $r_4^* = 2.21e-01$ for the fourth sequence, realize the minimum of the error between the experimental and the numerical area. This optimization operation has been performed using the Matlab function "fmincon".

We are now interested in choosing a sigmoid pattern compatible with the inhibition operation to improve the numerical area for sequences 3 and 6.

$$D(t) = \frac{3k}{2} - \frac{k}{1 + \alpha exp(-\lambda t)}, \qquad \lim_{t \to +\infty} D(t) = \frac{k}{2}$$

Numerical results obtained are less satisfactory than those obtained in the case of activation as shown in Figure 7.

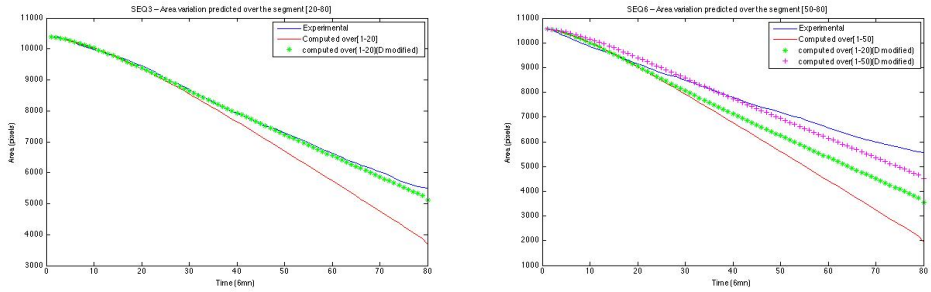


Figure 7. *The curves show the numerical areas obtained after inhibition with sequences 3 and 6 respectively*

These numerical results, illustrated by the green curve, are obtained with the optimal settings : $k_3^* = 1.32e-02$, $\lambda_3^* = 4.00e-02$, $\alpha_3^* = 3.00e+01$, $r_3^* = 2.99e-02$ for the third sequence and $k_6^* = 2.00e-02$, $\lambda_6^* = 4.00e-02$, $\alpha_6^* = 3.00e+01$, $r_6^* = 2.93e-02$ for the sixth sequence. If the result obtained with inhibition and sequence 3 is more or less acceptable, that of the sequence 6 is not at all acceptable. Hoping to get better results, we decided to also vary the parameter r over time.

Drawing on results in Figure 7, we choose the following variations of parameters r and D with respect to time :

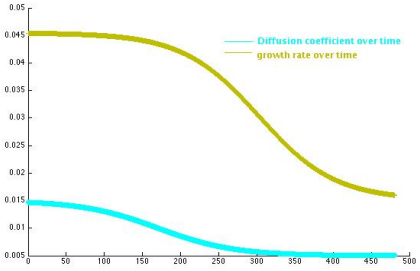


Figure 8. *Diffusion and proliferation coefficients time-dependent*

Numerical results obtained become very satisfactory as shown in the following figures.

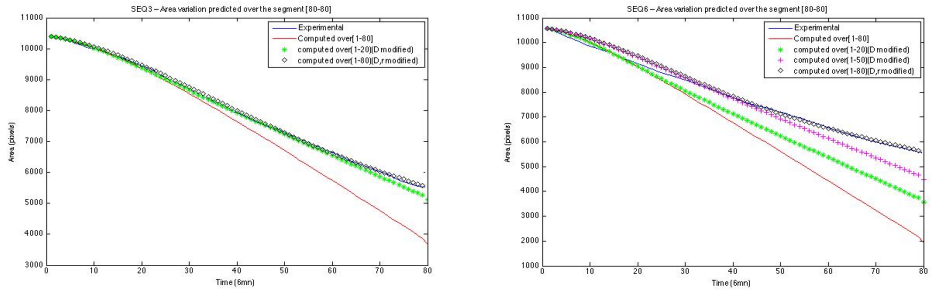


Figure 9. *The curves show the numerical areas obtained after inhibition with sequences 3 and 6 respectively*

In both cases, the numerical curve (shown in black diamonds) coincides so much with the experimental curve (shown in continu blue line) that one can not see the latter. These numerical results are obtained with the optimal settings : $k^*_{D3} = 2.00e-02$, $\lambda^*_{D3} = 4.00e-02$, $\alpha^*_{D3} = 3.4631e+03$, $k^*_{r3} = 1.72e-02$, $\lambda^*_{r3} = 1.00e-02$, $\alpha^*_{r3} = 12.1045$ for the third sequence and $k^*_{D6} = 2.00e-02$, $\lambda^*_{D6} = 4.00e-02$, $\alpha^*_{D6} = 3.4631e+03$, $k^*_{r6} = 1.93e-02$, $\lambda^*_{r6} = 1.05e-02$, $\alpha^*_{r6} = 14.5041$ for the sixth sequence.

3.2. Theoretical Results

We are interested in this section to the theoretical study of Fisher-KPP equation when the diffusion coefficient D and the proliferation rate r vary over time : estimate the difference of the wound surface variation for the two cases, constant and time-dependent. For a similar study we refer to [25] and [26].

3.2.1. Diffusion and proliferation coefficients time-dependent

Consider the two following problems :

$$(P_1)\begin{cases} \partial_t u_1 = D\Delta u_1 + r u_1(1-u_1), & (x,y,t)\in\Omega\times(0,\ T], \quad (P_1.1)\\ u_1(x,y,0) = u_{1,0}(x,y), & (x,y)\in\overline{\Omega},\\ u_1(x,y,t) = u_H(x,y,t), & (x,y,t)\in\Gamma_D\times(0,\ T],\\ \frac{\partial u_1}{\partial n}(x,y,t) = g(x,y,t), & (x,y,t)\in\Gamma_N\times(0,\ T]. \end{cases}$$

$$(P_2)\begin{cases} \partial_t u_2 = D(t)\Delta u_2 + r(t) u_2(1-u_2), & (x,y,t)\in\Omega\times(0,\ T], \quad (P_2.1)\\ u_2(x,y,0) = u_{2,0}(x,y), & (x,y)\in\overline{\Omega},\\ u_2(x,y,t) = u_H(x,y,t), & (x,y,t)\in\Gamma_D\times(0,\ T],\\ \frac{\partial u_2}{\partial n}(x,y,t) = g(x,y,t), & (x,y,t)\in\Gamma_N\times(0,\ T]. \end{cases}$$

Lemme 3.1 *Suppose that* $u_{1,0}, u_{2,0}\in H^1(\Omega)$, $u_1\in L^2(0,T;H^2(\Omega))$ *and* $u_2\in L^2(0,T;H^1(\Omega))$, *then we have the following estimate :*

$$\|u_2 - u_1\|_{0,\Omega} \le e^{L(t)}\|u_{2,0}-u_{1,0}\|_{0,\Omega} + e^{L(t)}\int_0^t e^{L(s)}\Big[|D(s)-D|\|\Delta u_1\|_{0,\Omega} + |\Omega||r(s)-r|\Big]ds. \qquad (3.1)$$

The area of the wound at the instant t, defined by the formula (2.5), yields

$$\begin{aligned} |W_2(t) - W_1(t)| &\leq \int_\Omega |u_2 - u_1| dx \\ &\leq \sqrt{|\Omega|} \|u_2 - u_1\|_{0,\Omega}, \end{aligned}$$

where $|\Omega|$ *denotes the measure of* Ω*.*

3.2.2. Estimate of $\|\Delta u_1\|_{0,\Omega}$

We are now seeking to estimate $\|\Delta u_1\|_{0,\Omega}$ as a function of the data $u_{1,0}$, D and r. In order to be reduced to a homogeneous problem, we make the following change of unknown : $w_1 = 1 - u_1$ in the problem (P_1). The new unknown w_1 is then the solution to the following problem :

$$(P_4) \begin{cases} \frac{\partial w_1}{\partial t}(x,y,t) - D\Delta w_1(x,y,t) + r w_1(1-w_1) = 0, \quad (x,y,t) \in \Omega \times]0,T[, \\ w_1(x,y,0) = w_0(x,y) = 1 - u_{1,0}, \quad (x,y) \in \overline{\Omega}, \\ w_1(x,y,t) = 0, \quad \text{dans} \quad \Gamma_V \times [0,T], \\ \frac{\partial w_1}{\partial n}(x,y,t) = 0, \quad \text{dans} \quad \Gamma_H \times [0,T]. \end{cases}$$

Lemme 3.2 *Suppose that* $u_{1,0} \in H^1(\Omega)$ *and* $u_1 \in L^2(0,T;H^2(\Omega))$*, we get the following estimate*

$$\|\Delta u_1\|^2_{0,\Omega} \leq 2 \sum_{n=1}^{+\infty} \lambda_n^2 \Big(\|1 - u_{1,0}\|^2_{0,\Omega} e^{-2D\lambda_n t} + r^2 |\Omega| t \int_0^t e^{-2D\lambda_n (t-s)} ds \Big), \quad (3.2)$$

where $(\lambda_n)_{n \in \mathbb{N}^*}$ *are eigenvalues of the following eigenvalue problem :*

$$(P_5) \begin{cases} -\Delta v = \lambda v, \quad \textit{in} \quad \Omega, \\ v = 0, \quad \textit{on} \quad \Gamma_V, \\ \frac{\partial v}{\partial n} = 0, \quad \textit{on} \quad \Gamma_H. \end{cases}$$

4. Conclusion and perspectives

In order to model cellular dynamics, a simple model of Fisher-KPP type, considering only the biological effect, was considered in the first step. The comparison of numerical results obtained, in the case where the proliferation and diffusion parameters are constant, with experimental results shows the insufficiency of Fisher-KPP model to accurately represent the activated and inhibited dynamics. Nevertheless, the activation and inhibition operations (D and r time-dependent) provide more effective results, which is coherent with the two estimations (3.1) and (3.2).

In order to better model the cellular dynamics, a coupled model is suggested. It consists at the Fisher-KPP equation coupled with the mechanics equation ; the behavior being purely elastic. Moreover, the numerical implementation of such coupled model is in progress.

5. Bibliographie

[1] MALANDAIN G.,« https ://interstices.info/jcms/$i_5$3813/les-mathematiques-cachees-de-la-medecine. 21/05/2010. »

[2] PAGE KAREN M., MAINI PHILIP K., A.M. MONK. NICHOLAS, « Complex pattern formation in reaction-diffusion systems with spatially varying parameters », *Physica*, vol. 202, n° , 2005.

[3] OLSEN L., MAINI P.K., SHERRATT J.A., « Spatially Varying Equilibria of Mechanical Models : Application to Dermal Wound Contraction », *Mathematical Bisciences*,vol. 147, n° 113, 1998.

[4] FENTEANY G., JANMEY P. A., STOSSEL T. P., « Signaling pathways and cell mechanics involved in wound closure by epithelia cell sheets », *Current Biology*, vol. 10, n° 831, 2000.

[5] BAO QI., HUGHES R.C., « Galectin-3 and polarized growth within collagen gels of wild-type and ricin-resistant mdck renal epithelial cells », *Glycobiology*, vol. 9, n° 5, 1999.

[6] QINGHUI MENG, JAMES M. MASON, DEBRA PORTI, ITZHAK D. GOLDBERG, ELIOT M. ROSEN, SAIJUN FAN, « Hepatocyte growth factor decreases sensitivity to chemotherapeutic agents and stimulates cell adhesion, invasion, and migration », *Biochem. Biophys. Res. Commun.*, vol. 274, n° 772, 2000.

[7] MURRAY J.D., « Mathematical Biology : I. An Introduction, Third Edition », *Springer,Interdisciplinary Applied Mathematics*, vol. 17, 2001.

[8] MURRAY J.D. , « Mathematical Biology : II. Spatial Models and Biomedical Applications », *Springer,Interdisciplinary Applied Mathematics*, vol. 18, 2011.

[9] GAFFNEY EAMONN A., MAINI PHILIP K., SHERRATT JONATHAN A., DALE PAUL D., « Wound healing in the corneal epithelium : Biological mechanisms and mathematical models », *J. Theor. Med.*, vol. 1, n° 13,1997.

[10] MAINI P.K., OLSEN L., SHERRATT J.A., « Mathematical models for cell-matrix interactions during dermal wound healing », *Int. J. Bifurcation Chaos Appl. Sci. Eng.*, vol. 12, n° 9, 2002.

[11] PAGE KAREN M., MAINI PHILIP K., MONK NICHOLAS A.M., « Complex pattern formation in reactionÐdiffusion systems with spatially varying parameters »,*Physica D*, vol. 202, n° 95, 2005.

[12] VEDULA S. R. K., LEONG M. C., LAI T. L. , HERSEN P., KABLA A. J., LIM C.T., LADOUX B., « Emerging modes of collective cell migration induced by geometrical constraints », *PNAS*, vol. 109, 2012.

[13] LEE P., WOLGEMUTH C. W., « Crawling Cells Can Close Wounds without Purse Strings or Signaling », *PLoS Computational Biology*, vol. 7, 2011.

[14] SAEZ A., ANON E., GHIBAUDO M., O DU ROURE, MEGLIO J-M DI., HERSEN P., SILBERZAN P., BUGUIN A., LADOUX B., « Traction forces exerted by epithelial cell sheets », *J. Phys. : Condens. Matter*, vol. 22, n° 9, 2010.

[15] KABLA A. J., « Collective cell migration : leadership, invasion and segregation », *J. R. Soc. Interface*, vol. 9, 2012.

[16] OLSEN L., MAINI P.K., SHERRATT J.A., « A mechanochemical model for normal and abnormal dermal wound repair », *Nonlinear Analysis, Theory, Methods & Applications*, vol. 30, n° 6, 1997.

[17] OLSEN L., MAINI P.K., SHERRATT J.A., « A Mechanochemical Model for Adult Dermal Wound Contraction and the Permanence of the Contracted Tissue Displacement Profile », *J. theor. Biol.*, vol. 177, 1995.

[18] PERELSON A. S., MAINI P. K., MURRAY J. D., HYMAN J. M. , OSTER G. F.G. F., « Nonlinear pattern selection in a mechanical model for morphogenesis », *Journal of Mathematical*

biology. Springer Veriag, vol. 24, n° 525, 1986.

[19] SHERRATT J. A., « Actin aggregation and embryonic epidermal wound healing », *Journal of Mathematical biology. Springer Veriag*, vol. 31, n° 703, 1993.

[20] GOTO Y., « A 2-dimensional mechanical model of the formation of a somite », *International journal of numerical analysis and modeling*, vol. 10, n° 1, 2013.

[21] HABBAL A., BARELLI H., MALANDAIN G., « Assessing the ability of the 2D Fisher-KPP equation to model cell-sheet would clousure », *Mathematical Biosciences*, vol. 252, n° 45, 2014.

[22] HANSEN C. M. , « Polymer science applied to biological problems : prediction to cytotoxic drug interactions with DNA », *European Polymer Journal*, vol. 44, 2008.

[23] HAMEL É. , « Modélisation mathématique de la déxpression synaptique et des périodes réfractaires pour le quantron », 2013.

[24] YEGANEFAR N., « Définitions et analyse de stabilité pour les systèmes à retard non linéaires », Novembre 2006.

[25] BREZIS H., « Analyse fonctionnelle : Théorie et applications », *Dunod*, 1999.

[26] RAVIART P.A., THOMAS J.M., « Introduction à l'analyse numérique des équations aux dérivées partielles », *Masson*, 1983.

Global weak solution to a 3-D Kazhikhov-Smagulov model with Korteweg stress

Caterina Calgaro[a] — Meriem Ezzoug[b,*] — Ezzeddine Zahrouni[b,c]

[a] Université Lille
CNRS, UMR 8524
Laboratoire Paul Painlevé
F-59000 Lille
France
caterina.calgaro@math.univ-lille1.fr

[b] Unité de Recherche : Multifractals et Ondelettes
Faculté des Sciences de Monastir
Université de Monastir
5019 Monastir
Tunisie
meriemezzoug@yahoo.fr

[c] Faculté des Sciences Économiques et de Gestion de Nabeul
Université de Carthage
8000 Nabeul
Tunisie
zahrouniez@gmail.com

[*] Corresponding author

ABSTRACT. In this article, we consider a multiphasic incompressible fluid model, called the Kazhikhov-Smagulov model, with a specific stress tensor which depends on density derivatives, introduced by Korteweg. We establish the existence of global weak solution to this model in a 3D bounded domain.

RÉSUMÉ. Dans cet article, nous considérons un modèle de fluide incompressible multiphasique, appelé modèle de Kazhikhov-Smagulov, avec un tenseur de contraintes spécifique qui dépend des dérivées d'ordre élevé de la densité, introduit par Korteweg. Nous établissons l'existence d'une solution faible globale pour ce modèle dans un domaine borné en 3D.

KEYWORDS : Kazhikhov-Smagulov model, Korteweg model, weak solution, global existence result.

MOTS-CLÉS : Modèle de Kazhikhov-Smagulov, modèle de Korteweg, solution faible, existence globale.

1. Introduction

We are concerned with systems of PDEs describing the evolution of mixture flows. Let Ω be a bounded open set in $\mathbb{R}^3$ with boundary Γ that is regular enough and let $\boldsymbol{n}$ be the outwards unit normal on the boundary Γ. We denote by $[0,T]$ the time interval, for $T>0$. The mixture of two fluids is described by the density $\rho(t,\boldsymbol{x}) \geq 0$, the velocity field $\boldsymbol{v}(t,\boldsymbol{x}) \in \mathbb{R}^3$ and the pressure $p(t,\boldsymbol{x})$, depending on the time and space variables $(t,\boldsymbol{x}) \in [0,T]\times\Omega$. According to [4, 7, 8], we consider the Korteweg equations for generalized incompressible fluids whose density and volume change with the concentration $\phi(t,\boldsymbol{x}) \geq 0$ and eventually the temperature, but not with pressure. In general, the velocity field $\boldsymbol{v}$ of such incompressible fluids is not solenoidal, $\operatorname{div}\boldsymbol{v} \neq 0$. Assuming that each fluid is incompressible, the mass density is conserved in the absence of diffusion. The theory of Korteweg, introduced in [9], considers the possibility that stresses are induced by gradients of concentration and density in a slow process of diffusion of incompressible miscible liquids. Such stresses could be important in regions of high gradients and they mimic the surface tension.

In order to model the fluid capillarity effects, Korteweg introduced in the usual compressible fluid model a specific stress tensor which depends on density derivatives. Following the rigorous formulation presented in [4] (see also [2]) and neglecting thermal fluctuations, the model reads

$$\begin{cases} \partial_t\rho + \operatorname{div}(\rho\boldsymbol{v}) = 0, \\ \partial_t(\rho\boldsymbol{v}) + \operatorname{div}(\rho\boldsymbol{v}\otimes\boldsymbol{v}) = \rho\boldsymbol{g} + \operatorname{div}(\boldsymbol{S}+\boldsymbol{K}), \end{cases} \tag{1}$$

where $\boldsymbol{g}$ stands for the gravity acceleration (but it can include further external forces). The viscous stress tensor $\boldsymbol{S}$ and the Korteweg stress tensor $\boldsymbol{K}$ are given by :

$$\begin{cases} \boldsymbol{S} &= (\nu\operatorname{div}\boldsymbol{v} - p)\boldsymbol{I} + 2\mu\boldsymbol{D}(\boldsymbol{v}), \\ \boldsymbol{K} &= (\alpha\Delta\rho + \beta|\nabla\rho|^2)\boldsymbol{I} + \delta(\nabla\rho\otimes\nabla\rho) + \gamma D_x^2\rho, \end{cases} \tag{2}$$

where $\boldsymbol{D}(\boldsymbol{v}) = (\nabla\boldsymbol{v} + \nabla\boldsymbol{v}^T)/2$ is the strain tensor and $D_x^2\rho$ is the hessian matrix of the density ρ. Here, the pressure p and the coefficients $\alpha,\beta,\gamma,\delta,\mu$ and ν are functions of ρ. The special case

$$\alpha = \kappa\rho, \quad \beta = \frac{\kappa}{2}, \quad \delta = -\kappa, \quad \gamma = 0,$$

for some constant $\kappa > 0$, corresponds precisely to Korteweg's original assumptions connected with the variational theory of Van Der Waals. In this case, the Korteweg stress tensor yields

$$\boldsymbol{K} = \frac{\kappa}{2}(\Delta\rho^2 - |\nabla\rho|^2)\boldsymbol{I} - \kappa(\nabla\rho\otimes\nabla\rho). \tag{3}$$

Writing

$$\operatorname{div}\boldsymbol{K} = \kappa\rho\nabla(\Delta\rho) = \kappa\nabla(\rho\Delta\rho) - \kappa\nabla\rho\Delta\rho, \tag{4}$$

and incorporating $\nabla(\rho\Delta\rho)$ in the pressure term, we obtain $-\kappa\nabla\rho\Delta\rho$ as a right hand side term in the momentum equation.

The Korteweg's theory can be applied to processes of slow diffusion on miscible incompressible fluids, for example, water and glycerin. The two fluids are characterized by their reference mass density : $\bar{\rho}_1$ the density of the dilute phase and $\bar{\rho}_2$ the density of the dense phase. We need the velocity field of each constituent : $\boldsymbol{v}_1(t,\boldsymbol{x})$ and $\boldsymbol{v}_2(t,\boldsymbol{x})$, respectively. We define the volume fraction of the dilute phase $0 \leq \phi(t,\boldsymbol{x}) \leq 1$:

$$\phi(t,\boldsymbol{x}) = \lim_{r\to 0} \frac{\text{Volume occupied at time } t \text{ by the dilute phase in } B(\boldsymbol{x},r)}{|B(\boldsymbol{x},r)|}.$$

Then, admitting that each fluid is incompressible and keeping a constant mass density, the density of the mixture is defined by

$$\rho(t,\boldsymbol{x}) = \underbrace{\bar{\rho}_2\big(1-\phi(t,\boldsymbol{x})\big)}_{:=\rho_2(t,\boldsymbol{x})} + \underbrace{\bar{\rho}_1\phi(t,\boldsymbol{x})}_{:=\rho_1(t,\boldsymbol{x})} = \bar{\rho}_2 + (\bar{\rho}_1-\bar{\rho}_2)\phi(t,\boldsymbol{x}).$$

Writing the mass conservation for the two phases, we obtain

$$\partial_t\rho + \operatorname{div}(\rho\boldsymbol{v}) = 0,$$

with $\rho\boldsymbol{v}(t,\boldsymbol{x}) = (\rho_2\boldsymbol{v}_2 + \rho_1\boldsymbol{v}_1)(t,\boldsymbol{x})$ presents the *mean mass velocity* $\boldsymbol{v}(t,\boldsymbol{x})$, which is not divergence free, $\operatorname{div}\boldsymbol{v} \neq 0$. Moreover, we define the *mean volume velocity*

$$\boldsymbol{u}(t,\boldsymbol{x}) = \big(1-\phi(t,\boldsymbol{x})\big)\boldsymbol{v}_2(t,\boldsymbol{x}) + \phi(t,\boldsymbol{x})\boldsymbol{v}_1(t,\boldsymbol{x}).$$

Applying the definitions, we verify that the velocity field $\boldsymbol{u}$ is solenoidal ($\operatorname{div}\boldsymbol{u} = 0$). According to Kazhikhov and Smagulov [10], we consider the following non-standard constraint associated to the pressure p :

$$\operatorname{div}\boldsymbol{v} = -\operatorname{div}(\lambda\nabla\ln(\rho)), \tag{5}$$

where $\lambda > 0$ is a diffusion coefficient. This Fick's law (5) describes the diffusive fluxes of one fluid into the other, see also [3]. Obviously, when we set

$$\boldsymbol{v} = \boldsymbol{u} - \lambda\nabla\ln(\rho), \tag{6}$$

the relation yields (5). The mixture density ρ verifies the mass conservation and we obtain

$$\partial_t\rho + \operatorname{div}(\rho\boldsymbol{u}) = \operatorname{div}(\lambda\nabla\rho). \tag{7}$$

For the momentum equation $(1)_2$, we start by developing each term using the relation (6), in order to eliminate $\boldsymbol{v}$. After some calculations and using (4), we get

$$\begin{aligned}\partial_t\big(\rho\boldsymbol{u}\big) \;&+\; \operatorname{div}\big(\rho\boldsymbol{u}\otimes\boldsymbol{u}\big) - \lambda\operatorname{div}\big(\nabla\rho\otimes\boldsymbol{u}\big) - \lambda\operatorname{div}\big(\boldsymbol{u}\otimes\nabla\rho\big)\\ &+\; \lambda\nabla\big(\boldsymbol{u}\cdot\nabla\rho\big) + \lambda\operatorname{div}\big(2\mu D_x^2\ln(\rho)\big) - \operatorname{div}\big(2\mu\boldsymbol{D}(\boldsymbol{u})\big) + \nabla p\\ &-\; \lambda^2\Big(\nabla\Delta\rho - \operatorname{div}\big(\frac{\nabla\rho\otimes\nabla\rho}{\rho}\big)\Big) = \rho\boldsymbol{g} + \kappa\nabla(\rho\Delta\rho) - \kappa\nabla\rho\Delta\rho.\end{aligned} \tag{8}$$

Choosing the dynamic viscosity μ constant, as in [6], we have $\operatorname{div}\big(2\mu\boldsymbol{D}(\boldsymbol{u})\big) = \mu\boldsymbol{\Delta}\boldsymbol{u}$ and $\operatorname{div}\big(2\mu D_x^2\ln(\rho)\big) = 2\mu\nabla\Delta\ln(\rho)$. Including all the gradient terms in the modified pressure

$$P = p + \lambda(\nu+2\mu)\Delta\ln(\rho) + \lambda\boldsymbol{u}\cdot\nabla\rho - \lambda^2\Delta\rho - \kappa\rho\Delta\rho.$$

Then, we obtain the Kazhikhov-Smagulov-Korteweg model in conservative form :

$$\begin{cases}\partial_t\big(\rho\boldsymbol{u}\big) + \operatorname{div}\big(\rho\boldsymbol{u}\otimes\boldsymbol{u}\big) - \lambda\operatorname{div}\big(\nabla\rho\otimes\boldsymbol{u}\big) - \lambda\operatorname{div}\big(\boldsymbol{u}\otimes\nabla\rho\big) - \mu\boldsymbol{\Delta}\boldsymbol{u}\\ \qquad\qquad +\nabla P + \lambda^2\operatorname{div}\big(\frac{\nabla\rho\otimes\nabla\rho}{\rho}\big) = \rho\boldsymbol{g} - \kappa\nabla\rho\Delta\rho,\\ \partial_t\rho + \operatorname{div}\big(\rho\boldsymbol{u}\big) = \lambda\Delta\rho,\\ \operatorname{div}\boldsymbol{u} = 0.\end{cases} \tag{9}$$

The tensorial product matrix of two vectors $\boldsymbol{a} = (a_i)_{i=1}^d$, $\boldsymbol{b} = (b_i)_{i=1}^d$ is denoted by $\boldsymbol{a} \otimes \boldsymbol{b}$ with cœfficients $(\boldsymbol{a} \otimes \boldsymbol{b})_{i,j} = a_i b_j$. Taking into account the equalities

$$\begin{aligned} \partial_t\big(\rho \boldsymbol{u}\big) + \operatorname{div}\big(\rho \boldsymbol{u} \otimes \boldsymbol{u}\big) - \lambda \operatorname{div}\big(\nabla\rho \otimes \boldsymbol{u}\big) &= \rho\partial_t \boldsymbol{u} + \rho(\boldsymbol{u}\cdot\nabla)\boldsymbol{u} - \lambda(\nabla\rho\cdot\nabla)\boldsymbol{u}, \\ -\lambda \operatorname{div}\big(\boldsymbol{u} \otimes \nabla\rho\big) = -\lambda(\boldsymbol{u}\cdot\nabla)\nabla\rho &= -\lambda\nabla(\boldsymbol{u}\cdot\nabla\rho) + \lambda \operatorname{div}\big(\rho\nabla\boldsymbol{u}^T\big). \end{aligned}$$

Then, denoting $\mathcal{Q}_T = (0,T)\times\Omega$, $\Sigma = (0,T)\times\Gamma$, the Kazhikhov-Smagulov-Korteweg (KSK) model can be written in $\mathcal{Q}_T$ as :

$$\left\{ \begin{array}{l} \rho\big(\partial_t \boldsymbol{u} + (\boldsymbol{u}\cdot\nabla)\boldsymbol{u}\big) - \lambda(\nabla\rho\cdot\nabla)\boldsymbol{u} + \lambda \operatorname{div}\big(\rho\nabla\boldsymbol{u}^T\big) - \mu\boldsymbol{\Delta}\boldsymbol{u} + \nabla P \\ \qquad\qquad +\lambda^2 \operatorname{div}\big(\dfrac{\nabla\rho\otimes\nabla\rho}{\rho}\big) = \rho\boldsymbol{g} - \kappa\Delta\rho\nabla\rho, \\ \partial_t\rho + \operatorname{div}\big(\rho\boldsymbol{u}\big) = \lambda\Delta\rho, \\ \operatorname{div}\boldsymbol{u} = 0. \end{array} \right. \tag{10}$$

The KSK model (10) is completed by the following boundary and initial conditions

$$\boldsymbol{u}(t,\boldsymbol{x}) = 0, \qquad \frac{\partial\rho}{\partial\boldsymbol{n}}(t,\boldsymbol{x}) = 0, \qquad (t,\boldsymbol{x}) \in \Sigma, \tag{11}$$

$$\boldsymbol{u}(0,\boldsymbol{x}) = \boldsymbol{u}_0(\boldsymbol{x}), \qquad \rho(0,\boldsymbol{x}) = \rho_0(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega, \tag{12}$$

with the compatibility condition $\operatorname{div}\boldsymbol{u}_0 = 0$, where $\rho_0 : \Omega \to \mathbb{R}$ and $\boldsymbol{u}_0 : \Omega \to \mathbb{R}^3$ are given functions. Throughout this work, we assume the hypothesis

$$0 < m \le \rho_0(\boldsymbol{x}) \le M < +\infty, \quad \boldsymbol{x} \in \Omega. \tag{13}$$

The paper is organized as follows. In Section 2 we present the main results about (10). After some preliminary results recalled in Section 3, the proof of existence of global weak solution for (10) is given in Section 4. The conclusions are summarized in Section 5.

2. Functional setup and main results

Let us introduce the following functional spaces (see [11, 13] for their properties):

$$\begin{aligned} \mathcal{V} &= \big\{\boldsymbol{u} \in \mathcal{D}(\Omega)^3 :\ \operatorname{div}\boldsymbol{u} = 0 \text{ in } \Omega\big\}, \\ \mathbf{V} &= \big\{\boldsymbol{u} \in \mathbf{H}_0^1(\Omega) :\ \operatorname{div}\boldsymbol{u} = 0 \text{ in } \Omega\big\}, \\ \mathbf{H} &= \big\{\boldsymbol{u} \in \mathbf{L}^2(\Omega) :\ \operatorname{div}\boldsymbol{u} = 0 \text{ in } \Omega,\ \boldsymbol{u}\cdot\boldsymbol{n} = 0 \text{ on } \Gamma\big\}, \\ H_N^s &= \left\{\rho \in H^s(\Omega) :\ \frac{\partial\rho}{\partial\boldsymbol{n}} = 0 \text{ on } \Gamma,\ \int_\Omega \rho(\boldsymbol{x})\,d\boldsymbol{x} = \int_\Omega \rho_0(\boldsymbol{x})\,d\boldsymbol{x}\right\}, \quad s \ge 2. \end{aligned}$$

The spaces $\mathbf{V}$ and $\mathbf{H}$ are the closures of $\mathcal{V}$ in $\mathbf{H}_0^1(\Omega)$ and $\mathbf{L}^2(\Omega)$, respectively.

Let us recall the definition of weak solution for the KSK model (10). Such class of solutions can be found in [1] for Kazhikhov-Smagulov type models and in [13] for the incompressible Navier-Stokes equations.

Definition 2.1 *A pair of functions $(\boldsymbol{u},\rho)$ is called a weak solution of problem (10),(11),(12) on Ω if and only if the following assumptions are satisfied :*

1) $\boldsymbol{u} \in L^\infty\big(0,T;\boldsymbol{H}\big) \cap L^2\big(0,T;\boldsymbol{V}\big)$, $\rho \in L^\infty\big(0,T;H^1(\Omega)\big) \cap L^2\big(0,T;H^2_N\big)$ *and*

$$0 < m \leq \rho(t,\boldsymbol{x}) \leq M < +\infty, \text{ a.e. } (t,\boldsymbol{x}) \in \mathcal{Q}_{_T}.$$

2) For all $\boldsymbol{\phi} \in \mathcal{C}^1\big([0,T];\boldsymbol{V}\big)$ *such that* $\boldsymbol{\phi}(T,.) = 0$*, one has :*

$$\begin{aligned}&\int_0^T \Big\{ -\big(\boldsymbol{u}, \rho\partial_t\boldsymbol{\phi} + \big((\rho\boldsymbol{u} - \lambda\nabla\rho)\cdot\nabla\big)\boldsymbol{\phi}\big) + \mu\big(\nabla\boldsymbol{u}, \nabla\boldsymbol{\phi}\big) - \lambda\big(\rho\nabla\boldsymbol{u}^T, \nabla\boldsymbol{\phi}\big)\Big\}dt \\ &-\lambda^2\int_0^T \big(\frac{1}{\rho}\nabla\rho\otimes\nabla\rho, \nabla\boldsymbol{\phi}\big)dt = \int_0^T \big(\rho\boldsymbol{g} - \kappa\Delta\rho\nabla\rho, \boldsymbol{\phi}\big)dt + \big(\rho_0\boldsymbol{u}_0, \boldsymbol{\phi}(0)\big).\end{aligned} \tag{14}$$

3) For all $\varphi \in \mathcal{C}^1\big([0,T];H^1(\Omega)\big)$ *such that* $\varphi(T,.) = 0$*, one has :*

$$\int_0^T \Big\{\big(\boldsymbol{u}\cdot\nabla\rho, \varphi\big) + \lambda\big(\nabla\rho, \nabla\varphi\big) - \big(\rho, \partial_t\varphi\big)\Big\}dt = \big(\rho_0, \varphi(0)\big). \tag{15}$$

REMARK. — The pressure P associated with the weak solution $(\boldsymbol{u}, \rho)$ can be obtained using (14) and the Rham's lemma [13].

We present the aim of this work about the Kazhikhov-Smagulov-Korteweg model (10). Under some assumption on the coefficients λ, μ, κ, we prove the global existence of weak solution of (10) for arbitrary initial data and external force field. Our main result reads :

Theorem 2.2 *Let* $\boldsymbol{u}_0 \in \boldsymbol{H}$, $\rho_0 \in H^1(\Omega)$ *satisfy (13),* $T > 0$ *and* $\boldsymbol{g} \in L^2\big(0,T;\boldsymbol{L}^2(\Omega)\big)$*. If* $\frac{\lambda}{\mu}\max(1, \frac{\lambda^2}{\kappa})$ *is sufficiently small, then there exists a weak solution* $(\boldsymbol{u}, \rho)$ *of (10) global in time such that*

$$\begin{aligned}&\boldsymbol{u} \in L^\infty\big(0,T;\boldsymbol{H}\big) \cap L^2\big(0,T;\boldsymbol{V}\big), \\ &\rho \in L^\infty\big(0,T;H^1(\Omega)\big) \cap L^2\big(0,T;H^2_N\big),\end{aligned}$$

with finite and uniformly bounded energy such that $\forall t \leq T$,

$$\begin{aligned}&\| \sqrt{\rho(t)}\boldsymbol{u}(t) \|^2_{L^2(\Omega)} + \kappa \| \nabla\rho(t) \|^2_{L^2(\Omega)} + \int_0^t \Big(\frac{\mu}{2} \| \nabla\boldsymbol{u}(s) \|^2_{L^2(\Omega)} + \kappa\lambda \| \Delta\rho(s) \|^2_{L^2(\Omega)} \Big)ds \\ &\leq \| \sqrt{\rho_0}\boldsymbol{u}_0 \|^2_{L^2(\Omega)} + \kappa \| \nabla\rho_0 \|^2_{L^2(\Omega)} + \frac{CM^2}{\mu}\int_0^T \| \boldsymbol{g}(s) \|^2_{L^2(\Omega)}\, ds.\end{aligned}$$

3. Preliminary results

Given the initial density ρ_0 and the velocity field $\boldsymbol{u}$, we find the density ρ as solution of the following Neumann problem :

$$\begin{cases} \partial_t\rho + \boldsymbol{u}\cdot\nabla\rho = \lambda\,\Delta\rho & \text{in } \mathcal{Q}_{_T}, \\ \rho(0,\boldsymbol{x}) = \rho_0(\boldsymbol{x}) & \text{in } \Omega, \\ \dfrac{\partial\rho}{\partial\boldsymbol{n}} = 0 & \text{on } \Sigma. \end{cases} \tag{16}$$

The density ρ satisfies the maximum principle. This result is classical (see [1]).

Proposition 3.1 *If $(\boldsymbol{u}, \rho)$ is a weak solution of (10), then*

$$0 < m \leq \rho(t, \boldsymbol{x}) \leq M < +\infty \qquad \textit{a.e. } (t, \boldsymbol{x}) \in \mathcal{Q}_T. \tag{17}$$

Proposition 3.2 *Let $\rho_0 \in H^1(\Omega)$ verify (13) and $\boldsymbol{u} \in \mathcal{C}\big([0,T]; \boldsymbol{V} \cap \boldsymbol{H}^2(\Omega)\big)$. Then there exists a unique solution ρ of (16) such that*

$$\rho \in L^\infty\big(0,T; H^1(\Omega)\big) \cap L^2\big(0,T; H^2_N\big).$$

Moreover, we have

$$\sup_{0\leq t\leq T} \| \rho(t) \|^2_{L^2(\Omega)} \leq \| \rho_0 \|^2_{H^1(\Omega)}, \tag{18}$$

$$\int_0^T \| \nabla\rho(t) \|^2_{L^2(\Omega)} \, dt \leq \frac{1}{2\lambda} \| \rho_0 \|^2_{H^1(\Omega)}, \tag{19}$$

$$\sup_{0\leq t\leq T} \| \nabla\rho(t) \|^2_{L^2(\Omega)} \leq C_\lambda \| \rho_0 \|^2_{H^1(\Omega)} \big(1 + \sup_{0\leq t\leq T} \| \boldsymbol{u}(t) \|^2_{L^\infty(\Omega)}\big), \tag{20}$$

$$\int_0^T \| \Delta\rho(t) \|^2_{L^2(\Omega)} \, dt \leq \frac{C_\lambda}{\lambda} \| \rho_0 \|^2_{H^1(\Omega)} \big(1 + \sup_{0\leq t\leq T} \| \boldsymbol{u}(t) \|^2_{L^\infty(\Omega)}\big), \tag{21}$$

where C_λ is a positive constant depending only on λ.

Given $\rho_0 \in H^1(\Omega)$ satisfying (13) and $\boldsymbol{u} \in \mathcal{C}\big([0,T]; \mathbf{V} \cap \mathbf{H}^2(\Omega)\big)$, let ρ the solution obtained by Proposition 3.2. Therefore, it is clear that the following map is well defined

$$\mathcal{S} : \mathcal{C}\big([0,T]; \mathbf{V} \cap \mathbf{H}^2(\Omega)\big) \longrightarrow L^\infty\big(0,T; H^1(\Omega)\big) \cap L^2\big(0,T; H^2_N\big),$$

such that $\rho = \mathcal{S}\boldsymbol{u}$ is well defined.

Proposition 3.3 *Let $\rho_0 \in H^1(\Omega)$ verify (13) and $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathcal{C}\big([0,T]; \boldsymbol{V} \cap \boldsymbol{H}^2(\Omega)\big)$. Set $\rho = \rho_1 - \rho_2 = \mathcal{S}\boldsymbol{u}_1 - \mathcal{S}\boldsymbol{u}_2$ and $\boldsymbol{u} = \boldsymbol{u}_1 - \boldsymbol{u}_2$, we have the following estimates :*

$$\sup_{0\leq t\leq T} \| \rho(t) \|^2_{L^2(\Omega)} + \lambda \int_0^T \| \nabla\rho(t) \|^2_{L^2(\Omega)} \, dt \leq \frac{M^2}{\lambda} T \sup_{0\leq t\leq T} \| \boldsymbol{u}(t) \|^2_{L^2(\Omega)}, \tag{22}$$

$$\begin{aligned} &\sup_{0\leq t\leq T} \| \nabla\rho(t) \|^2_{L^2(\Omega)} + \lambda \int_0^T \| \Delta\rho(t) \|^2_{L^2(\Omega)} \, dt \\ &\leq \frac{2T}{\lambda} \sup_{0\leq t\leq T} \| \nabla\rho_1 \|^2_{L^2(\Omega)} \sup_{0\leq t\leq T} \| \boldsymbol{u} \|^2_{L^\infty(\Omega)} + \frac{2M^2T}{\lambda^3} \sup_{0\leq t\leq T} \| \boldsymbol{u}_2 \|^2_{L^\infty(\Omega)} \sup_{0\leq t\leq T} \| \boldsymbol{u} \|^2_{L^2(\Omega)}. \end{aligned} \tag{23}$$

We recall that there exists an orthonormal basis of $\mathbf{L}^2(\Omega)$ defined by

$$\begin{aligned} &\boldsymbol{\omega}_k \in \mathbf{V} \cap \mathbf{H}^2(\Omega) \\ &-\mathbb{P}\boldsymbol{\Delta}\boldsymbol{\omega}_k = \lambda_k \, \boldsymbol{\omega}_k \qquad \text{on } \Omega, \end{aligned}$$

where $\mathbb{P}$ is the orthogonal projection operator of $\mathbf{L}^2(\Omega)$ onto $\mathbf{H}$. For any $n \in \mathbb{N}^*$, we define by $\mathbf{X_n}$ the finite dimensional subspace of $\mathbf{H}$ such that

$$\mathbf{X_n} = \mathcal{V}\text{ect}\{\boldsymbol{\omega}_k, k = 1, \ldots, n\},$$

and we consider the orthogonal projection $\mathbb{P}_n : \mathbf{L}^2(\Omega) \to \mathbf{X_n}$ defined by

$$\forall \boldsymbol{w} \in \mathbf{H}, \qquad \big(\mathbb{P}_n \boldsymbol{w},\ \boldsymbol{v}\big) \ = \ \big(\boldsymbol{w},\ \boldsymbol{v}\big), \qquad \forall \boldsymbol{v} \in \mathbf{X_n}. \tag{24}$$

As in [5], we introduce a family of operators $\mathcal{M}[\rho] : \mathbf{X_n} \longrightarrow \mathbf{X_n}$ defined by

$$\big(\mathcal{M}[\rho]\boldsymbol{v},\ \boldsymbol{\omega}\big) \ = \ \int_\Omega \rho\ \boldsymbol{v} \cdot \boldsymbol{\omega}\ d\boldsymbol{x} \qquad \text{for all } \boldsymbol{v}, \boldsymbol{\omega} \in \mathbf{X_n}. \tag{25}$$

If $\rho \in L^\infty(\Omega)$, then $\mathcal{M}[\rho]$ is well defined. Moreover, let $m > 0$, we set

$$\mathcal{D} \ = \ \Big\{ \rho \in L^\infty(\Omega);\ \ \rho(\boldsymbol{x}) \geq m > 0 \Big\}.$$

Proposition 3.4 *$\mathcal{M}[\rho]$ is one-to-one and its inverse verifies*

$$\| \ \mathcal{M}[\rho]^{-1} \ \|_{\mathcal{L}(\mathbf{X_n}, \mathbf{X_n})} \ \leq \ \big(\inf_{\boldsymbol{x} \in \Omega} \rho(\boldsymbol{x}) \big)^{-1} \qquad \forall \rho \in \mathcal{D}, \tag{26}$$

$$\| \ \mathcal{M}[\rho_1]^{-1} - \mathcal{M}[\rho_2]^{-1} \ \|_{\mathcal{L}(\mathbf{X_n}, \mathbf{X_n})} \ \leq \ \frac{C_n}{m^2} \ \| \ \rho_1 - \rho_2 \ \|_{L^2(\Omega)} \quad \forall \rho_1, \rho_2 \in \mathcal{D}, \tag{27}$$

where C_n is a constant depending on the dimension of the space $\mathbf{X_n}$.

4. Proof of Theorem 2.2

4.1. Faedo-Galerkin method

We are looking for the approximate solutions

$$(\boldsymbol{u}_n, \rho_n) \in \mathcal{C}\big([0,T]; \mathbf{X_n}\big) \times \mathcal{C}\big([0,T]; H^1(\Omega) \cap H^2_N\big)$$

satisfying

$$\begin{cases} \displaystyle\int_\Omega \partial_t\big(\rho_n \boldsymbol{u}_n\big) \cdot \boldsymbol{v} d\boldsymbol{x} + \int_\Omega \rho_n (\boldsymbol{u}_n \cdot \nabla)\boldsymbol{u}_n \cdot \boldsymbol{v} d\boldsymbol{x} - \lambda \int_\Omega (\nabla \rho_n \cdot \nabla)\boldsymbol{u}_n \cdot \boldsymbol{v} d\boldsymbol{x} \\ \displaystyle + \int_\Omega (\boldsymbol{u}_n \cdot \nabla \rho_n)\ \boldsymbol{u}_n \cdot \boldsymbol{v} d\boldsymbol{x} - \lambda \int_\Omega \Delta \rho_n \boldsymbol{u}_n \cdot \boldsymbol{v} d\boldsymbol{x} - \mu \int_\Omega \boldsymbol{\Delta} \boldsymbol{u}_n \cdot \boldsymbol{v} d\boldsymbol{x} \\ \displaystyle + \lambda \int_\Omega \mathrm{div}\ \big(\rho_n \nabla \boldsymbol{u}_n^T\big) \cdot \boldsymbol{v} d\boldsymbol{x} + \lambda^2 \int_\Omega \mathrm{div}\ \Big(\frac{\nabla \rho_n \otimes \nabla \rho_n}{\rho_n}\Big) \cdot \boldsymbol{v} d\boldsymbol{x} \\ \displaystyle = \int_\Omega \rho_n \boldsymbol{g} \cdot \boldsymbol{v} d\boldsymbol{x} - \kappa \int_\Omega \Delta \rho_n \nabla \rho_n \cdot \boldsymbol{v} d\boldsymbol{x}, \qquad \forall \boldsymbol{v} \in \mathbf{X_n}, \\ \displaystyle \int_\Omega \partial_t\big(\rho_n\big)\ \eta\ d\boldsymbol{x} + \int_\Omega \boldsymbol{u}_n \cdot \nabla \rho_n\ \eta\ d\boldsymbol{x} = \lambda \int_\Omega \Delta \rho_n\ \eta\ d\boldsymbol{x}, \qquad \forall \eta \in H^1(\Omega), \\ \boldsymbol{u}_n(0) = \boldsymbol{u}_{0n} = \mathbb{P}_n \boldsymbol{u}_0, \\ \rho_n(0) = \rho_0. \end{cases} \tag{28}$$

We set

$$\begin{aligned} \mathcal{N}[\boldsymbol{u}_n, \rho_n] = &-\big((\rho_n \boldsymbol{u}_n - \lambda \nabla \rho_n) \cdot \nabla\big)\boldsymbol{u}_n - (\boldsymbol{u}_n \cdot \nabla \rho_n)\boldsymbol{u}_n + \lambda \Delta \rho_n \boldsymbol{u}_n \\ &+ \mu \boldsymbol{\Delta} \boldsymbol{u}_n - \lambda\, \mathrm{div}\big(\rho_n \nabla \boldsymbol{u}_n^T\big) - \lambda^2 \mathrm{div}\Big(\frac{\nabla \rho_n \otimes \nabla \rho_n}{\rho_n}\Big) - \kappa \Delta \rho_n \nabla \rho_n + \rho_n \boldsymbol{g}. \end{aligned} \tag{29}$$

Taking $(28)_1$ with $\boldsymbol{v} = \boldsymbol{\omega}_k$, for $k = 1, \ldots, n$, and integrating in time between 0 and $t \leq T$, the solution $\boldsymbol{u}_n$ verifies the following integral equations for $k = 1, \ldots, n$:

$$\int_\Omega \rho_n(t)\boldsymbol{u}_n(t) \cdot \boldsymbol{\omega}_k \, d\boldsymbol{x} \; = \int_\Omega \boldsymbol{q}_0 \cdot \boldsymbol{\omega}_k \, d\boldsymbol{x} \; + \int_0^t \int_\Omega \mathcal{N}[\boldsymbol{u}_n, \rho_n] \cdot \boldsymbol{\omega}_k \, d\boldsymbol{x} \, ds, \tag{30}$$

where $\rho_n = \mathcal{S}\boldsymbol{u}_n$ and $\boldsymbol{q}_0 = \rho_0 \boldsymbol{u}_{0n}$. Thanks to (24) and (25), we rewrite (30) as follows :

$$\Big(\mathcal{M}[\rho_n(t)]\boldsymbol{u}_n(t), \; \boldsymbol{\omega}_k\Big) = \Big(\mathbb{P}_n \boldsymbol{q}_0, \; \boldsymbol{\omega}_k\Big) + \Big(\mathbb{P}_n \int_0^t \mathcal{N}[\boldsymbol{u}_n(s), \rho_n(s)] \, ds, \; \boldsymbol{\omega}_k\Big),$$

for $k = 1, \ldots, n$. Since $\mathcal{M}[\rho_n]$ is invertible, then the resulting equation reads

$$\boldsymbol{u}_n \in \mathcal{C}\big([0,T]; \mathbf{X_n}\big), \quad \boldsymbol{u}_n(t) = \mathcal{M}[\rho_n(t)]^{-1}\mathbb{P}_n\Big(\boldsymbol{q}_0 + \int_0^t \mathcal{N}[\boldsymbol{u}_n(s), \rho_n(s)] \, ds\Big). \tag{31}$$

Hence, $\boldsymbol{u}_n$ appears as a fixed point of a suitable functional Ψ

$$\begin{array}{rccc} \Psi : & \mathcal{C}\big([0,T]; \mathbf{X_n}\big) & \longrightarrow & \mathcal{C}\big([0,T]; \mathbf{X_n}\big) \\ & \boldsymbol{u}_n & \longmapsto & \Psi\big(\boldsymbol{u}_n\big) \end{array}$$

defined by

$$\Psi\big(\boldsymbol{u}_n\big)(t) = \mathcal{M}[\rho_n(t)]^{-1}\mathbb{P}_n\Big(\boldsymbol{q}_0 + \int_0^t \mathcal{N}[\boldsymbol{u}_n(s), \rho_n(s)] \, ds\Big), \qquad \text{for all } t \in [0,T].$$

Let $\mathbf{X_T}$ be the Banach space $\mathcal{C}\big([0,T]; \mathbf{X_n}\big)$ endowed with the norm

$$\| \, \boldsymbol{u}_n \, \|_{\mathbf{X_T}} = \sup_{0 \leq t \leq T} \| \, \boldsymbol{u}_n(t) \, \|_{\mathbf{L}^2(\Omega)} \, .$$

In order to apply the Banach fixed point theorem, we establish some uniform estimates for Ψ. With Propositions 3.2, 3.3 and 3.4 in mind, we have the following :

Proposition 4.1 *There exists a constant $C > 0$ depending on n, λ, μ, κ, M, m, $\| \, \rho_0 \, \|_{H^1(\Omega)}$, $\| \, \boldsymbol{g} \, \|_{L^2(0,T;L^2(\Omega))}$, such that for all $\boldsymbol{u}_n \in \mathbf{X_T}$,*

$$\| \, \Psi\big(\boldsymbol{u}_n\big) \, \|_{\mathbf{X_T}} \leq \; \frac{M}{m} \, \| \, \boldsymbol{u}_0 \, \|_{L^2(\Omega)} + C \max(T, T^{\frac{1}{4}})\Big(1 + \| \, \boldsymbol{u}_n \, \|^2_{\mathbf{X_T}} \, \Big), \tag{32}$$

and for all $\boldsymbol{u}_n^1, \boldsymbol{u}_n^2 \in \mathbf{X_T}$,

$$\begin{aligned} \| \, \Psi\big(\boldsymbol{u}_n^1\big) - \Psi\big(\boldsymbol{u}_n^2\big) \, \|_{\mathbf{X_T}} \; \leq \; & C \max(T, T^{\frac{1}{4}})\Big(1 + \| \, \boldsymbol{u}_0 \, \|_{L^2(\Omega)} \\ & + \; \| \, \boldsymbol{u}_n^1 \, \|^2_{\mathbf{X_T}} + \| \, \boldsymbol{u}_n^2 \, \|^2_{\mathbf{X_T}} \, \Big) \, \| \, \boldsymbol{u}_n^1 - \boldsymbol{u}_n^2 \, \|_{\mathbf{X_T}} \, . \end{aligned} \tag{33}$$

At this stage, we set $R = 2\dfrac{M}{m} \, \| \, \boldsymbol{u}_0 \, \|_{L^2(\Omega)}$ and $\mathcal{B}_R^T = \Big\{ \boldsymbol{u} \in \mathbf{X_T}, \;\; \| \, \boldsymbol{u} \, \|_{\mathbf{X_T}} \leq R \Big\}$.

Proposition 4.2 *There exists $T_n \in]0,1[$ small enough and $\boldsymbol{u}_n \in \mathcal{B}_R^{T_n}$ such that*

$$\boldsymbol{u}_n = \Psi(\boldsymbol{u}_n).$$

Proof. Let $0 < T_n < 1$ such that

$$\max\Big(CT_n^{\frac{1}{4}}\big[R+\frac{1}{R}\big],\ CT_n^{\frac{1}{4}}\big[1+\parallel \boldsymbol{u}_0 \parallel_{L^2(\Omega)} +2R^2\big]\Big) \leq \frac{1}{2}.$$

Thanks to Proposition 4.1, we verify that Ψ is a contraction mapping on $\mathcal{B}_R^{T_n}$ and we conclude the existence of a unique fixed point of Ψ. ■

It is clear that $\boldsymbol{u}_n$ the fixed point of Ψ, obtained in Proposition 4.2, implies that $(\boldsymbol{u}_n, \rho_n = \mathcal{S}\boldsymbol{u}_n)$ is a local solution of the Galerkin approximate problem (28). Now, we will prove that this local solution is in fact a global one. For this, we establish some uniform estimates for $(\boldsymbol{u}_n, \rho_n)$ with respect to time.

Proposition 4.3 *If* $\dfrac{\lambda}{\mu}\max(1, \dfrac{\lambda^2}{\kappa})$ *small enough, there exists a constant* $C > 0$ *depending on* $\rho_0, \boldsymbol{u}_0, \boldsymbol{g}, M, \mu, \kappa$, *such that for all* $t \in [0, T_n)$

$$m \ \parallel \boldsymbol{u}_n(t) \parallel^2_{L^2(\Omega)} + \frac{\mu}{2}\int_0^t \parallel \nabla \boldsymbol{u}_n(s) \parallel^2_{L^2(\Omega)} ds \ \leq \ C, \tag{34}$$

$$\kappa \ \parallel \nabla \rho_n(t) \parallel^2_{L^2(\Omega)} + \kappa\lambda \int_0^t \parallel \Delta \rho_n(s) \parallel^2_{L^2(\Omega)} ds \ \leq \ C. \tag{35}$$

Evidently, thanks to the previous Proposition 4.3, we have the following :

Corollary 4.4 $(\boldsymbol{u}_n, \rho_n)$ *is a global solution of (28) and for all* $T > 0$,

$$(\boldsymbol{u}_n)_n \ \text{is bounded in} \ L^\infty\big(0,T;\boldsymbol{H}\big) \cap L^2\big(0,T;\boldsymbol{V}\big), \tag{36}$$

$$(\rho_n)_n \ \text{is bounded in} \ L^\infty\big(0,T;H^1(\Omega)\big) \cap L^2\big(0,T;H^2_N\big). \tag{37}$$

4.2. Uniform estimates for time derivatives

In this section, we establish uniform estimates for time derivatives $\partial_t \rho_n$ and $\partial_t \boldsymbol{u}_n$.

Proposition 4.5 *Let* $T > 0$. *The sequence* $(\partial_t \rho_n)_n$ *is bounded in* $L^{4/3}\big(0,T;L^2(\Omega)\big)$.

Proof. Taking the L^2-norm of $\partial_t \rho_n$. Applying the Hölder and Gagliardo-Nirenberg inequalities and the inequality : $\parallel \nabla\rho \parallel_{L^4(\Omega)} \leq \ C_0 \parallel \rho \parallel^{1/2}_{L^\infty(\Omega)} \parallel \Delta\rho \parallel^{1/2}_{L^2(\Omega)}$, we get

$$\parallel \partial_t \rho_n \parallel_{L^2(\Omega)} \leq \ \lambda \parallel \Delta\rho_n \parallel_{L^2(\Omega)} + C \parallel \boldsymbol{u}_n \parallel^{1/4}_{L^2(\Omega)} \parallel \nabla \boldsymbol{u}_n \parallel^{3/4}_{L^2(\Omega)} \parallel \rho_n \parallel^{1/2}_{L^\infty(\Omega)} \parallel \Delta\rho_n \parallel^{1/2}_{L^2(\Omega)} .$$

By the uniform estimate (34) and (17), we get

$$\parallel \partial_t \rho_n \parallel_{L^2(\Omega)} \leq \ \lambda \ \parallel \Delta\rho_n \parallel_{L^2(\Omega)} + C \ \parallel \nabla \boldsymbol{u}_n \parallel^{3/4}_{L^2(\Omega)} \parallel \Delta\rho_n \parallel^{1/2}_{L^2(\Omega)} . \tag{38}$$

Next, applying the Young inequality $ab \leq \dfrac{1}{2}(a^2 + b^2)$ in (38), we get

$$\parallel \partial_t \rho_n \parallel_{L^2(\Omega)} \leq \ \lambda \ \parallel \Delta\rho_n \parallel_{L^2(\Omega)} + C \ \parallel \nabla \boldsymbol{u}_n \parallel^{3/2}_{L^2(\Omega)} .$$

Thanks to the uniform time estimates (34) and (35), we deduce that $\parallel \partial_t \rho_n \parallel_{L^2(\Omega)}$ is bounded in $L^{4/3}\big(0,T\big)$. ■

Now, by following [1], we establish an estimation of the fractional time derivative of $\boldsymbol{u}_n$.

Proposition 4.6 *Let $0 < \delta < T$ such that*

$$\int_0^{T-\delta} \| \boldsymbol{u}_n(t+\delta) - \boldsymbol{u}_n(t) \|^2_{L^2(\Omega)} \, dt \leq C \, \delta^{\frac{1}{4}}, \tag{39}$$

where C a constant independent of n and δ.

Proof. For all functions $\boldsymbol{\phi} \in \mathbf{X_T}$, the approximate solution $(\boldsymbol{u}_n, \rho_n)$ verifies :

$$\begin{aligned} &\frac{d}{d\tau}\int_\Omega \rho_n \boldsymbol{u}_n \cdot \boldsymbol{\phi} \, d\boldsymbol{x} - \int_\Omega \rho_n \boldsymbol{u}_n \cdot \frac{\partial \boldsymbol{\phi}}{\partial \tau} \, d\boldsymbol{x} - \int_\Omega \rho_n (\boldsymbol{u}_n \cdot \nabla)\boldsymbol{\phi} \cdot \boldsymbol{u}_n \, d\boldsymbol{x} \\ &+ \mu \int_\Omega \nabla \boldsymbol{u}_n : \nabla \boldsymbol{\phi} \, d\boldsymbol{x} + \lambda \int_\Omega (\nabla \rho_n \cdot \nabla)\boldsymbol{\phi} \cdot \boldsymbol{u}_n \, d\boldsymbol{x} - \lambda \int_\Omega \rho_n \nabla \boldsymbol{u}_n^T : \nabla \boldsymbol{\phi} \, d\boldsymbol{x} \\ &- \lambda^2 \int_\Omega \frac{\nabla \rho_n \otimes \nabla \rho_n}{\rho_n} : \nabla \boldsymbol{\phi} \, d\boldsymbol{x} = \int_\Omega \rho_n \boldsymbol{g} \cdot \boldsymbol{\phi} \, d\boldsymbol{x} - \kappa \int_\Omega \Delta \rho_n \nabla \rho_n \cdot \boldsymbol{\phi} \, d\boldsymbol{x}. \end{aligned} \tag{40}$$

Integrating (40) with respect to τ between t and $t+\delta$, and taking $\boldsymbol{\phi} = \boldsymbol{u}_n(t+\delta) - \boldsymbol{u}_n(t)$

$$\begin{aligned} &\int_\Omega \big[\rho_n(t+\delta)\, \boldsymbol{u}_n(t+\delta) - \rho_n(t)\, \boldsymbol{u}_n(t)\big]\big[\boldsymbol{u}_n(t+\delta) - \boldsymbol{u}_n(t)\big] \, d\boldsymbol{x} \\ &= \int_t^{t+\delta} \int_\Omega \Big(\rho_n(\tau)\, \boldsymbol{g}(\tau) - \kappa\, \Delta \rho_n(\tau)\, \nabla \rho_n(\tau)\Big) \cdot \big(\boldsymbol{u}_n(t+\delta) - \boldsymbol{u}_n(t)\big) \, d\boldsymbol{x}\, d\tau \\ &+ \int_t^{t+\delta} \int_\Omega \Big(\big(\rho_n(\tau)\boldsymbol{u}_n(\tau) - \lambda \nabla \rho_n(\tau)\big) \cdot \nabla\Big)\big(\boldsymbol{u}_n(t+\delta) - \boldsymbol{u}_n(t)\big) \cdot \boldsymbol{u}_n(\tau) \, d\boldsymbol{x}\, d\tau \\ &- \int_t^{t+\delta} \int_\Omega \Big(\mu\, \nabla \boldsymbol{u}_n(\tau) - \lambda\, \rho_n(\tau) \nabla \boldsymbol{u}_n^T(\tau)\Big) : \nabla\big(\boldsymbol{u}_n(t+\delta) - \boldsymbol{u}_n(t)\big) \, d\boldsymbol{x}\, d\tau \\ &+ \lambda^2 \int_t^{t+\delta} \int_\Omega \frac{\nabla \rho_n(\tau) \otimes \nabla \rho_n(\tau)}{\rho_n(\tau)} : \nabla\big(\boldsymbol{u}_n(t+\delta) - \boldsymbol{u}_n(t)\big) \, d\boldsymbol{x}\, d\tau. \end{aligned} \tag{41}$$

Using the following identity

$$\rho_n(t+\delta)\, \boldsymbol{u}_n(t+\delta) - \rho_n(t)\, \boldsymbol{u}_n(t) = \rho_n(t+\delta)\big[\boldsymbol{u}_n(t+\delta) - \boldsymbol{u}_n(t)\big] + \big[\rho_n(t+\delta) - \rho_n(t)\big]\boldsymbol{u}_n(t),$$

then, (41) becomes

$$\begin{aligned} &\| \sqrt{\rho_n(t+\delta)}\big[\boldsymbol{u}_n(t+\delta) - \boldsymbol{u}_n(t)\big] \|^2_{L^2(\Omega)} \\ &= -\int_\Omega \big[\rho_n(t+\delta) - \rho_n(t)\big]\big[\boldsymbol{u}_n(t+\delta) - \boldsymbol{u}_n(t)\big] \cdot \boldsymbol{u}_n(t) \, d\boldsymbol{x} \\ &\quad + \int_t^{t+\delta} \int_\Omega \Big(\rho_n(\tau)\, \boldsymbol{g}(\tau) - \kappa\, \Delta \rho_n(\tau)\, \nabla \rho_n(\tau)\Big) \cdot \big(\boldsymbol{u}_n(t+\delta) - \boldsymbol{u}_n(t)\big) \, d\boldsymbol{x}\, d\tau \\ &\quad + \int_t^{t+\delta} \int_\Omega \Big(\big(\rho_n(\tau)\boldsymbol{u}_n(\tau) - \lambda \nabla \rho_n(\tau)\big) \cdot \nabla\Big)\big(\boldsymbol{u}_n(t+\delta) - \boldsymbol{u}_n(t)\big) \cdot \boldsymbol{u}_n(\tau) \, d\boldsymbol{x}\, d\tau \\ &\quad - \int_t^{t+\delta} \int_\Omega \Big(\mu\, \nabla \boldsymbol{u}_n(\tau) - \lambda\, \rho_n(\tau) \nabla \boldsymbol{u}_n^T(\tau)\Big) : \nabla\big(\boldsymbol{u}_n(t+\delta) - \boldsymbol{u}_n(t)\big) \, d\boldsymbol{x}\, d\tau \\ &\quad + \lambda^2 \int_t^{t+\delta} \int_\Omega \frac{\nabla \rho_n(\tau) \otimes \nabla \rho_n(\tau)}{\rho_n(\tau)} : \nabla\big(\boldsymbol{u}_n(t+\delta) - \boldsymbol{u}_n(t)\big) \, d\boldsymbol{x}\, d\tau \\ &= I_1(t) + I_2(t) + I_3(t) + I_4(t) + I_5(t) + I_6(t) + I_7(t) + I_8(t). \end{aligned} \tag{42}$$

Let us estimate $I_1(t)$. Applying the Hölder inequality, we get

$$|I_1(t)| \leq \| \rho_n(t+\delta) - \rho_n(t) \|_{L^2(\Omega)} \| \boldsymbol{u}_n(t+\delta) - \boldsymbol{u}_n(t) \|_{L^4(\Omega)} \| \boldsymbol{u}_n(t) \|_{L^4(\Omega)}.$$

In particular, we write

$$\rho_n(t+\delta)-\rho_n(t)=\int_t^{t+\delta}\frac{\partial\rho_n}{\partial\tau}\,d\tau.$$

Using the Hölder and Young inequalities and the embedding $H^1(\Omega)\subset L^4(\Omega)$, we obtain

$$|I_1(t)|\ \leq\ C\delta^{\frac{1}{4}}\Big(\int_t^{t+\delta}\ \|\ \frac{\partial\rho_n}{\partial\tau}\ \|^{\frac{4}{3}}_{L^2(\Omega)}\,d\tau\Big)^{\frac{3}{4}}\Big(\ \|\ \nabla\boldsymbol{u}_n(t+\delta)\ \|^2_{L^2(\Omega)}+\|\ \nabla\boldsymbol{u}_n(t)\ \|^2_{L^2(\Omega)}\ \Big).$$

In the same way, we verify the following estimations :

$$|I_2(t)|\ \leq\ C\delta^{\frac{1}{2}}\Big(\int_t^{t+\delta}\ \|\ \boldsymbol{g}(\tau)\ \|^2_{L^2(\Omega)}\,d\tau\Big)^{\frac{1}{2}}\Big(\ \|\ \nabla\boldsymbol{u}_n(t+\delta)\ \|^2_{L^2(\Omega)}+\|\ \nabla\boldsymbol{u}_n(t)\ \|^2_{L^2(\Omega)}\ \Big),$$

$$|I_3(t)|\ \leq\ C\delta^{\frac{1}{4}}\Big(\int_t^{t+\delta}\ \|\ \Delta\rho_n(\tau)\ \|^2_{L^2(\Omega)}\,d\tau\Big)^{\frac{3}{4}}\Big(\ \|\ \nabla\boldsymbol{u}_n(t+\delta)\ \|^2_{L^2(\Omega)}+\|\ \nabla\boldsymbol{u}_n(t)\ \|^2_{L^2(\Omega)}\ \Big).$$

Similarly, one can obtain the desired estimates of $I_j(t)$ terms, for $j=4,\ldots,8$.
At last, if we choose $0<\delta<1$ and taking into account Propositions 4.3 and 4.5, then by gathering together all the above estimates, we rewrite (42) as follows :

$$\|\ \sqrt{\rho_n(t+\delta)}\big[\boldsymbol{u}_n(t+\delta)-\boldsymbol{u}_n(t)\big]\ \|^2_{L^2(\Omega)}\leq C\delta^{\frac{1}{4}}\Big(\ \|\ \nabla\boldsymbol{u}_n(t+\delta)\ \|^2_{L^2(\Omega)}\ +\|\ \nabla\boldsymbol{u}_n(t)\ \|^2_{L^2(\Omega)}\ \Big).$$

Thanks to the lower bound of ρ_n and Proposition 4.3, we finish the proof. ■

4.3. The existence of solution $(\boldsymbol{u},\rho)$

The final step to complete this study is to employ the previous uniform estimates in order to pass to the limit in the approximate problem (28). When $n\to+\infty$, we have

$$\boldsymbol{u}_{0n}\ \longrightarrow\ \boldsymbol{u}_0\ \text{ in }\mathbf{H}\ \text{ strongly.}$$

Thanks to (36) and (37), choosing the subsequences $(\boldsymbol{u}_n)_n$ and $(\rho_n)_n$ such that

$$\begin{array}{rcll}
\boldsymbol{u}_n & \longrightarrow & \boldsymbol{u} \quad \text{in } L^2\big(0,T;\mathbf{V}\big) & \text{weakly,}\\
\boldsymbol{u}_n & \longrightarrow & \boldsymbol{u} \quad \text{in } L^\infty\big(0,T;\mathbf{H}\big) & \text{weakly-star,}
\end{array}$$

and

$$\begin{array}{rcll}
\rho_n & \longrightarrow & \rho \quad \text{in } L^2\big(0,T;H^2_N\big) & \text{weakly,}\\
\rho_n & \longrightarrow & \rho \quad \text{in } L^\infty\big(0,T;H^1(\Omega)\big) & \text{weakly-star,}\\
\partial_t\,\rho_n & \longrightarrow & \partial_t\,\rho \quad \text{in } L^{4/3}\big(0,T;L^2(\Omega)\big) & \text{weakly.}
\end{array}$$

We are able to pass to the limit in the linear terms of (28), thanks to these above convergence results. Now, to ensure the passage to the limit in the nonlinear terms of (28), it is necessary to use the following strong convergence :

Proposition 4.7 *There exists a subsequence $(\boldsymbol{u}_n,\rho_n)_n$ which converges strongly to $(\boldsymbol{u},\rho)$ in $L^2\big(0,T;\boldsymbol{L}^2(\Omega)\big)\times L^2\big(0,T;H^1(\Omega)\big)$. Moreover, $(\boldsymbol{u},\rho)$ is a weak solution of (10).*

Proof. Applying some compactness theorems [13, Chap.3, Theorem 2.1] for ρ_n and [12, Theorem 5] for $\boldsymbol{u}_n$ and using Propositions 4.5 and 4.6, we get to the desired result. ■

5. Conclusions

In this paper, we study the system of PDEs derived from the compressible Navier-Stokes equations with presence of a specific Korteweg stress tensor, called the Kazhikhov-Smagulov-Korteweg (KSK) model. We arrive at verify the existence of a weak solution $(\boldsymbol{u}, \rho)$ of the KSK model (10) global in time with finite and uniformly bounded energy. Then, we conclude the proof of Theorem 2.2, the main result of this paper.

6. References

[1] D. BRESCH, E.H. ESSOUFI, M. SY, "Effects of density dependent viscosities on multiphasic incompressible fluid models", *J. Math. Fluid Mech.*, vol. 9, num. 3, p. 377-397, 2007.

[2] D. BRESCH, B. DESJARDINS, C.K. LIN, "On some compressible fluid models: Korteweg, lubrication and shallow water systems", *Comm. Partial Diff. Eqs.*, vol. 28, num. 3-4, p. 843-868, 2003.

[3] C. CALGARO, E. CREUSÉ, T. GOUDON, "Modeling and simulation of mixture flows: Application to powder-snow avalanches", *Computers and Fluids*, vol. 107, p. 100-122, 2015.

[4] J.E. DUNN, J. SERRIN, "On the thermomechanics of interstitial working", *Arch. Rational Mech. Anal.*, vol. 88, num. 2, p. 95-133, 1985.

[5] E. FEIREISL, A. NOVOTNÝ, H. PETZELTOVÁ, "On the existence of globally defined weak solutions to the Navier-Stokes equations", *J. Math. Fluid Mech.*, vol. 3, p. 358-392, 2001.

[6] F. FRANCHI, B. STRAUGHAN, "A comparison of Graffi and Kazhikov-Smagulov models for top heavy pollution instability", *Adv. in Water Resources*, vol. 24, p. 585-594, 2001.

[7] P. GALDI, D.D. JOSEPH, L. PREZIOSI, S. RIONERO, "Mathematical problems for miscible, incompressible fluids with Korteweg stresses", *European J. of Mech. B-Fluids*, vol. 10, num. 3, p. 253-267, 1991.

[8] D.D. JOSEPH, "Fluid dynamics of two miscible liquids with diffusion and gradient stresses", *European J. of Mech. B-Fluids*, vol. 6, p. 565-596, 1990.

[9] D.J. KORTEWEG, "Sur la forme que prennent les équations du mouvement des fluides si l'on tient compte des forces capillaires causées par des variations de densité considérables mais continues et sur la théorie de la capillarité dans l'hypothèse d'une variation continue de la densité", *Archives Néerlandaises des Sciences Exactes et Naturelles, Séries II*, vol. 6, p. 1-24, 1901.

[10] A. KAZHIKHOV, SH. SMAGULOV, "The correctness of boundary value problems in a diffusion model of an inhomogeneous fluid", *Sov. Phys. Dokl.*, vol. 22, num. 1, p. 249-252, 1977.

[11] J.L. LIONS, "Quelques méthodes de résolution des problèmes aux limites non linéaires", *Dunod, Gauthier-Villars, Paris*, 1969.

[12] J. SIMON, "Compact sets in the space $L^p\big(0,T;B\big)$", *Ann. Mat. Pura Appl.*, vol. 146, p. 65-96, 1987.

[13] R. TEMAM, "Navier-Stokes equations, theory and numerical analysis", *Revised Edition, Studies in mathematics and its applications vol. 2, North Holland Publishing Company-Amsterdam, New York*, 1984.

Theoretical Analysis of a Water Wave Model using the Diffusive Approach

Olivier Goubet[a] — Imen Manoubi [b,*]

[a] Université de Picardie Jules Verne
33 rue Saint-Leu 80039
Amiens, France.
olivier.goubet@u-picardie.fr

[b] UR Multifractales et Ondelettes
Faculté des Sciences de Monastir
Avenue de l'environnement 5019 Monastir, Tunisie
imen.manoubi@yahoo.fr

[*] Corresponding author

ABSTRACT. In this paper, we theoretically study the water wave model with a nonlocal viscous term

$$u_t + u_x + \beta u_{xxx} + \frac{\sqrt{\nu}}{\sqrt{\pi}} \frac{\partial}{\partial t} \int_0^t \frac{u(s)}{\sqrt{t-s}} ds + u u_x = \nu u_{xx},$$

where the Riemann-Liouville half-order derivative $\frac{1}{\sqrt{\pi}} \frac{\partial}{\partial t} \int_0^t \frac{u(s)}{\sqrt{t-s}} ds$ is represented with a diffusive realization.

RÉSUMÉ. Dans cet article, nous étudions théoriquement le modèle visqueux asymptotique

$$u_t + u_x + \beta u_{xxx} + \frac{\sqrt{\nu}}{\sqrt{\pi}} \frac{\partial}{\partial t} \int_0^t \frac{u(s)}{\sqrt{t-s}} ds + u u_x = \nu u_{xx},$$

où la demi-dérivée de Riemann-Liouville $\frac{1}{\sqrt{\pi}} \frac{\partial}{\partial t} \int_0^t \frac{u(s)}{\sqrt{t-s}} ds$ est représentée à l'aide d'une réalisation diffusive.

KEYWORDS : nonlocal viscous model, Riemann-Liouville half derivative, diffusive realization

MOTS-CLÉS : modèle visqueux non local, demi-dérivée de Riemann-Liouville, réalisation diffusive

1. Introduction

1.1. State of the art.

The modeling and the mathematical analysis of viscosity in water wave propagation are challenging issues. In the last decade, P. Liu and T. Orfila [8], and D. Dutykh and F. Dias [6] have independently derived viscous asymptotic models for transient long-wave propagation on viscous shallow water. These effects appear as nonlocal terms in the form of convolution integrals. A one-dimensional nonlinear system is presented in [5].
In their recent work [4], M. Chen et al. investigated theoretically and numerically the decay rate for solutions to the following water wave model with a nonlocal viscous dispersive term as follows

$$u_t + u_x + \beta u_{xxx} + \frac{\sqrt{\nu}}{\sqrt{\pi}} \int_0^t \frac{u_t(s)}{\sqrt{t-s}} ds + u u_x = \alpha u_{xx}, \tag{1}$$

where $\frac{1}{\sqrt{\pi}} \int_0^t \frac{u_t(s)}{\sqrt{t-s}} ds$ represents the Caputo half-derivative in time. Here u is the horizontal velocity of the fluid, $-\alpha u_{xx}$ is the usual diffusion, βu_{xxx} is the geometric dispersion and $\frac{1}{\sqrt{\pi}} \int_0^t \frac{u_t(s)}{\sqrt{t-s}} ds$ stands for the nonlocal diffusive-dispersive term. We denote as β, ν and α the parameters dedicated to balance or unbalance the effects of viscosity and dispersion against nonlinear effects. Particularly, the authors in [4] consider (1) with $\beta = 0$ supplemented with the initial condition $u_0 \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. They proved that if $\|u_0\|_{L^1(\mathbb{R})}$ is small enough, then there exists a unique global solution $u \in C(\mathbb{R}_+; L^2_x(\mathbb{R})) \cap C^1(\mathbb{R}_+; H^{-2}_x(\mathbb{R}))$. In addition, u satisfies

$$t^{1/4} \|u(t,\cdot)\|_{L^2_x(\mathbb{R})} + t^{1/2} \|u(t,\cdot)\|_{L^\infty_x(\mathbb{R})} < C(u_0). \tag{2}$$

In order to study the effects of the nonlocal term on the existence and on the decay rate of the solutions, the second author considered in her recent work [10] a derived model from (1) where the fractional term is described by the Riemann-Liouville half derivative instead of that of Caputo, namely

$$u_t + u_x + \beta u_{xxx} + \frac{\sqrt{\nu}}{\sqrt{\pi}} \frac{\partial}{\partial t} \int_0^t \frac{u(s)}{\sqrt{t-s}} ds + u u_x = \alpha u_{xx}. \tag{3}$$

She proved the local and global existence of solutions to problem (3) when $\beta = 0$ using a fixed point theorem. Then she studied theoretically the decay rate of the solutions in this case. Precisely, she stated the following theorem

Theorem 1.1 (I. Manoubi, 2014) *Let $u_0 \in L^2(\mathbb{R})$, then there exists a unique local solution $u \in C([0,T); L^2_x(\mathbb{R}))$ of* (3).
Moreover for $u_0 \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, there exists a positive constant $C_0 > 0$ that depends on u_0 such that if $\|u_0\|_{L^1(\mathbb{R})}$ is small enough, there exists a unique global solution $u \in C(\mathbb{R}_+; L^2_x(\mathbb{R})) \cap C^{1/2}(\mathbb{R}_+; H^{-2}_x(\mathbb{R}))$ of (3) *given by*

$$u(t,x) = [K_{RL}(t,\cdot) \star u_0](x) - N \circledast u^2(t,x), \tag{4}$$

where K_{RL} and N are given by

$$K_{RL}(t,x) = \frac{1}{2\sqrt{\pi t}} e^{-\frac{x^2}{4t}} e^{-x^-} (1 - \frac{1}{2} \int_0^{+\infty} e^{-\frac{\mu^2}{4t} - \frac{\mu|x|}{2t} - \frac{\mu}{2}} d\mu),$$

and

$$N(t,x)=\frac{1}{4\sqrt{\pi t}}\partial_x\left(e^{-\frac{x^2}{4t}}e^{-x^-}\left(1-\frac{1}{2}\int_0^{+\infty}e^{-\frac{\mu^2}{4t}-\frac{\mu|x|}{2}-\frac{\mu}{2}}d\mu\right)\right).$$

with $x^-=\frac{|x|-x}{2}=max(-x,0)$, $\star$ *represents the usual convolution product and* $\circledast$ *is the time-space convolution product defined by*

$$v\circledast w(t,x)=\int_0^t\int_{\mathbb{R}}v(t-s,x-y)w(s,y)dsdy.$$

whenever the integrals make sense. In addition, we have the following estimate

$$\max(t^{1/4},t^{3/4})\|u(t,\cdot)\|_{L^2_x(\mathbb{R})}+\max(t^{1/2},t)\|u(t,\cdot)\|_{L^\infty_x(\mathbb{R})}\leq C_0. \tag{5}$$

The proof of this theorem is presented in [10].
However, all these results are performed assuming a smallness condition on the initial data. In order to remove this smallness condition and to investigate the model (3) for a large class of initial data, we introduce here the concept of diffusive realizations for the half-order derivative. This approach was initially developed by Montseny [12], Montseny et al. [14, 15] and Staffans [16]. Diffusive realization make possible to represent nonlocal in time operators, and more generally *causal* pseudo-differential operators, in a state space model formulation where the state belongs to an appropriate Hilbert space. Different applications of this approach can be found in [1, 7, 11, 13].
In this article, we assume that the effects of the geometric dispersion in (3) is less important than the viscosity effects (i.e we take $\beta=0$ in (3)) and we assume that the other constants are normalized. Thus, our model is reduced as follows

$$u_t+u_x+\frac{1}{\sqrt{\pi}}\frac{\partial}{\partial t}\int_0^t\frac{u(s)}{\sqrt{t-s}}ds+uu_x=u_{xx}. \tag{6}$$

We prove the well posedness of the model (6) for all initial data $u_0\in H^1(\mathbb{R})$ using the diffusive realization. To this end, we complete the introduction as follows. We first introduce the diffusive formulation of the half-order Riemann-Liouville derivative. Then, we deduce the mathematical model that derives from (6) using the diffusive approach. Finally, we present the main results of this article.
We note that one can consider the general case of the Caputo or Riemann-Liouville fractional derivative of order α where $0<\alpha<1$. Comparing the effects of these nonlocal terms with our results is a challenging issue and it may be the subject of a future work. However, choosing another definition of fractional derivative like Atangana-Baleanu derivative or Caputo-Fabrizio derivative in equation (1) must be justified.

1.2. Diffusive formulation of the model

In the literature, there are several diffusive realizations of the Riemann-Liouville half-order derivative. We recall in the following some of these formulations.
First, the diagonal form of the diffusive realization of $D^{1/2}u(t)$ which will be used in the remaining of this article is given for all $t>0$ by

$$\begin{cases}\partial_t\psi(t,\sigma)=-\sigma\psi(t,\sigma)+u(t),\ \ \psi(0,\sigma)=0,\ \ \sigma\in\mathbb{R}^+\\ D^{1/2}u(t)=\displaystyle\int_0^{+\infty}\frac{1}{\pi\sqrt{\sigma}}\partial_t\psi(t,\sigma)d\sigma.\end{cases} \tag{7}$$

Second, the PDE-form of the diffusive realization of $D^{1/2}u(t)$ is given for all $t > 0$ by

$$\begin{cases} \partial_t \Phi(t,y) = \Phi_{yy}(t,y) + u(t) \otimes \delta_{y=0}, \;\; \Phi(0,y) = 0, \;\; y \in \mathbb{R}, \\ D^{1/2}u(t) = 2 < \delta_{y=0}, \partial_t \Phi(t,y) >_{\mathcal{D}',\mathcal{D}} = 2\dfrac{d}{dt}\Phi(t,0). \end{cases} \tag{8}$$

where $\delta_{y=0}$ is the Dirac delta function at $y = 0$ and $u(t) \otimes \delta_{y=0}$ is the tensorial product in the distributions sense of the applications $t \mapsto u(t)$ and $y \mapsto \delta_{y=0}$.
Finally, another form of the diffusive realization of $D^{1/2}u(t)$ is given for all $t > 0$ by

$$\begin{cases} \partial_t \phi(t,\sigma) = -\sigma^2 \phi(t,\sigma) + \dfrac{2}{\pi}u(t), \;\; \phi(0,\sigma) = 0, \;\; \sigma \geq 0, \\ D^{1/2}u(t) = \displaystyle\int_0^{+\infty} (\frac{2}{\pi}u(t) - \sigma^2 \phi(t,\sigma)) d\sigma. \end{cases} \tag{9}$$

We note that the author has used the diffusive realizations (8) and (9) in her PhD Thesis to study mathematically and numerically the integro-differential equation (3) when $\beta = 0$, $\nu = \alpha = 1$. For more details, we refer the readers to [9]. In the following, we describe the mathematical framework. Thanks to the diffusive realization (7), the problem (6) is written as follows

$$\begin{cases} u_t(t,x) + u_x(t,x) + \displaystyle\int_0^{+\infty} (u(t,x) - \sigma\psi(t,x,\sigma)) \frac{d\sigma}{\pi\sqrt{\sigma}} \\ \qquad + u(t,x)u_x(t,x) = u_{xx}(t,x), \;\; t > 0, \; x \in \mathbb{R}, \\ \psi_t(t,x,\sigma) = -\sigma\psi(t,x,\sigma) + u(t,x), \;\; t > 0, \; x \in \mathbb{R}, \; \sigma \geq 0, \\ u(0,x) = u_0(x), \;\; x \in \mathbb{R}, \\ \psi(0,x,\sigma) = 0, \;\; x \in \mathbb{R}, \sigma \geq 0. \end{cases} \tag{10}$$

Then, We rewrite the system (10) as a first-order semi-linear differential equation as follows

$$\begin{cases} X_t + \mathcal{A}X = F(X), \\ X(0) = X_0, \end{cases} \tag{11}$$

where $X = (u, \psi)^T$, $X_0 = (u_0, 0)^T$ and

$$\mathcal{A}X = \begin{pmatrix} \displaystyle\int_0^{+\infty} (u - \sigma\psi) \frac{d\sigma}{\pi\sqrt{\sigma}} - u_{xx} \\ -u + \sigma\psi \end{pmatrix},$$

$$F(X) = \begin{pmatrix} -u_x - uu_x \\ 0 \end{pmatrix}. \tag{12}$$

1.3. Main results.

We introduce our functional space . First, we define the positive measure dN on $\mathbb{R}_+$ by

$$dN(\sigma) = \frac{d\sigma}{\pi\sqrt{\sigma}}.$$

Hence, dN satisfies

$$C_N = \int_0^{+\infty} \frac{dN(\sigma)}{1+\sigma} = 1. \tag{13}$$

Then, we define the spaces

$$\begin{aligned} H_N &= L^2(\mathbb{R}_+, dN), \\ \widetilde{H}_N &= L^2(\mathbb{R}_+, \sigma dN), \\ V &= L^2(\mathbb{R}_+, (1+\sigma)dN). \end{aligned}$$

We suppose that (11) has a regular solution. The following result holds.

Proposition 1.2 *The energy function associated to* (11)

$$\mathcal{E}(t) = \frac{1}{2}\|u(t)\|^2_{L^2_x} + \frac{1}{2}\|\psi(t)\|^2_{L^2(\mathbb{R},\widetilde{H}_N)}, \tag{14}$$

satisfies the following energetic equilibrium

$$\frac{1}{2}\frac{d}{dt}\mathcal{E}(t) = -\int_{\mathbb{R}} \|u(t,x) - \sigma\psi(t,x,\sigma)\|^2_{H_N} dx - \int_{\mathbb{R}} |u_x(t,x)|^2 dx. \tag{15}$$

The natural energy space of the solution X is

$$\mathcal{H} = L^2(\mathbb{R}) \times L^2(\mathbb{R}, \widetilde{H}_N),$$

endowed with the scalar product $(\cdot,\cdot)_{\mathcal{H}}$ defined for all $X = (u,\psi)^T$ and $Y = (v,\chi)^T$ in $\mathcal{H}$ by

$$(X,Y)_{\mathcal{H}} = (u,v)_{L^2(\mathbb{R})} + \int_{\mathbb{R}} (\psi,\chi)_{\widetilde{H}_N} dx.$$

Moreover, we define the following Hilbert space

$$\mathcal{V} = H^2(\mathbb{R}) \times L^2(\mathbb{R}, \widetilde{H}_N).$$

We state the main result of this article.

Theorem 1.3 *For all $u_0 \in H^1(\mathbb{R})$, there exists a unique global solution $X \in C([0,+\infty[, D(\mathcal{A}^{1/2}))$ of* (11) *such that $X_0 = \begin{pmatrix} u_0 \\ 0 \end{pmatrix}$ and*

$$X(t) = \Phi(X)(t). \tag{16}$$

2. Proof of Theorem 1.3.

2.1. The linear problem

We first consider the following linear problem associated to (11)

$$\begin{cases} X_t + \mathcal{A}X = 0 \ \ \forall t > 0, \\ X(0) = X_0, \end{cases} \tag{17}$$

where $X = \begin{pmatrix} u \\ \psi \end{pmatrix}$, $X_0 = \begin{pmatrix} u_0 \\ 0 \end{pmatrix}$ and

$$\mathcal{A}X = \begin{pmatrix} \displaystyle\int_0^{+\infty} (u - \sigma\psi)\frac{d\sigma}{\pi\sqrt{\sigma}} - u_{xx} \\ -u + \sigma\psi \end{pmatrix}, \tag{18}$$

We can establish the following properties of the operator $\mathcal{A}$.

Proposition 2.1 *The domain $D(\mathcal{A})$ of the operator $\mathcal{A}$ in $\mathcal{H}$ is given by*

$$D(\mathcal{A}) = \{(u, \psi) \in \mathcal{V}; u - \sigma\psi \in L^2(\mathbb{R}, V)\}.$$

We define the norm of $X \in D(\mathcal{A})$ by

$$\|X\|_{D(\mathcal{A})} = \left(\|X\|_{\mathcal{H}}^2 + \|\mathcal{A}X\|_{\mathcal{H}}^2\right)^{1/2}.$$

Moreover $\mathcal{A} : D(\mathcal{A}) \subset \mathcal{H} \longrightarrow \mathcal{H}$ is well-defined and bounded on $D(\mathcal{A})$.

Lemma 2.2 *The domain $D(\mathcal{A}^{1/2})$ of the operator $\mathcal{A}^{1/2}$ in $\mathcal{H}$ is given by*

$$D(\mathcal{A}^{1/2}) = \{(u, \psi) \in H^1(\mathbb{R}) \times L^2(\mathbb{R}, \widetilde{H}_N) \text{ and } u - \sigma\psi \in L^2(\mathbb{R}, H_N)\}.$$

equipped by the norm

$$\|X\|_{D(\mathcal{A}^{1/2})} = \left(\int_{\mathbb{R}} \|u - \sigma\psi\|_{H_N}^2 dx + \|u_x\|_{L^2(\mathbb{R})}^2\right)^{1/2}.$$

Proposition 2.3 *The operator $\mathcal{A}$ is maximal monotone and symmetric. Thus $\mathcal{A}$ is auto-adjoint.*

In the following, we state results based on the Hille-Yosida Theorem [2, 3].

Proposition 2.4 (Hille-Yosida) *For all $X_0 \in D(\mathcal{A})$, there exists a unique solution*

$$X \in C^1(]0, +\infty[, \mathcal{H}) \bigcap C([0, +\infty[, D(\mathcal{A}))$$

of (17)*. Moreover, formally we have*

$$X(t) = e^{-t\mathcal{A}} X_0.$$

Proposition 2.5 *For all $X_0 \in \mathcal{H}$, there exists a unique solution*

$$X \in C([0, +\infty[, \mathcal{H}) \bigcap C^1(]0, +\infty[, \mathcal{H}) \bigcap C(]0, +\infty[, D(\mathcal{A}))$$

of (17)*.*

Proposition 2.6 *For all $X_0 \in D(\mathcal{A}^{1/2})$, equation* (17) *has a unique solution*

$$X \in C([0, +\infty[, D(\mathcal{A}^{1/2})).$$

Proof. Let $X_0 \in D(\mathcal{A}^{1/2})$. We consider the following problem

$$\begin{cases} \mathcal{H}_1 = D(\mathcal{A}^{1/2}), \\ \mathcal{A}_1 X = \mathcal{A}X \quad \text{pour} \quad X \in \mathcal{H}_1, \\ X_t + \mathcal{A}_1 X = 0 \ \ \forall t > 0, \\ X(0) = X_0 \in \mathcal{H}_1. \end{cases} \tag{19}$$

Then $\mathcal{A}_1$ is unbounded operator and $D(\mathcal{A}_1)$ is its domain in $\mathcal{H}_1$. By construction, $\mathcal{A}_1$ is a self-adjoint operator. Moreover,

$$\begin{aligned} D(\mathcal{A}_1) =& \{X \in \mathcal{H}_1; \mathcal{A}_1 X \in \mathcal{H}_1\} \\ =& \{X \in D(\mathcal{A}^{1/2}); \mathcal{A}X \in D(\mathcal{A}^{1/2})\} \\ =& \{X \in D(\mathcal{A}); (\mathcal{A}^2 X, \mathcal{A}X) < \infty\} = D(\mathcal{A}^{3/2}). \end{aligned}$$

In addition, $\mathcal{A}_1$ is a maximal and monotone operator. In fact,

$$(\mathcal{A}_1 X, X)_{\mathcal{H}_1} = (\mathcal{A}X, X)_{D(\mathcal{A}^{1/2})} = (\mathcal{A}\mathcal{A}X, X)_{\mathcal{H}}.$$

Since $\mathcal{A}$ is self-adjoint then

$$(\mathcal{A}_1 X, X)_{\mathcal{H}_1} = (\mathcal{A}X, \mathcal{A}X)_{\mathcal{H}} = \|\mathcal{A}X\|^2_{\mathcal{H}} \geq 0.$$

We deduce that $\mathcal{A}_1$ is monotone. Now, we establish that $\mathcal{A}_1$ is maximal. Let $Y \in \mathcal{H}_1 = D(\mathcal{A}^{1/2})$ and we establish that there exists $X \in D(\mathcal{A}_1)$ such that $(I + \mathcal{A}_1)X = Y$. Since $\mathcal{H}_1 \subset \mathcal{H}$ then there exists $X \in D(\mathcal{A})$ such that

$$(I + \mathcal{A})X = X + \mathcal{A}X = Y.$$

In particular, since $D(\mathcal{A})) \subset D(\mathcal{A}^{1/2})$, then

$$X \in D(\mathcal{A}^{1/2}) \text{ et } Y \in D(\mathcal{A}^{1/2}),$$

This implies that

$$X \in D(\mathcal{A}^{1/2}) \text{ et } \mathcal{A}X \in D(\mathcal{A}^{1/2}).$$

We conclude that $X \in D(\mathcal{A}^{3/2}) = D(\mathcal{A}_1)$ and verifies $(I + \mathcal{A})X = (I + \mathcal{A}_1)X = Y$. Hence, using Hille-Yosida Theorem, we conclude that there exists a unique solution of (19)

$$X(t) = e^{-t\mathcal{A}_1} X_0 \in C^0([0, +\infty[, \mathcal{H}_1).$$

Moreover, $D(\mathcal{A}^{1/2}) = \mathcal{H}_1 \subset \mathcal{H}$ and using the uniqueness of the solution of (17), we deduce that

$$X(t) = e^{-t\mathcal{A}} X_0 \in C^0([0, +\infty[, D(\mathcal{A}^{1/2})).$$

■

We have the following uniform estimates.

Proposition 2.7 *First,*

$$\forall X_0 \in \mathcal{H}, \forall t > 0, \ \ \|e^{-t\mathcal{A}} X_0\|_{\mathcal{H}} \leq \|X_0\|_{\mathcal{H}}. \tag{20}$$

Second,

$$\forall X_0 \in D(\mathcal{A}^{1/2}), \forall t > 0, \ \ \|e^{-t\mathcal{A}} X_0\|_{D(\mathcal{A}^{1/2})} \leq \|X_0\|_{D(\mathcal{A}^{1/2})}. \tag{21}$$

Finally,

$$\exists C > 0, \forall X_0 \in \mathcal{H}, \forall t > 0, \ \ \|e^{-t\mathcal{A}} X_0\|_{D(\mathcal{A}^{1/2})} \leq \frac{C}{\sqrt{t}} \|X_0\|_{\mathcal{H}}. \tag{22}$$

2.2. Resolution in $H^1(\mathbb{R})$.

In this subsection, we focus on the problem (11). Formally, if X is a solution of (11) then X satisfies the Duhamel form as follows

$$X(t) = e^{-t\mathcal{A}} X_0 + \int_0^t e^{-(t-s)\mathcal{A}} F(X(s)) ds, \tag{23}$$

Hence X is considered as a fixed point of the functional Φ defined by (23) as

$$\Phi(X)(t,x) = e^{-t\mathcal{A}} X_0 + \int_0^t e^{-(t-s)\mathcal{A}} F(X(s)) ds. \tag{24}$$

In the sequel, we establish Theorem 1.3. To this end, we start by proving the local existence of the solution of (16) using the fixed point Theorem.
Local existence. First we have the following result.

Proposition 2.8 *The function $F : D(\mathcal{A}^{1/2}) \subset \mathcal{H} \to \mathcal{H}$, given by* (12)*, is locally lipschitz continuous on $D(\mathcal{A}^{1/2})$. Moreover, for all $X, Y \in D(\mathcal{A}^{1/2})$ we have*

$$\|F(X) - F(Y)\|_{\mathcal{H}} \leq (\|X\|_{D(\mathcal{A}^{1/2})} + \|Y\|_{D(\mathcal{A}^{1/2})} + 1) \|X - Y\|_{D(\mathcal{A}^{1/2})}.$$

Corollory 2.9 *Since $F(0) = 0$, we deduce that*

$$\forall X \in D(\mathcal{A}^{1/2}), \ \|F(X)\|_{\mathcal{H}} \leq \|X\|^2_{D(\mathcal{A}^{1/2})} + \|X\|_{D(\mathcal{A}^{1/2})}. \tag{25}$$

Let $T > 0$ and set

$$E_T = C([0,T], D(\mathcal{A}^{1/2})),$$

E_T is a Banach space when endowed with the norm

$$\|X\|_{E_T} := \sup_{t \in [0,T]} \|X(t)\|_{D(\mathcal{A}^{1/2})}.$$

In the following, we state some properties satisfied by the functional Φ on E_T with $X_0 \in D(\mathcal{A}^{1/2})$). Let $X \in E_T$, we have

$$\begin{aligned} \|\Phi(X)(t)\|_{D(\mathcal{A}^{1/2})} &\leq \|e^{-t\mathcal{A}} X_0\|_{D(\mathcal{A}^{1/2})} + \int_0^t \|e^{-(t-s)\mathcal{A}} F(X(s))\|_{D(\mathcal{A}^{1/2})} ds \\ &\leq \|X_0\|_{D(\mathcal{A}^{1/2})} + \int_0^t \frac{c}{\sqrt{t-s}} \|F(X(s))\|_{\mathcal{H}} ds. \end{aligned}$$

Thanks to (25), we obtain

$$\|\Phi(X)(t)\|_{D(\mathcal{A}^{1/2})} \leq \|X_0\|_{D(\mathcal{A}^{1/2})} + \int_0^t \frac{c}{\sqrt{t-s}}(\|X(t)\|^2_{D(\mathcal{A}^{1/2})} + \|X(t)\|_{D(\mathcal{A}^{1/2})})ds.$$

Hence, for all $t \in [0,T]$ we have

$$\|\Phi(X)(t)\|_{D(\mathcal{A}^{1/2})} \leq \|X_0\|_{D(\mathcal{A}^{1/2})} + C_1\sqrt{T}(\|X\|^2_{E_T} + \|X\|_{E_T}). \tag{26}$$

Moreover, let X and $Y \in E_T$ then

$$\begin{aligned}
\|\Phi(X)(t) - \Phi(Y)(t)\|_{D(\mathcal{A}^{1/2})} &= \| \int_0^t e^{-(t-s)\mathcal{A}}F(X(s))ds - \int_0^t e^{-(t-s)\mathcal{A}}F(Y(s))ds\|_{D(\mathcal{A}^{1/2})} \\
&\leq \int_0^t \|e^{-(t-s)\mathcal{A}}(F(X(s)) - F(Y(s)))\|_{D(\mathcal{A}^{1/2})}ds \\
&\leq \int_0^t \frac{c}{\sqrt{t-s}}\|F(X(s)) - F(Y(s))\|_{\mathcal{H}}ds \\
&\leq \int_0^t \frac{c(K)}{\sqrt{t-s}}\|X(s) - Y(s)\|_{D(\mathcal{A}^{1/2})}ds \\
&\leq c(K)\sqrt{t}\|X - Y\|_{E_T}.
\end{aligned}$$

Here K is the constant of Lipschitz of F on the ball B. Hence, for all $t \in [0,T]$

$$\|\Phi(X)(t) - \Phi(Y)(t)\|_{D(\mathcal{A}^{1/2})} \leq C_2\sqrt{T}\|X - Y\|_{E_T}. \tag{27}$$

Also, we show that if $X_0 \in D(\mathcal{A}^{1/2}))$ then Φ is well defined. Next we define a set B invariant under the action of Φ. Therefore we take $R = 2\|X_0\|_{D(\mathcal{A}^{1/2})}$. Let $B(0,R)$ the closed ball in E_T of radius R centered at the origin. Thanks to (26) and (27), we get

$$\forall X \in B,\ \ \|\Phi(X)(t)\|_{D(\mathcal{A}^{1/2})} \leq \frac{R}{2} + C_1\sqrt{T}(R^2 + R),$$

$$\forall X, Y \in B,\ \ \|\Phi(X)(t) - \Phi(Y)(t)\|_{D(\mathcal{A}^{1/2})} \leq C_2\sqrt{T}\|X - Y\|_{E_T}.$$

Finally, we choose T small such that $\max(C_1R, C_2)\sqrt{T} \leq \frac{1}{2}$. Hence, with this choice, we get $\Phi(B) \subset B$ and thus the map Φ is a contraction on B. Using the fixed point Theorem, we deduce that there exists a unique fixed point X of the functional Φ on B. Moreover, $X \in C([0,T], D(\mathcal{A}^{1/2}))$.
In the following, we establish the global existence of the solution of (16).
Global existence. We take the scalar product in $\mathcal{H}$ of (11) with X, we get

$$(X_t, X)_{\mathcal{H}} + (\mathcal{A}X, X)_{\mathcal{H}} = (F(X), X)_{\mathcal{H}}. \tag{28}$$

We observe that

$$(F(X), X)_{\mathcal{H}} = \left(\begin{pmatrix} -u_x - uu_x \\ 0 \end{pmatrix}, \begin{pmatrix} u \\ \psi \end{pmatrix}\right)_{\mathcal{H}} = \int_{\mathbb{R}}(-u_x - uu_x)udx = 0.$$

Hence (28) is written as

$$\frac{1}{2}\frac{d}{dt}\|X(t)\|^2_{\mathcal{H}} + \|X(t)\|^2_{D(\mathcal{A}^{1/2})} = 0.$$

We deduce that there exists a constant $C > 0$ such that

$$\forall t > 0, \ \int_0^t \|X(s)\|^2_{D(\mathcal{A}^{1/2})} ds \leq C\|X_0\|^2_{\mathcal{H}} = C\|u_0\|^2_{L^2(\mathbb{R})}. \tag{29}$$

Moreover, we take the scalar product in $\mathcal{H}$ of (11) with $\mathcal{A}X$. Since $\mathcal{A}$ is self-adjoint, we get

$$\frac{1}{2}\frac{d}{dt}\|X\|^2_{D(\mathcal{A}^{1/2})} + \|\mathcal{A}X\|^2_{\mathcal{H}} = (F(X), \mathcal{A}X)_{\mathcal{H}}. \tag{30}$$

Moreover, using Cauchy-Schwarz inequality and the estimation (25), we get

$$(F(X), \mathcal{A}X)_{\mathcal{H}} \leq \|F(X)\|_{\mathcal{H}}\|\mathcal{A}X\|_{\mathcal{H}} \leq (\|X\|^2_{D(\mathcal{A}^{1/2})} + \|X\|_{D(\mathcal{A}^{1/2})})\|\mathcal{A}X\|_{\mathcal{H}}.$$

Using Young inequality, we obtain

$$\begin{aligned}(F(X), \mathcal{A}X)_{\mathcal{H}} &\leq c(\|X\|^2_{D(\mathcal{A}^{1/2})} + \|X\|_{D(\mathcal{A}^{1/2})})^2 + \frac{1}{2}\|\mathcal{A}X\|^2_{\mathcal{H}} \\ &\leq c\|X\|^4_{D(\mathcal{A}^{1/2})} + \|X\|^2_{D(\mathcal{A}^{1/2})} + \frac{1}{2}\|\mathcal{A}X\|^2_{\mathcal{H}}.\end{aligned}$$

We deduce using (30) that for all $t \in [0, T]$

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|X(t)\|^2_{D(\mathcal{A}^{1/2})} &\leq c\|X(t)\|^4_{D(\mathcal{A}^{1/2})} + \|X(t)\|^2_{D(\mathcal{A}^{1/2})} \\ &\leq c\|X(t)\|^2_{D(\mathcal{A}^{1/2})}(\|X(t)\|^2_{D(\mathcal{A}^{1/2})} + 1).\end{aligned}$$

Then using Gronwall inequality, we get for all $t \in [0, T]$

$$\begin{aligned}\|X(t)\|^2_{D(\mathcal{A}^{1/2})} &\leq \|X_0\|^2_{D(\mathcal{A}^{1/2})} \exp(c\int_0^t (\|X(s)\|^2_{D(\mathcal{A}^{1/2})} + 1)ds) \\ &\leq \|X_0\|^2_{D(\mathcal{A}^{1/2})} \exp(ct) \exp(c\int_0^t \|X(s)\|^2_{D(\mathcal{A}^{1/2})} ds).\end{aligned}$$

Finally, taking in account the estimation (29), we deduce that for all $t \in [0, T]$

$$\|X(t)\|^2_{D(\mathcal{A}^{1/2})} \leq Ce^{ct}\|X_0\|^2_{D(\mathcal{A}^{1/2})}.$$

Let $T_{max} \in]0, +\infty]$ be the maximal existence time of the solution X of (11). We have

$$\forall t < T_{max}, \ \ \|X(t)\|^2_{D(\mathcal{A}^{1/2})} \leq Ce^{ct}\|X_0\|^2_{D(\mathcal{A}^{1/2})}. \tag{31}$$

We deduce that $T_{max} = +\infty$.

3. Conclusion

In this article, we investigate theoretically the well-posedness of an asymptotical water wave model with a nonlocal viscous term described by the Riemann-Liouville half derivative. Here we present the half-derivative using a diffusive realization. We proved the existence and the uniqueness of solutions for all initial data $u_0 \in H^1(\mathbb{R})$. A challenging issue is to study theoretically and numerically the decay rate of solutions for this class of initial data. This question will be the subject of a future work.

4. References

[1] J. AUDOUNET, V. GIOVANGIGLI, J. ROQUEJOFFRE, "A threshold phenomenon in the propagation of a point-source initiated flame", *Physica D*, vol. 121, 1998.

[2] H. BREZIS, " Analyse Fonctionnelle : Théorie et Applications ", *Editions Masson, Paris*, 1983.

[3] T. CAZENAVE , A. HARAUX, " An Introduction to Semilinear Evolution Equations ", *Oxford University Press, Oxford*, 1998.

[4] M. CHEN, S. DumontL. DUPAIGNE, O. GOUBET " Decay of solutions to a water wave model with a nonlocal viscous dispersive term ", *Discrete Contin. Dyn. Syst.*, vol. 27, 2010.

[5] D. DUTYKH, " Viscous-potential free-surface flows and long wave modelling ", *Eur. J. Mech. B Fluids*, vol. 28, 2009.

[6] D. DUTYKH , F. DIAS, " Viscous potential free surface flows in a fluid layer of finite depth *C.R.A.S, Série I*, vol. 345, 2007.

[7] T. HELIE, D. MATIGNON, " Diffusive reprentations for the analysis and simulation of flared acoustic pipes with visco-thermal losse ", *Mathematical Models and Methods in Applied Sciences*, vol. 16, 2006.

[8] P. LIU, A. ORFILA " Viscous effects on transient long wave propagation " *J. Fluid Mech.*, vol. 520 , 2004.

[9] I. MANOUBI, " Modèle visqueux asymptotique pour la propagation d'une onde dans un canal " *PhD Thesis*, 2014.

[10] I. MANOUBI, " Theoretical and numerical analysis of the decay rate of solutions to a water wave model with a nonlocal viscous dispersive term with Riemann-Liouville half derivative " *Discrete Contin. Dyn. Syst*, vol. 19, 2014.

[11] D. MATIGNON, C. PRIEUR, " Asymptotic stability of linear conservative systems when coupled with diffusive systems", *ESAIM: Control, Optim. and Calc. of Var.*, vol. 11, 2005.

[12] G. MONTSENY , " Diffusion monodimensionnelle et intégration d'ordre $1/2$ ", *Internal LAAS Report N.* 91232, 1991.

[13] G. MONTSENY, J. AUDOUNET , D. MATIGNON, " Diffusive representation for pseudodifferentially damped nonlinear systems ", *Nonlinear control in the year 2000*, vol. 2, 2000.

[14] G. MONTSENY, J. AUDOUNET , B. MBODGE, " Modèle simple d'amortisseur viscoélastique. Application à une corde vibrante " , *Lecture notes in Control and Information Sciences, Eds. RF.Curtain, A.Bensoussan, JL.Lions-Springer Verlag*, vol. 185 (1993),

[15] G. MONTSENY, J. AUDOUNET , B. MBODGE, " Optimal models of fractional integrators and application to systems with fading memory ", *IEEE International Conference on Systems, Man and Cybernetics, Le Touquet France*, 1993.

[16] O. STAFFANS , " Well-posedness and stabilizability of a viscoelastic equation in energy space ", *Transactions of the American Mathematical Society*, vol. 345, 1994.

ARIMA-CARI'2016

Mathematical modeling of fouling membrane in an anaerobic membrane bioreactor

BENYAHIA Boumediene[(a)], CHARFI Amine[(b)], HARMAND Jérôme[(c)], BEN AMAR Nihel[(b)], CHERKI Brahim[(a)]

[(a)] Tlemcen Automatics Laboratory, University of Tlemcen, B.P. 230, Tlemcen 13000, ALGERIA (b.benyahia.ut1@gmail.com ; b.cherki@gmail.com)

[(b)] ENIT-LAMSIN, BP 37, 1002 Tunis and, INSAT Centre Urbain Nord BP 676 - 1080, University of Tunis El Manar, Tunis Cedex, TUNISIA
(amine.charfi@ymail.com ; nihel.benamar@insat.rnu.tn)

[(c)] INRA, UR050, LBE-INRA, Avenue des étangs, Narbonne F-11100, FRANCE
(jerome.harmand@supagro.inra.fr)

RÉSUMÉ. Dans ce travail, nous proposons un modèle mathématique simple de colmatage membranaire et nous le couplons à un modèle simple de digestion anaérobie (BioRéacteur Membranaire Anaérobie). Par simulation numérique, nous étudions le comportement qualitatif du modèle et nous montrons des résultats préliminaires sur des stratégies de contrôle possible pour limiter le colmatage.

ABSTRACT. This paper deals with the development of a simple model of membrane fouling and its integration with a simple anaerobic digestion model (Anaerobic Membrane BioReactor). Using numerical simulations, we investigate the qualitative behavior of the model and we show some preliminary results of possible control strategies to limit fouling.

MOTS-CLÉS : BioRéacteur Membranaire Anaérobie, Modélisation des MBRs, Colmatage de membrane, SMP, Traitement des eaux usées

KEYWORDS : Anaerobic Membrane BioReactor, MBR modeling, Membrane Fouling, SMP, Wastewater treatment

1. Introduction

Anaerobic Membrane BioReactor (AnMBR) is an interesting wastewater treatment technology, allowing to obtain a highly purified effluent. Such processes have integrated models : biological dynamics models coupled to membrane filtration models. In MBRs, specific components as Soluble Microbial Product (SMP) dynamics play an important role in membrane fouling [2] and they must be added to the process model, as it was proposed in [1], in order to properly describe the entire MBR dynamics. If a number of such integrated models have been proposed for aerobic MBRs (cf. for instance [3, 4]), very few have been proposed for AnMBRs for studying their behavior or control purpose ([5], [6], [7]). The aim of the present paper is to propose a simple and generic membrane fouling model which the usefulness is illustrated in coupling it with a simple anaerobic model [1], to completely describe an AnMBR for control design purposes. Qualitative behavior of the system is investigated and some control strategies are discussed.

2. Mathematical model

The idea is to adapt the model proposed in [8], which it is not suitable for control purposes since it is too complicated, in order to include a feedback of the decreasing flux due to membrane fouling into the actual output flow rate $Q_{out}(t)$ leaving the MBR. We propose to consider $Q_{out}(t)$ as a decreasing function of the total mass solids attached onto the membrane surface and of the solute (as SMP) deposited inside the pores, which are the two main membrane fouling mechanisms considered in this work. Under some realist assumptions used for building the membrane model, this later is given in the following for two functioning periods : filtration and relaxation.

2.1. Fouling model for the filtration phase ($\Delta P > 0$)

The filtration phase model is given by equations (1)-(5). It predicts the output flow rate Q_{out} as a decreasing function : when the permeate flux dramatically decreases, the process must be stopped and backwash or cleaning of the membrane must be realized.

$$\dot{m} = \delta Q_{out}(C_s S_T + C_x X_T + C_{smp} SMP), \quad [1]$$

$$\dot{S_p} = \delta' Q_{out}\Big(\beta.SMP + f(S_T)\Big), \quad [2]$$

$$R = \alpha\frac{m}{A} + \alpha'\frac{V_p S_p}{\epsilon A}, \quad [3]$$

$$A = \frac{A_0}{1 + \frac{m}{\sigma} + \frac{S_p}{\sigma'}}, \quad [4]$$

$$Q_{out} = J.A = \frac{\Delta P.A}{\mu\big(R_0 + R\big)}. \quad [5]$$

Where $m(t)$ the mass of solids attached onto the membrane surface, $S_p(t)$ the particles (as SMP) retained inside the membrane pores. Dynamics of these variables depend on soluble components $S_T(t)$, particulate components $X_T(t)$ and $SMP(t)$, all coming from reactional medium, with C_s, C_x and C_{smp} are weighting parameters used to model the contribution of each component to the membrane fouling, β the fraction of SMP leaving

the MBR (see [1] for more details), $f(S_T)$ is a function used to model the contribution of S_T to the pores clogging, δ and δ' are weighting parameter used to calibrate the rate of the fouling (cake formation and pores clogging). $R(t)$ the total fouling resistance defined as the sum of the cake resistance ($R_m(t) = \alpha\frac{m}{A}$) depending essentially on $m(t)$, and the pores clogging resistance ($R_s(t) = \alpha'\frac{V_p S_p}{\epsilon A}$) which is assumed to be due mainly to $S_p(t)$, with $A(t)$ the total membrane area, ϵA the porous surface of A, V_p the total volume of the pores, α and α' the specific resistances, A_0 the initial membrane surface, σ and σ' parameters in appropriate units. $J(t)$ the permeate flux, $\Delta P(t)$ the transmembrane pressure, μ the permeate viscosity and R_0 the intrinsic membrane resistance.

We consider that the total filtering membrane surface $A(t)$, is not constant during a filtration period nor after several filtration/stop cycles : it is described in a very general way as a decreasing function of $m(t)$ and $S_p(t)$, as the possible function of (4). Here, $A(t)$ tends to zero as $m(t)$ and/or $S_p(t)$ tend to infinity. The function (4) is also able to model the fact that the initial filtering surface A_0 is not totally recovered after a backwash or a chemical cleaning, because it will be small remaining quantities of $m(t)$ and $S_p(t)$ which are not detached, causing an irreversible fouling effect, and thus $A(t) < A_0$.

2.2. Fouling model for the relaxation phase ($\Delta P = 0$)

The flux is simply stopped ($\Delta P = 0$) allowing the natural detachment of matters and particles. The model is simply given by :

$$\dot{m} = -f_m(m), \quad [6]$$

$$\dot{S_p} = -f_s(S_p), \quad [7]$$

For instance, we can choose $f_m(m) = \omega m(t)$ and $f_s(S_p) = \omega' S_p(t)$, with ω and ω' positive constants to be adjusted with respect to experimental data. The relaxation time is neglected compared to the filtration time and it is expected that one has always a certain percentage of attached matter which may remain onto the membrane surface and/or blocked inside the pores, yielding to irreversible fouling.

3. Investigating of qualitative behavior

To investigate the qualitative behavior of the system, we must integrate the fouling model (1)-(7) with a biological anaerobic model as illustrated in Fig. 1. For the biological compartment, we suggest to use the AM2b model which includes SMP dynamics and that has been precisely developed for control purposes [1]. Whatever the considered biological model, its output variables (soluble and particulate matters S_T, X_T, SMP, ...) are injected as inputs for the model (1)-(5).

We perform numerical simulations using parameters values given in Table 1, and we consider two functioning phases : filtration for $2h$ and relaxation for $5min$. Such sequence is probably not optimized and is quite far from an optimal adjustment, which remains an open problem of fouling control.

Simulation results are reported in Fig. 2, where we have plotted the dynamic evolution of the attached mass $m(t)$ on the membrane surface, the blocked soluble matter $S_p(t)$ (SMP in the majority) inside the pores, the fouling resistances $R_m(t)$, $R_s(t)$ and $R(t)$, the output flow rate $Q_{out}(t)$, the permeate flux $J(t)$ and the membrane surface $A(t)$. Dynamic responses are simulated for three different values of both parameters $\delta = (5; 25; 50)$ and $\delta' = (0.1; 0.75; 1.5)$, to emphasize effects of deposited and blocked matter rates on the

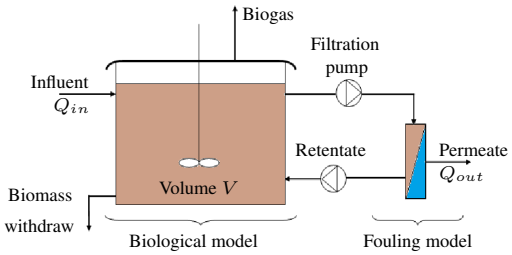


Figure 1. *Schematic representation of the proposed AnMBR model*

Table 1. *Parameter values used in simulations*

Parameter	value	Parameter	value	Parameter	value	Parameter	value
β	0.6	σ'	10	C_x	0.05	μ	0.001
V	50	α	1e10	C_{smp}	0.005	A_0	1
V_p	1.4	α'	1e10	δ	5,20,50	ΔP	0.25
σ	10	C_s	0.005	δ'	0.1,0.75,1.5	R_0	1.11e13

fouling dynamic. These rates depend on many parameters as concentrations of soluble and particulate matters, characteristics of mixed liquor and its viscosity or still temperature and matters specific capability to contribute to fouling.

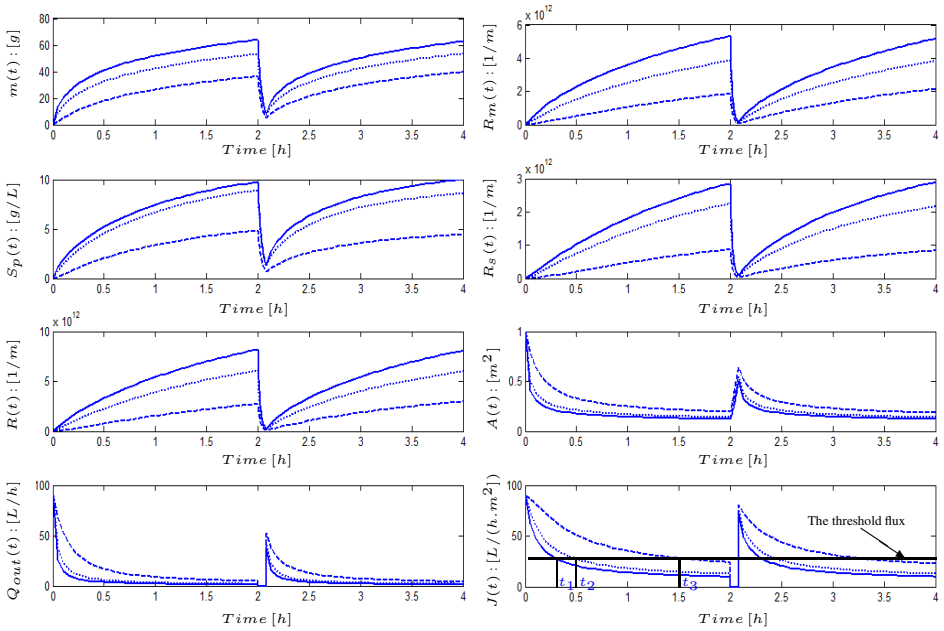


Figure 2. *Simulation results of the membrane fouling model for both phases (filtration and backwash).*

During the first minutes of the filtration process, the fouling is fast and significant. All variables have fast dynamics (increasing or decreasing) at the beginning and then attain progressively (with a decreasing rate) their equilibria (steady state). This can be explained by the fast clogging of pores which occurs firstly, before that the cake formation increases

in a second time and prevents pores fouling (slow fouling phenomenon). We emphasize here that the useful filtering surface $A(t)$, the output flow $Q_{out}(t)$ and the permeate flux $J(t)$, decrease significantly, especially during first minutes of filtration as it is often the case in practice.

The trajectories of the main variables are plotted in the case of a slight and strong fouling. Solids plots correspond to a strong fouling due, for example, to a high concentration of solid matter. Dashed and doted plots correspond to a slower and softer fouling respectively : slower the fouling, longer the time period the process may operate without switching in a relaxation mode. For instance, if we define a threshold flux over which the process can operate (see sub-figure in bottom-right), then the process will be stopped very often and be switched in relaxation phase for strong fouling (t_1 is small, solids plots). In the case of slower fouling, the process will be switched less frequently to relaxation mode (t_3 is large, dashed plots). Such simulations show that δ and δ' may be adjusted to match a large range of experimental data.

4. Preliminary results on some control strategies

Membrane fouling is the major drawback of MBRs and one important challenge is to propose new control strategies to minimize fouling and improve treatment efficiency. Very often, the control strategies are tuned heuristically and use available process actuators : gas sparging, intermittent filtration and backwash (or relaxation). In the following, we investigate in simulation the influence of the previous filtering parameters on the flux production and process performances, by using the simple model (1)-(5) and (6)-(7).

4.1. Influence of the gas sparging

In this section, we investigate how gas sparging can be used for limiting membrane fouling. To do so, we need to modify the proposed model (1)-(5) in adding negative terms on the right sides of equations (1) and (2). This way, the fouling rates are reduced by gas sparging as illustrated by equations (8) and (9), where functions $f(m)$ and $g(S_p)$ are positive and depending on the intensity of gas sparging (parameter control).

$$\dot{m} = \delta Q_{out}(C_s S_T + C_x X_T + C_{smp} SMP) - f(m), \quad [8]$$

$$\dot{S}_p = \delta' Q_{out}\Big(\beta SMP + f(S_T)\Big) - g(S_p). \quad [9]$$

A first simple form of $f(m)$ and $g(S_p)$ which is already used in the literature is $k_m m$ and $k_{S_p} S_p$, which represent quantities of m and S_p detached by shear forces caused by membrane scouring, where k_m and k_{S_p} depend on the intensity of injected bubbles used to detach fouling [5]. Fig. 3, illustrates time evolution of the flux $J(t)$ with respect to different values of k_m (here $k_{S_p} = 0$, it is assumed that the irreversible fouling detachment is neglected, since it is not significantly affected by gas sparging). It can be seen that $m(t)$ and $R_m(t)$ are inversely proportional to the control parameter k_m, for higher values of this later, accumulated matter on the membrane surface and its corresponding resistance take small values. Output flow $Q_{out}(t)$ and permeate flux $J(t)$ are increasing proportionally to k_m during the first minutes of filtration (until *0.6h*). On the other hand, one sees on Fig. 3, that deposited matters $S_p(t)$ inside the pores and its relative resistance $R_s(S_p)$ are proportional to k_m and inversely proportional to $m(t)$. If the value of this parameter increases, then the quantity of $S_p(t)$ and the value of $R_s(S_p)$ increase likewise leading

to a flux loss at the end of the filtration time (around steady-state). One can explain this result as follows : it is known in the literature that the cake layer formed by $m(t)$ represents a second biological membrane, preventing the pores fouling by $S_p(t)$ [4]. When this layer detaches, more particles of different sizes go through pores and cause further fouling. Which control strategy can favour the cake formation until acceptable level, to protect pores from fouling, but at the same time, without influencing permeate flux ? This question, actually, remains open.

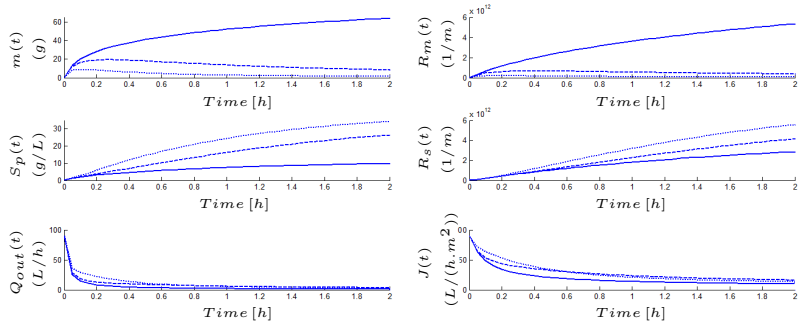


Figure 3. *Results simulation of the membrane fouling model with control terms using (8)-(9), solid :* $k_m = 0$*, dash :* $k_m = 5$*, dot :* $k_m = 25$*, (*$k_{S_p} = 0$*).*

4.2. Influence of the number of filtration/relaxation (backwash) cycles per time unit

Given a sufficiently large time horizon, what is the optimal number of filtration/relaxation or backwash cycles allowing a higher mean value for the MBR output flux ? To illustrate the importance of this functioning mode, we are particularly interested by the mean value J_{mean} of the produced flux on the given period of $2h$ on which, we performed numerical simulations by changing the number of filtration/relaxation cycles with a constant ratio between filtration time and relaxation time $\alpha_t = \frac{T_{filtr}}{T_{Relax}} = 7$ for all cycles. On Fig. 4, results are given for :

– 1 cycle : $T_{filtr} = 105mn, T_{Relax} = 15mn \Rightarrow J_{mean} = 17.9\ L/(h.m^2)$,

– 2 cycles : $T_{filtr} = 52.2mn, T_{Relax} = 7.5mn \Rightarrow J_{mean} = 22.9\ L/(h.m^2)$:

– 5 cycles : $T_{filtr} = 21mn, T_{Relax} = 3mn \Rightarrow J_{mean} = 29\ L/(h.m^2)$,

– 10 cycles : $T_{filtr} = 10.5mn, T_{Relax} = 1.5mn \Rightarrow J_{mean} = 31.5\ L/(h.m^2)$.

It can be seen that higher the number of cycles, higher the produced mean flux on the given period. A functioning frequency of 10 filtration cycles appears to be the best strategy, since it produces the higher mean flux $J_{mean} = 31.5\ L/(h.m^2)$. But if the number of intermittent filtration cycles is too large on the considered functioning period, then it can damage the process by forcing it to operate very frequently in On/Off mode. It is thus suggested not to wait too long before proceeding to the membrane cleaning by relaxation (or backwash) and to find the best ratio for operated time by benefit in terms of flux produced.

4.3. Coupling sparging gas and intermittent filtration controls

Our idea here is to minimize the energy consumption when using gas sparging and the flux loss (resp. the permeate loss) when the process is in relaxation mode. In others words,

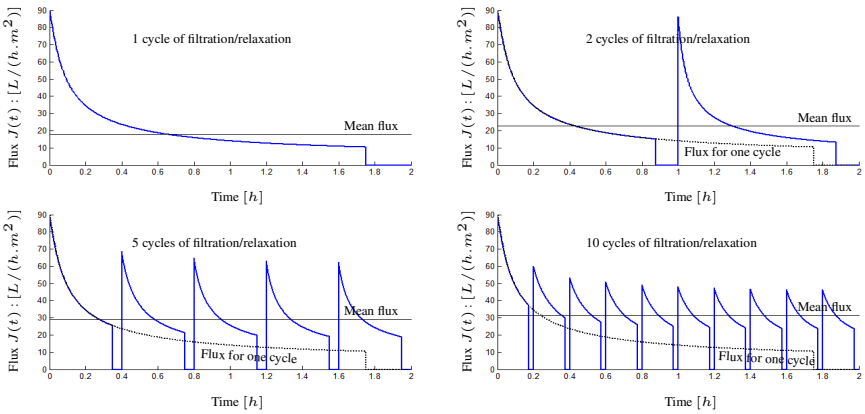


Figure 4. *Results simulation of different numbers of filtration/relaxation cycles.*

instead of using gas sparging and intermittent filtration simultaneously, we propose to use them sequentially for the following reasons :

– Gas sparging is used to detach the matter deposited on the membrane at the beginning of the filtration (fouling is soft and not yet dense).

– Intermittent relaxation is used to detach a denser fouling (strong), which can occur after an enough long functioning time.

To illustrate this idea, we performed numerical simulations plotted in Fig. 5. The system is first simulated without any control (black plot). Then this reference scenario is compared with the proposed coupled control (blue plot). It means that gas sparging is first applied until the flux reaches the threshold flux (here $J_s = 18\ L/(h.m^2)$). At this instant ($t = 0.64h$), we apply intermittent control with $k_m = 5$ in the equation (8), where $f(m) = k_m m$ with 4 cycles.

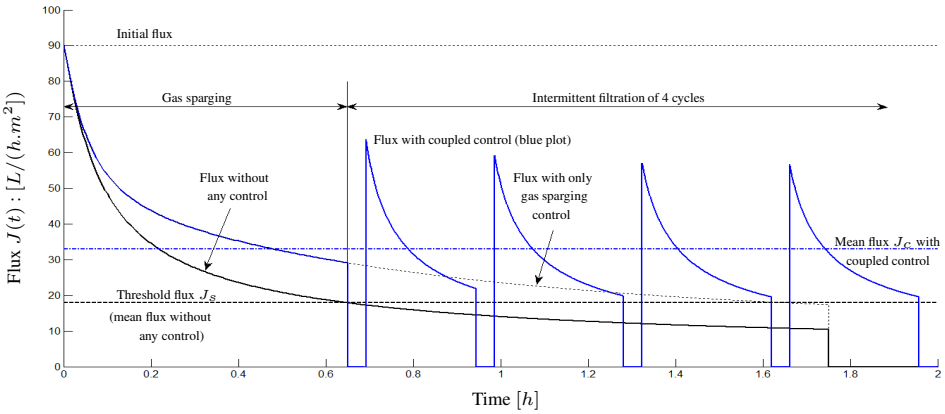


Figure 5. *Coupling control based on gas sparging and intermittent filtration.*

Simulations show that this control strategy allows one to increase the mean production flux to 33 $L/(h.m^2)$, whereas the mean flux without control was 18 $L/(h.m^2)$. As it is noticed in Fig. 5, when applying the gas sparging control, it has increased favorably the permeate flux on the control period (until $0.64h$). It should be noticed that even if we applied only the gas sparging all along the functioning period, without using inter-

mittent filtration cycles (see black doted plot), the mean flux is $28.76\ L/(h.m^2)$, lower than the produced flux when the two techniques are used together (see blue dashed plot). Thus, intermittent filtration was an appropriate control strategy to obtain over the whole functioning period a maximum of flux, while optimizing the energy.

Our study on control strategy is obviously inline with other studies as the work presented in [7]. Their main purpose was to investigate and select the best operating conditions in terms of aeration intensity, duration of filtration/backwashing cycles and number of membrane cleaning to optimize energy demand and operational costs.

5. Conclusion

In this paper we proposed a simple fouling model of AnMBR. The model was developed under certain classical hypotheses on the membrane fouling phenomena, by taking into account two fouling mechanisms and, was coupled with a reduced order anaerobic digestion model. It was shown by simulation that the proposed model can predict quite well the fouling behavior for the considered AnMBR. In a second part of the paper, preliminary results were obtained about the results of different control strategies over a given time period : at the beginning stage of the process functioning, it appeared useful to use the gas sparging and the intermittent filtration at the end of the considered time period. Based on these results, we proposed to couple control benefits in order to produce the maximum mean flux over the total considered functioning period.

6. Bibliographie

[1] BENYAHIA, B., SARI, T., CHERKI, B, AND HARMAND, J., « Anaerobic membrane bioreactor modeling in the presence of Soluble Microbial Products (SMP) - the Anaerobic Model AM2b », *Chemical Engineering Journal*, vol. 228, pp 1011–1022, 2013.

[2] MENG, F., CHAE, S.R., DREWS, A., KRAUME, M., SHIN, H.S. AND YANG, F., « Recent advances in membrane bioreactors (MBRs) : Membrane fouling and membrane material », *Water Research*, vol. 43, pp 1489–1512, 2009.

[3] LEE, Y., CHO, J., SEO, Y., LEE, J.W. AND AHN, K.H., « Modeling of submerged membrane bioreactor process for wastewater treatment », *Desalination*, vol. 146, pp 451-457, 2002.

[4] DI BELLA, G., MANNINA, G. AND VIVIANI, G., « An integrated model for physical-biological wastewater organic removal in a submerged membrane bioreactor : Model development and parameter estimation », *Journal of membrane science*, vol. 322(1), pp 1–12, 2008.

[5] LIANG, S., SONG, L., TAO, G., KEKRE, K.A. AND SEAH, H., « A modeling study of fouling development in membrane bioreactors for wastewater treatment », *Water environment research*, vol. 78(8), pp 857–863, 2006.

[6] ROBLES, A., RUANO, MV., RIBES, J., SECO, A. AND FERRER, J., « A filtration model applied to submerged anaerobic MBRs (SAnMBRs) », *Journal of membrane science*, vol. 444, pp 139–147, 2013.

[7] MANNINA, G., COSENZA, A., « The fouling phenomenon in membrane bioreactors : assessment of different strategies for energy saving », *Journal of membrane science*, vol. 444, pp 332–344, 2013.

[8] LI, X., WANG, X., « Modelling of membrane fouling in a submerged membrane bioreactor », *Journal of membrane science*, vol. 278, pp 151–161, 2006.

Mathematical modelling of intra-clonal heterogeneity in multiple myeloma

A. Bouchnita[1,2], F. E. Belmaati[1], R. Aboulaich[1], R. Ellaia[1], V. Volpert[2]

[1]Laboratoire d'Etude et de Recherche en Mathématiques Appliquées (LERMA)
Engineering Mohammadia School
Mohammed V University
Rabat - Agdal
Morocco
e-mails: fati.belmaati@gmail.com, aboulaich@gmail.com, rachid.ellaia@gmail.com

[2]Institut Camille Jordan (ICJ)
Université Lyon 1
Villeurbanne
France
e-mails: anass.bouchnita@univ-lyon1.fr, volpert@math.univ-lyon1.fr

RÉSUMÉ. Cette étude est consacrée à la modélisation mathématique de l'hétérogénéité intra-clonale du myélome multiple (MM) et de sa résistance aux médicaments qui en résulte. Pour explorer les mécanismes inhérents qui régulent ce processus, nous développons un modèle hybride multi-échelles de la croissance des tumeurs MM dans la moelle. Les cellules malignes sont représentées par approche individuelle. L'action du traitement est introduite. La tumeur consiste en des clones en compétition. Le taux de division des cellules dans un clone dépend de sa compétition avec les autres. Nous étudions la dynamique de l'hétérogénéité intra-clonale dans le MM et nous décrivons son rôle dans l'émergence de phénotypes plus résistants au traitement.

ABSTRACT. This study is devoted to the mathematical modelling of multiple myeloma (MM) intra-clonal heterogeneity and the resulting drug resistance. To explore the underlying mechanisms of intra-clonal heterogeneity, we develop a multi-scale hybrid model of MM tumor growth in the bone marrow. Malignant plasma cells are represented by individual based approach. Drug action is introduced and its concentration inside each cell is described by an ordinary differential equation. The tumor consists of competing clones. The rate of cell division in each clone depends on the competition with the other clones. We study the dynamics of intra-clonal heterogeneity in MM and describe its role in the emergence of drug resisting phenotypes.

MOTS-CLÉS : myélome multiple; hétérogénéité intra-clonale; résistance aux médicaments, modélisation mathématique

KEYWORDS : multiple myeloma; intra-clonal heterogeneity; drug resistance; mathematical modelling

1. Introduction

Multiple myeloma (MM) is a malignancy characterized by the infiltration of cancerous plasma cells into the bone marrow. These cells form multiple tumors that expand and secrete apoptosis inducing cytokines which eliminate erythroid cells resulting in anemia. As in other cancers, MM cells undergo various mutations and the tumor is formed by different clones [1]. This feature is known as intra-clonal heterogeneity. It is related to the adaptation and natural selection of cancer cells. Malignant cells compete for limited nutrients, and more adapted cells survive and multiply. In addition to this selective pressure, cancer treatment can act as an additional factor which favors the survival of some clones more than others. While there are efficient treatment regimens of MM, drug resistance remains the major concern. In this regard, the resisting clones may be initially present in the first cells that infiltrate the bone marrow, but they can also emerge during treatment leading to relapse. The emergence of novel clones is due to the MM progression in branching pattern discussed below.

Mathematical models of cancer growth and intra-clonal heterogeneity falls in three main categories. The first one is continuous models. These are deterministic models that use partial differential equations to describe cancer development [7] and treatment [10]. Another type of models uses the discrete approach to describe cancer growth. These can be lattice [12] or off-lattice models [9]. The question of stress-induced drug resistance in tumors was also studied in some works [7]. Finally, hybrid models combine continuous and discrete approaches where cells are considered as individual objects, intracellular concentrations are described with ordinary differential equations and extracellular concentrations with partial differential equations [4].

Modelling methods previously developed to study hematopoiesis and blood diseases [5, 6] will be adapted in this work to study MM intra-clonal heterogeneity and drug resistance. In this approach, each cell is represented as an elastic sphere that can move due to the interaction with other cells. Cells can also divide or die by apoptosis. Each cell is characterized by its genotype which can change because of the mutations. When a cell divides, the daughter cells inherit the genotype of the mother cell with small random mutations. This leads to the emergence of new clones in the process of tumor growth. We use this approach to model the intra-clonal heterogeneity of MM. Furthermore, we apply it to study the emergence of drug resisting clones during chemotherapy.

2. The model

We consider a square computational domain with the side equal to 100 length units corresponding to 10 microns. Cells are represented by elastic spheres with initial diameters equal to one unit. They are removed from the domain when they reach its boundaries. We consider an initial tumor consisting of 208 malignant cells as initial condition with the same genotype. In the process of tumor growth, they can change their genotype due to mutations. Their rate of apoptosis depends on the competition between clones for resources.

2.1. Cells motion

We model cells as elastic spheres with an incompressible inner part and compressible outer part. Since cells divide, they push each other and can change their position. Cell motion is des-

cribed by Newton's second law for their centers. Let x_i be the coordinate of the center of the ith cell (two-component vector). Then we have the following equation for its motion :

$$m\ddot{x_i} + m\mu\dot{x_i} - \sum_{j \neq i} f_{ij} = 0, \tag{1}$$

where

$$f_{ij} = \begin{cases} K\dfrac{h_0 - h_{ij}}{h_{ij} - (h_0 - h_1)}, & h_0 - h_1 < h_{ij} < h_0 \\ \\ 0, & h_{ij} \geq h_0 \end{cases}. \tag{2}$$

Here f_{ij} is the force acting between cells i and j, h_{ij} is the distance between their centers, h_0 is the sum of their radii, K is a positive parameter and h_1 represents the incompressible part of each cell. The second term in Eq. (1) describes the friction by the surrounding medium. Cell radius increases in the process of cell division. More detailed description of the method can be found in [8].

2.2. Cells division and mutations

When the malignant cell reaches the end of its life cycle, it has two possible fates. Either it divides and self-renews giving rise to two daughter cells or it dies by apoptosis. The apoptosis probability is determined by cell genotype.

We characterize cell genotype by a real variable z. Let z_m be a cell genotype before division. After cell division, the genotype of the daughter cells can take three values,z_m, $z_m + \epsilon$, $z_m - \epsilon$ where ϵ is a small positive number. The choice between these three values is random with equal probability. Thus, the genotype of the daughter cell can be the same as the genotype of the mother cell or it differs from it by ϵ. This difference describes small random mutations after each division. If all cells have initially the same genotype z_0, then cell density distribution $u(z,t)$ with respect to the genotype becomes wider with time. The evolution of the function $u(z,t)$ can be described by the diffusion equation.

The probability of cell apoptosis depends on its genotype. We define viable cell clones by some intervals of genotype where apoptosis probability is less than the probablity of self-renewal. Consider the function $p(z)$ which determines the probability of apoptosis depending on the genotype. We set $p(z) = p_0$ for $z \in [a_i, b_i]$ and $p = p_1$ outside these intervals (Figure 1, a). Here $[a_i, b_i]$ with $i = 1..4$ are the intervals of genotype characterizing different clones, p_0 is the basic level of apoptosis of these clones. The ordering and distance between the clones in the function $p(z)$ mimic the moment of apparition of clones in experiments [13]. We consider the value p_0 sufficiently small in order for these cells to survive and multiply, p_1 is sufficiently close to 1. Then cell clones will survive while cells with different genotypes can appear due to mutations but they will mostly die after some time due to apoptosis.

Cell competition for resources increases their apoptosis. Hence apoptosis probability depends not only on cell genotype but also on the quantity of cells for different genotypes. We will specify this dependence below in the case of multiple myeloma.

In application to multiple myeloma, we will consider four cell mutations observed experimentally : ATM, FSIP2, GLMN, CLTC [13]. As a result, different clones emerge as shown in Figure 1, b. We denote these clones as c_1, c_2, c_3 and c_4. Clones c_1 and c_2 are sufficiently close to each other and they compete between themselves. Similarly, clones c_3 and c_4 are in competition

between each other [13]. We suppose that c_1 and c_2 do not compete with $c3$ and $c4$. We define the probability p_i of cell apoptosis for each clone as follows :

$$p_1 = p_0+2\alpha(u_1+u_2),\ p_2 = p_0+\alpha(u_1+u_2),\ p_3 = p_0+\alpha(u_3+u_4),\ p_4 = p_0+\alpha(u_3+u_4). \quad (3)$$

Here u_i are cell densities for each clone, $u_1 + u_1 + u_3 + u_4 = 1$, p_0 is the probability of cell apoptosis without competition for resources taken equal to 0.2, α is a positive number equal to 0.04. We note that apoptosis probability of the clone c_1 is greater than that of other clones. According to the biological data it is less adapted to the environment than the others. Apoptosis probabilities and the genotypes corresponding to different clones will be chosen in numerical simulations in order to fit the experimental data.

2.3. MM therapy and drug resistance

Multiple myeloma is treated by chemotherapy with myeloma specific drugs (thalidomide, lenalidomide and bortezomib), which kill malignant cells and do not influence other hematopoietic cells. Though chemotherapy treatment is efficient in reducing the number of MM cells, it does not eradicate them completely. In order to avoid relapse, chemotherapy is usually followed by bone marrow transplantation.

The intracellular drug concentration q_i in the i^{th} cell is described by the equation :

$$\frac{dq}{dt} = k_1 Q(t) - k_2 q, \quad (4)$$

where $Q(t)$ is the drug concentration in the bone marrow. We take it constant and equal to 0.7 for t during the administration and 0 elsewhere. The treatment is administrated in the first two week of each cycle of 28 days during a four cycle protocol after 25 days of tumor development. It depends on time according to the treatment protocol and it is supposed to be equally distributed in space. The first term in the right-hand side of this equation describes drug influx and the second term its degradation and efflux. The coefficients k_1 and k_2 can be different for different clones. If the intracellular drug concentration reaches some critical value q*, then the cell dies. In numerical simulations dead cells are removed from the computational domain.

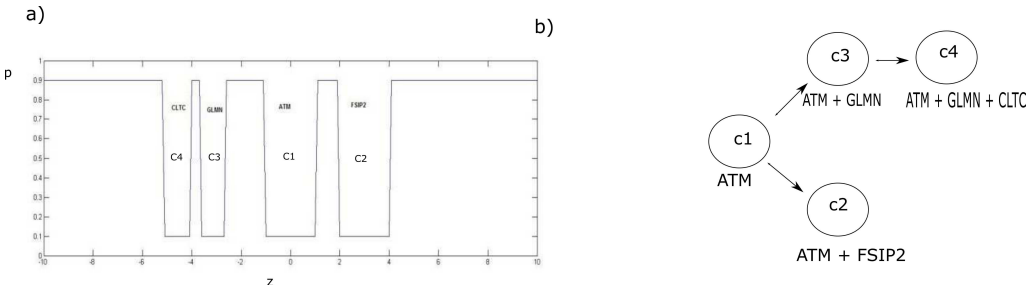


Figure 1. *(a) The apoptosis probability $p(z)$ as a function of genotype z. The four clones are shown. The values of their apoptosis probabilities (shown in dashed lines) are not fixed and depend on cell densities. (b) The branching pattern of multiple myeloma intra-clonal heterogeneity.*

3. Results

3.1. Intra-clonal heterogeneity and clones competition dynamics in multiple myeloma

MM is a genetically complex malignancy characterized by intra-clonal heterogeneity. Malignant myeloma cells undergo a number of mutations as the cancer progresses. We will compare here the results of our modeling with the biological data presented in [13]. In this work, MM intra-clonal heterogeneity and the presence of different coexisting clones were shown in the sequencing data. Furthermore, it was proven that more competitive clones emerge in the process of tumor growth. We use the genetic function model described in the previous section. We consider a population of malignant cells which initially belongs to clone c_1. As the simulation progresses, new clones emerge. The size of the clone c_1 population increases in the beginning. After some time, as clone c_2 emerges and starts expanding, clone c_1 declines since its apoptosis rate is greater than for clone c_2 (Figure 2, a). Clone c_3 emerges independently of clone c_2 and later than c_2

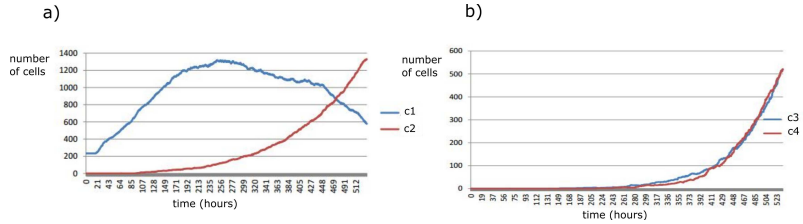


Figure 2. *Size of cell populations for clones c_1 and c_2 (a) and for clones c_3 and c_4 (b) over time.*

since its genetic distance from clone c_1 is larger. Clone c_4 appears from c_3 due to an additional mutation. As we discussed above, clones c_1 and c_2 compete with each other as well as clones c_3 and c_4. The numbers of cells in these clone in time are shown in Figure 2 and snapshots of growing tumor in Figure 3.

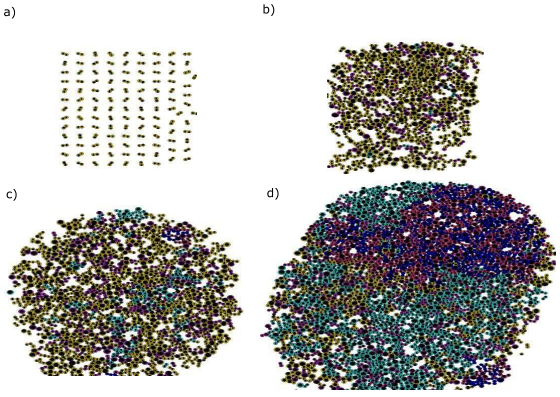


Figure 3. *Snapshots of the simulation with different stages of MM progression : (a) the initial cell population belongs to clone c_1 (yellow cells), (b) emergence of clone c_2 (cyan) followed by appearance of cells c_3 (magneta), (c) clones c_2 and c_3 form sub-populations across the tumor, (d) the tumor now consists primarily of clones c_2, c_3 and recently emerged clone c_4 (blue). The few cells that do not belong to any clone are also shown (purple).*

3.2. Intra-clonal heterogeneity role in MM drug resistance

To assess the tumor response to therapy, we suppose that the toxic effect of the drug on MM cells is different for each clone. Therefore the coefficients k_1 and k_2 in Eq. 4 depend on clone type. We suppose that the administrated drugs are more prone to eliminate the initial clone c_1 but are less efficient in eliminating the cells of $c_2,\ c_3,\ c_4$. We set $k_{c1,1} > k_{ci,1}, i = 2, 3, 4$. Treatment is administrated when tumor is formed and clone $c1$ is predominant while the other clones are only emerging. The overall population of malignant cells is compared with the population of clone c_1 in Figure 4.

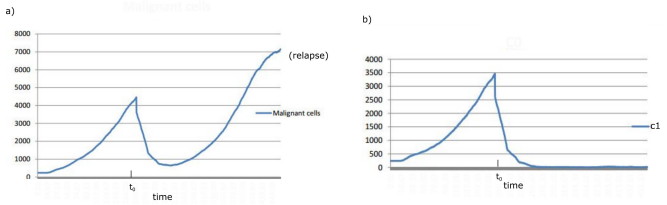


Figure 4. *The total population of malignant cells (left) and the population of the clone $c1$ cells (right) over time. Clone $c1$ disappears due to treatment while other more resistant clones emerge and multiply in spite of treatment.*

At the pre-treatment stage, the tumor grows with an exponential rate. Other clones have emerged from the initial cells and, thus, the tumor is no longer homogenous. By the end of the first cycle of therapy, the cells of the clone c_1 were completely eliminated while cells from the other clones have survived. The remaining cells form separate niches. Each niche consists of cells of the same clone. These cells take advantage of the rest period between chemotherapy cycles to divide and form independent tumors. These recently formed tumors are more resistant to treatment and they keep growing even after the beginning of the new cycle of therapy. After some time they form a single large tumor. Different stages of tumor grows are shown in Figure 5.

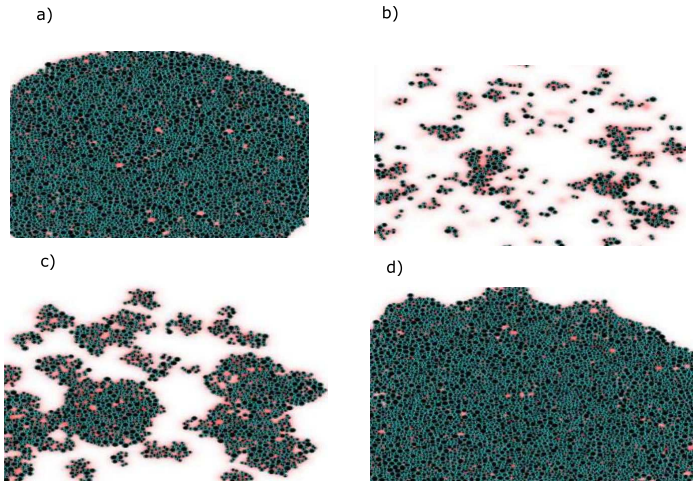


Figure 5. *Snapshots of a simulation of myeloma tumor growth under treatment : (a) the tumor reaches its maximal mass before the treatment, (b) the drugs eliminate the cells of clone c_1, the cells belonging to other clones survive and form separate niches, (c) the niches formed by the remaining cells consolidate and form independant tumors, (d) the tumors keep growing and join together in a one single tumor.*

4. Discussion

The heterogeneous nature of MM and drug resistance of the emerging clones represents a difficulty in the MM therapy. Different clones have different sensitivities to treatment and to the other components of the microenvironment. The heterogeneous property of MM usually leads to the relapse when treatment is finished. To understand the dynamics of clones competition and its impact on therapy resistance, we have developed a multi-scale model of myeloma tumor growth. We used this model to simulate the emergence of cell clones as observed in [13]. The model reproduces these phenomena not only qualitatively but also quantitatively. To quantify the results of the simulations and to compare them with the experiments, we introduce a mutation frequency variable (m) that corresponds to a scaling from 1 to 0 of the genetic variable z. It represents the inverse of the number of mutations undergone by the cell. We show the kernel density plot based on this variable in Figure 6. This plot allows the estimation of the general distribution of global mutational frequency in a population using a sample of cells. The results are in good agreement with the experimental data (Figure 4, b in [13]).

Biological observations show that cancer and mutations are reversible[11]. Hence the emergence of resistant clones is a reversible process. This property is taken into account in our model and it was observed in the simulations when new clones emerge. It can also be related to relapse when eliminated clones reemerge after the end of treatment. In order to prevent relapse, new therapeutical strategies were developed in MM treatment. In this context, sequential therapy was used as an induction followed by consolidation and maintenance [3]. In the induction phase, a part of the tumor is surgically removed to reduce its mass. Consolidation therapy is then used to eliminate cells belonging to all different clones. The remaining clonal cells are treated by maintenance therapy in which treatment is modified in order to eradicate the different clones.

The model presented here reproduces the main features of MM intra-clonal heterogeneity. More detailed intracellular and extracellular regulations and their influence on the emergence and competition of different clones will be studied in subsequent works.

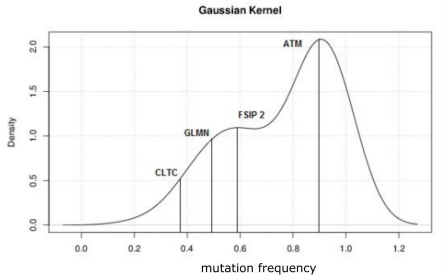


Figure 6. *Kernel density plot of heterogenous MM population at a certain moment of time during simulation. This distribution is similar to the experimentally observed distribution in [13].*

5. Bibliographie

[1] Anderson, K. C. "New insights into therapeutic targets in myeloma." ASH Education Program Book 2011.1 (2011) : 184-190.

[2] Barlogie, B., et al. "Total therapy with tandem transplants for newly diagnosed multiple myeloma." Blood 93.1 (1999) : 55-65.

[3] Brioli, A., et al. "The impact of intra-clonal heterogeneity on the treatment of multiple myeloma." British journal of haematology 165.4 (2014) : 441-454.

[4] Basanta, D., et al. "The Role of Transforming Growth Factor-β-Mediated Tumor-Stroma Interactions in Prostate Cancer Progression : An Integrative Approach." Cancer research 69.17 (2009) : 7111-7120.

[5] Bouchnita, A., et al. "Normal erythropoiesis and development of multiple myeloma." ITM Web of Conferences. Vol. 5. EDP Sciences, 2015.

[6] Bouchnita, A., et al. "Bone marrow infiltration by multiple myeloma causes anemia by reversible disruption of erythropoiesis." American journal of hematology (2016).

[7] Chisholm, R. H., et al. "Emergence of drug tolerance in cancer cell populations : An evolutionary outcome of selection, nongenetic instability, and stress-induced adaptation." Cancer research 75.6 (2015) : 930-939.

[8] Eymard, N., et al. "The role of spatial organization of cells in erythropoiesis." J. Math. Biol (2014).

[9] Galle, J., et al. "Individual cell-based models of tumor-environment interactions : Multiple effects of CD97 on tumor invasion." The American journal of pathology 169.5 (2006) : 1802-1811.

[10] Jackson, T. L., and Helen M. B. "A mathematical model to study the effects of drug resistance and vasculature on the response of solid tumors to chemotherapy." Mathematical biosciences 164.1 (2000) : 17-38.

[11] Keats, J. J., et al. "Clonal competition with alternating dominance in multiple myeloma." Blood 120.5 (2012) : 1067-1076.

[12] Piotrowska, M. J., and Simon D. A. "A quantitative cellular automaton model of in vitro multicellular spheroid tumour growth." Journal of theoretical biology 258.2 (2009) : 165-178.

[13] Walker, B. A., et al. "Intraclonal heterogeneity and distinct molecular mechanisms characterize the development of t (4 ; 14) and t (11 ; 14) myeloma." Blood 120.5 (2012) : 1077-1086.

What is the impact of disease-induced death in a Predator-Prey model experiencing an infectious disease ?

Valaire Yatat Djeumen[a, d, e, *]- JJ. Tewa[b, d, e]- S. Bowong[c, d, e]

a,* Department of Mathematics, University of Yaoundé I, PO Box 812 Yaoundé, Cameroon, yatat.valaire@gmail.com, Corresponding author, Tel.+(237) 675 30 57 26
b National Advanced School of Engineering University of Yaoundé I, Department of Mathematics and Physics P.O. Box 8390 Yaoundé, Cameroon, tewajules@gmail.com
c Department of Mathematics and Computer Science, Faculty of Science, University of Douala, P.O. Box 24157 Douala, Cameroon, sbowong@gmail.com
d UMI 209 IRD/UPMC UMMISCO, University of Yaoundé I, Faculty of Science, LIRIMA Project team GRIMCAPE, University of Yaoundé I, Faculty of Science P.O. Box 812, Yaoundé, Cameroon
e CETIC project, University of Yaoundé I, Yaoundé, Cameroon

RÉSUMÉ. Dans ce travail, nous discutons de l'incidence que peut avoir la surmortalité due à une maladie infectieuse sur la dynamique d'un modèle Proie-Prédateur de type Leslie-Gower avec maladie chez les Proies. La maladie infectieuse a le formalisme épidémiologique SIS (Susceptible-Infecté-Susceptible). Nous procédons à une analyze qualitative du modèle nous permettant de calculer des seuils écologiques qui résument les résultats de stabilité des différents équilibres. Nous mettons en exergue des conditions pour lesquelles la maladie disparaîtrait de la commuanauté ou deviendrait endémique. Finalement, nous présentons des simulations numériques qui illustrent nos résultats analytiques.

ABSTRACT. In this paper, we discuss the incidence of disease-induced death in a Leslie-Gower Prey-Predator model subjects to an infectious disease affecting only Preys. The infectious disease has the epidemiological SIS (Susceptible-Infectious-Susceptible) formalism. We carry out a qualitative analysis through which we compute ecological thresholds involving biological parameters of Preys, Predators and disease dynamic. We further investigate stability results of model steady states. We further highlight conditions, involving ecological thresholds, under which disease will disappear from the community or will become endemic. Finally, we show some numerical simulations in order to illustrate our analytical results.

MOTS-CLÉS : Modélisation, Maladie infectieuse, Surmortalité due à la maladie, Analyse qualitative

KEYWORDS : Modelling, Infectious disease, disease-induced death, Qualitative analysis

1. Introduction

A Leslie-Gower Predator-Prey model is a two species food chain with the particularity that the carrying capacity of Predator population is proportional to the number of Preys i.e. when there is a few quantity of Preys, predation is negligible so Predators find alternative foods ([10]). Since Predators and Preys that are involved in this model can be subjected to infectious disease, a major issue in mathematical modelling is to understand the effects of infectious diseases in regulating natural populations, decreasing their population sizes or reducing their natural fluctuations ([2], [9], [10], [8]). Many studies have been carried out in order to analyze the influence of infectious disease in Predator-Prey dynamics through mathematical modelling. Generally, there are more macroparasitic infections which can affect only preys, only predators or preys and predators. According to several epidemiological models and studies, infectious disease is able to leads a sur-mortality in the host population ([1], [3]).

Disease-induced death has been identify by a wide of authors as able to lead the so-called 'backward bifurcation' in epidemiological models ([1], [3] and references therein). Recall that in mathematical modelling theory, a backward bifurcation occurs when the disease-free equilibrium and the endemic equilibrium are simultaneously stable when a given threshold takes some values ([1] [3]). In other words, the infectious disease will not die out from the population. From public health policies, backward bifurcation is the worth think that can happen.

Based on that observations, a natural question that concerns the modelling of Predator-Prey dynamics experiencing infectious disease is : what is the incidence of disease-induced death in the outcomes of the model ? Despite the fact that there exist several study on Predator-Prey modelling in presence of infectious disease, this particular question has been scarcely addressed. Therefore, this paper aims to give an answer to that question at least for the particular case of the Leslie-Gower Predator-Prey model that has been widely study in the literature ([10] and references therein). For the authors knowledge, this paper is the first that addresses the question of taking into account or not disease-induced death in eco-epidemiological models.

2. The model formulation

Following ([6], [7]), the Leslie-Gower Predator-Prey model is given by

$$\begin{cases} \dot{H}(t) = (r_1 - a_1 P(t) - b_1 H(t))H(t), \qquad \dot{P}(t) = \left(r_2 - a_2 \dfrac{P(t)}{H(t)}\right) P(t), \\ H(0) > 0, \quad P(0) \geq 0 \end{cases} \qquad [1]$$

where H denotes the Prey population, P the Predator population, r_1 the intrinsic growth rate of the Preys, r_2 is the intrinsic growth rate of the Predators, a_1 is the predation rate per unit of time, $K = \dfrac{r_1}{b_1}$ is the carrying capacity of the Prey's environment and $\dfrac{r_2}{a_2}H$ is the "carrying capacity" of the Predator's environment which is proportional to the number of Prey.

The major objective here is to combine the preceding model (1) and an epidemiological SIS compartmental model, in order to analyze the influence of SIS infectious disease in a Predator-Prey community. The following hypothesis hold true in our model
(H1) The disease transmission follows the mass action law.

(H2) There is a disease-induced death for infectious populations.
(H3) The infected population do not become immune.
(H4) It is assumed that Predator cannot distinguish the infectious and healthy Preys.
(H5) We assume that only susceptible Preys are capable of reproducing.
Note that assumptions (H3)-(H5) was already described in [10]. Recall that irrespective to [10] our model acknowledges a major mechanism of infectious disease dynamic : the disease-induced death of infectious individuals.

3. Mathematical analysis

We start this study by recalling some meaningful results of model (1). The following results hold for system (1).

Theorem 3.1 *1) The nonnegative orthant $\mathbf{R}^2_+$ is positively invariant by system (1).*

2) Let $\varepsilon > 0$, the set $D = \left\{(H,P) : 0 < H \leq K+\varepsilon, 0 \leq P \leq \frac{r_2}{a_2}(K+\varepsilon)\right\}$ is a feasible region for system (1).

3) System (1) don't admit periodic solutions.

4) The predator-free equilibrium $E_1 = \left(\frac{r_1}{b_1}, 0\right) = (K,0)$ is a saddle point with stability for Prey population and instability for Predator population.

5) The coexistence equilibrium $E_2 = (H^*, P^*) = \left(\frac{r_1 a_2}{a_1 r_2 + a_2 b_1}, \frac{r_1 r_2}{a_1 r_2 + b_1 a_2}\right)$ is globally asymptotically stable (GAS).

Proof 3.1 *See Appendix A.*

Now we reach the step of the formulation and the study of the eco-epidemiological Predator-Prey model. For this purpose, let the variables S and I denote respectively the susceptible and infectious in Prey population. We further assume a density-dependent demographic mechanisms (birth and death) for Preys ([2]). Specifically, the parameter $0 \leq \theta \leq 1$ is such that $b - \frac{r_1 \theta H}{K}$ is the birth rate coefficient, $\mu + \frac{(1-\theta) r_1 H}{K}$ is the mortality rate, $r_1 = b - \mu$ is the intrinsic growth rate of Preys. The restricted growth in the logistic equation is due to a density-dependent death rate when $\theta = 0$, is due to a density-dependent birth rate when $\theta = 1$, and is due to a combination of these when $0 < \theta < 1$. σ denotes the recovery rate of infectious Preys. λ is the adequate contact rate between susceptibles and infectious in Prey that leads to disease transmission while d denotes the disease-induced death rate.
Based on these biological premise together with assumptions (H1)-(H6), the Leslie-Gower Predator-Prey model when the disease is present in Preys reads as

$$\left\{\begin{array}{rcl} \dot{H} &=& r_1\left(1-\frac{H}{K}\right)H - a_1 PH - dI, \\ \dot{S} &=& \left(b - r_1\theta\frac{H}{K}\right)H - \left[\mu + \frac{(1-\theta)r_1 H}{K}\right]S - \lambda SI + \sigma I - a_1 SP, \\ \dot{I} &=& \lambda SI - \sigma I - \left[\mu + \frac{(1-\theta)r_1 H}{K}\right]I - a_1 IP - dI, \\ \dot{P} &=& \left(r_2 - \frac{a_2 P}{H}\right)P, \end{array}\right. \quad [2]$$

Using the fact that $H = S + I$, (2) is reduced to

$$\left\{\begin{array}{lcl} \dot{H} & = & r_1\left(1-\dfrac{H}{K}\right)H - a_1PH - dI, \\ \dot{I} & = & \lambda(H-I)I - \sigma I - \left[\mu + \dfrac{(1-\theta)r_1H}{K}\right]I - a_1IP - dI, \\ \dot{P} & = & \left(r_2 - \dfrac{a_2P}{H}\right)P, \\ H(0) & > & 0, \quad I(0) \geq 0, \quad P(0) \geq 0. \end{array}\right. \qquad [3]$$

Using a similar reasoning as in Theorem 3.1, the following results hold for system (3).

Lemma 3.1 *1) The nonnegative orthant $\mathbf{R}^3_+$ is positively invariant by system (3).*

2) Let $\varepsilon > 0$, the set D defined as

$$D = \left\{(H, I, P) : 0 < H \leq K + \varepsilon, 0 \leq I \leq H, 0 \leq P \leq \frac{r_2}{a_2}(K+\varepsilon)\right\}$$

is a feasible region for system (3).

In order to analyze the impact of the disease-induced death rate on the outcomes of model (3), in the sequel, we will distinguish to cases. First, the case where $d = 0$ and second, $d > 0$.

3.1. The eco-epidemiological model without disease-induced death

Here we start, by assuming that the infectious disease does not lead supplement deaths. Therefore we should set $d = 0$ in model (3). Let

$$\mathcal{R}_1 = \frac{\lambda K}{\sigma + \mu + (1-\theta)r_1}, \quad \mathcal{Q}_1 = \frac{\lambda H^*}{\sigma + \mu + (1-\theta)b_1H^* + a_1P^*},$$

where H^* and P^* are given in Theorem 3.1. Setting the right hand side of model (3) equal to zero leads the following result.

Lemma 3.2 *Model (3) admits at most four equilibria :*

1) The point $E_1 = (K, 0, 0)$. That is, both Predators and disease die out.

2) When $\mathcal{R}_1$>1, the point $E_2 = \left(K, K\left(1 - \dfrac{1}{\mathcal{R}_1}\right), 0\right)$ is ecologically meaningful. In other words, Predators die out but disease persists in Preys.

3) The point $E_3 = (H^*, 0, P^*)$. There is a coexistence between Preys and Predators while disease dies out.

4) When $\mathcal{Q}_1 > 1$, the endemic point $E_4 = (H^*, I_e, P^*)$ with $I_e = H^*\left(1 - \dfrac{1}{\mathcal{Q}_1}\right)$ is ecologically meaningful.

Now we turn to investigate asymptotic stability results of equilibria of system (3). We first investigate local stability properties and further characterize their global asymptotic stability properties. To address local stability properties, we will compute jacobian matrix of system (3) at any of its equilibria. Recall that an equilibrium is locally asymptotically stable (LAS) whenever its jacobian matrix has eigenvalues with real part lying in negative real axis.

Theorem 3.2 *The following result holds for system (3).*

1) Both E_1 and E_2 are unstable.

2) If $\mathcal{Q}_1 < 1$ then E_3 is LAS.

3) Assume that the endemic equilibrium E_4 exists, that is, $\mathcal{Q}_1 > 1$ then it is LAS.

Proof 3.2 *See Appendix B.*

Remark 3.1 *At this step, it is not possible to conclude about what are the outcomes of model (3) when the threshold $\mathcal{Q}_1$ take the critical value 1. This issue will be addressed in the next result. Moreover, $\mathcal{Q}_1$ can be seen as the basic reproduction number of Preys when Predators are present while $\mathcal{R}_1$ can be seen as the basic reproduction number of Preys in absence of Predator. Recall that the basic reproduction number is the number of secondary infectious individuals that can be generated by an infectious individual, all over it infectious time, when he is in a population of susceptible individuals.*

We also derive the following result

Theorem 3.3 *1) If $\mathcal{Q}_1 \leq 1$ then E_3 is globally asymptotically stable (GAS).*

2) Assume that the endemic equilibrium E_4 exists, that is, $\mathcal{Q}_1 > 1$ then it is GAS.

Proof 3.3 *See Appendix C.*

At this step, we have characterized, from a qualitative point of view, the outcomes of model (3) when there is no disease-induced death. In the next section, we will carry out a similar study in order to obtain elements to characterize the impact of the disease-induced death in the Leslie-Gower Predator-Prey model with disease in Preys.

3.2. The eco-epidemiological model with disease-induced death

This section is devoted to the study of model (3) with $d > 0$. As the starting point, we computed its equilibria. To achieve that objective, we set the right hand side of system (3) equal to zero. Let $\mathcal{R}_2 = \dfrac{\lambda(db_1a_2 + r_1r_2a_1)}{b_1d(a_1r_2 + (1-\theta)a_2b_1)}$. The following result is valid.

Lemma 3.3 *Model (3) admits at most four equilibria :*

1) The point $e_1 = (K, 0, 0)$. Both Predators and disease die out.

2) Assume $\mathcal{R}_1 > 1$ and let $0 < \overline{H} \leq K$ the positive solution of

$$-b_1\lambda H^2 + H(\lambda(r_1 - d) + db_1(1-\theta)) + d(\sigma + \mu + d) = 0. \quad [4]$$

Let also $\mathcal{Q}_2 = \dfrac{\lambda\overline{H}}{\sigma + \mu + (1-\theta)b_1\overline{H} + d}$. Therefore, if $\mathcal{Q}_2 > 1$ then the point $e_2 = \left(\overline{H}, \overline{H}\left(1 - \dfrac{1}{\mathcal{Q}_2}\right), 0\right)$ is a meaningful equilibrium. In other words, Predators die out but disease persists in Preys.

3) The point $e_3 = (H^*, 0, P^*)$. There is a coexistence between Preys and Predators while disease dies out.

4) Suppose that $\mathcal{R}_2 > 1$ and let $0 < H^\dagger \leq K$ the positive solution of

$$-\lambda\left(b_1 + \frac{a_1r_2}{a_2}\right)H^2 + H\left(\lambda(r_1 - d) + db_1(1-\theta) + d\frac{a_1r_2}{a_2}\right) + d(\sigma + \mu + d) = 0. \quad [5]$$

Let also $P^\dagger = \frac{r_2}{a_2}H^\dagger$ *and* $\mathcal{Q}_3 = \frac{\lambda H^\dagger}{\sigma+\mu+(1-\theta)b_1H^\dagger+d+a_1P^\dagger}$. *Therefore, if* $\mathcal{Q}_3 > 1$ *then the point* $e_4 = \left(H^\dagger, H^\dagger\left(1-\frac{1}{\mathcal{Q}_3}\right), P^\dagger\right)$ *is a meaningful equilibrium. It denotes the endemicity of the disease in Preys coexisting with Predators.*

Remark 3.2 *We stress the fact that in Lemma 3.3, assumptions* $\mathcal{R}_1 > 1$ *and* $\mathcal{R}_2 > 1$ *are necessary and sufficient to have the positive solution of (4) and (5), respectively, in the feasible domain. That is, lower than* K.
At this step, a first observation that can be made while comparing model (3) without and with disease-induced death is the complexity of computations of equilibria in the latter case.

Now we reach the step of characterizing the stability property of various equilibria. As previously (see Theorem 3.2), we will achieve that goal by characterizing the real parts of eigenvalues of the jacobian matrices computed at any of these equilibria. The following results address that issue. Theorem 3.4 is obtained similarly as Theorem 3.2, so we omit the proof.

Theorem 3.4 *The following result holds for system (3).*

1) Both e_1 *and* e_2 *are unstable.*

2) Let $\mathcal{Q}_3^* = \frac{\lambda H^*}{\sigma+\mu+(1-\theta)b_1H^*+d+a_1P^*}$. *If* $\mathcal{Q}_3^* < 1$ *then* e_3 *is LAS.*

The next result addresses the asymptotic stability of the endemic equilibrium.

Theorem 3.5 *Assume that the endemic equilibrium* $e_4 = \left(H^\dagger, H^\dagger\left(1-\frac{1}{\mathcal{Q}_3}\right), P^\dagger\right)$ *exists, that is* $\mathcal{R}_2 > 1$ *and* $\mathcal{Q}_3 > 1$. *Then is LAS.*

Proof 3.4 *See Appendix D.*

Remark 3.3 *From a qualitative point of view, one can conclude that irrespective of epidemiological models ([1], [3]), the Leslie-Gower Predator-Prey model experiencing infectious disease in Preys present similar results without and with disease-induced death. We observe in this study that the disease-induced death only leads more complexity in terms of analytical treatments of the model.*

4. Numerical simulations

In this section, we provide numerical simulations using an implicit nonstandard algorithm (see [10]) to illustrate and validate analytical results obtained in the previous sections. Indeed, as mentioned in [10], standard numerical methods (Euler, Runge Kutta methods, etc.) included in software package such as Scilab and Matlab sometimes present spurious behaviors which are not in adequacy with the continuous system properties that they aim to approximate i.e., lead to negative solutions, exhibit numerical instabilities, or even converge to the wrong equilibrium for certain values of the time discretization or the model parameters ([10]). Moreover, parameter values have been chosen in such a way

that they obey the conditions for stability or bifurcation. For our numerical treatments, we consider parameter values summarized in Table 1.

Tableau 1. *Parameter values for the Leslie-Gower predator-prey models*

Parameter	Value	Reference
r_1	1	Sharma et al. (2015) [8]
r_2	0.2	Sharma et al. (2015) [8]
a_1	0.1	Tewa et al. (2012) [9]
a_2	0.4	Sharma et al. (2015) [8]
b_1	0.01	Assumed
σ	0.1	Assumed
μ	0.2	Assumed
θ	0.8	Tewa et al. (2012) [9]

Figure 1 illustrates the coexistence of Preys and Predators in the disease-free case.

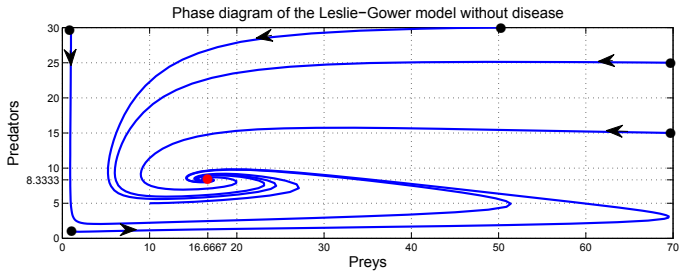


Figure 1. *Predators and Preys coexist in the disease-free case.*

When there is no disease-induced death and as we saw in Theorem 3.3, page 5, the threshold $\mathcal{Q}_1$ captures the whole dynamic of model 3. We illustrate it in figure 2.

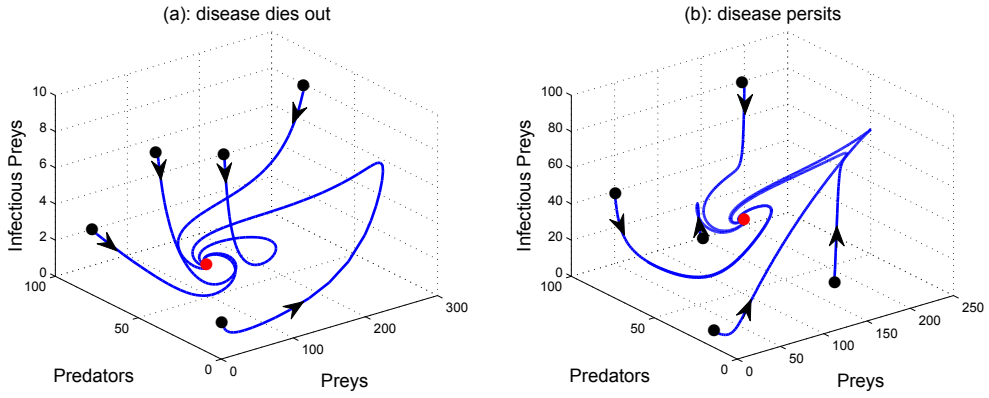


Figure 2. *Disease dies out ($\mathcal{Q}_1$<1) or persists ($\mathcal{Q}_1 > 1$). In panel (a), $\lambda = 0.006$, $d = 0$, in panel (b), $\lambda = 0.2$, $d = 0$. The rest of parameter values in Table 1.*

5. Conclusion

In this paper we carry out the study of a Leslie-Gower Predator-prey model experiencing an infectious disease only in Preys. We distinguished the cases where the model acknowledges or not a disease-induced death. Our qualitative analysis have highlighted several thresholds that summarize the whole dynamics of the model. We further compute conditions, involving afore-mentioned thresholds, under which the infectious disease will disappear or will become endemic in the community. Moreover, we can also conclude that, from a qualitative point of view, disease-induced death has not incidence in the outcomes of the model, irrespective of epidemiological finding ([1], [3]). However, this finding should be improved by the study of several other eco-epidemiological models. At this step, we just have a first indication, a first study and it remains to be validated by several others works. This paper just gives an insight concerning the question of taking into account or not disease-induced death in eco-epidemiological models. We finally illustrate our theoretical results with relevant numerical simulations.

6. Bibliographie

[1] C. CASTILLO-CHAVEZ GAO, B. SONG, « Dynamical models of of turbeculosis and their applications », *Math. Biosc. Eng.*, vol. 1, 2004, pp. 361-404.

[2] L.Q. GAO, H.W. HETHCOTE, « Disease transmission models with density-dependent demographics », *J. Math. Biol.*, vol. 30, 1992, pp. 717-731.

[3] A. HAMADJAM, J.C. KAMGANG, L.N. NKAMBA , D. TIEUDJO, L. EMINI, « Modeling the Dynamics of Arboviral Diseases with Vaccination Perspective », *Biomath*, vol. 4, 2015.

[4] H. HETHCOTE , W. WANG, L. HAN, Z. MA, « A predator-prey model with infected prey« , *Theo. Pop. Biol.* vol. 66, pp. 259-268, 2004.

[5] A. KOROBEINIKOV, « A Lyapunov function for Leslie-Gower predator-prey models », *Appl. Math. Let.*, vol. 14, pp. 697-699, 2001.

[6] P.H. LESLIE, « Some further notes on the use of matrices in population mathematics », *Bioetrika*, vol. 35, pp. 231-245, 1948.

[7] P.H. LESLIE, « A stochastic model for studying the properties of certain biological systems by numerical methods », *Bioetrika*, vol. 45, pp. 16-31, 1958.

[8] S. SHARMA, G.P. SAMANTA, « A Leslie-Gower predator-prey model with disease in prey incorporating a prey refuge », *Chaos, Solitons & Fractals*, vol. 70, 2015, pp. 39-84.

[9] JJ. TEWA, V. YATAT, S. BOWONG, « Predator-prey model with Holling response function of type II and SIS infectious disease », *App. Math. Mod.*, vol. 37, 2012, pp. 4825-4841.

[10] V. YATAT, JJ. TEWA, S. BOWONG, « Dynamic behaviors of a Leslie-Gower Predator-Prey model subject to a SIS infectious disease and Nonstandard Numerical Schemes », *Proceedings CARI*, 2014, pp. 9-17.

A. Proof of Theorem 3.1

From system (1), one has for all $t \geq 0$,

$$\begin{aligned} H(t) &= H(0)\exp\left(\int_0^t (r_1 - a_1P(s) - b_1H(s))\,ds\right) > 0 \\ P(t) &= P(0)\exp\left(\int_0^t \left(r_2 - \frac{a_2P(s)}{H(s)}\right)ds\right) \geq 0. \end{aligned} \quad [6]$$

Therefore, part 1 holds.

To prove part 2 we need to establish that the set D is a positively invariant and absorbing set. Let $([0,T), X = (H,P))$ be the maximal solution of the Cauchy problem (1) with $0 < T \leq +\infty$. Let $t_1 \in [0,T)$. It suffices to show that

– if $H(t_1) \leq K$ then for all $t \in [t_1,T)$, $H(t) \leq K$

– if $P(t_1) \leq \frac{r_2}{a_2}K$ then for all $t \in [t_1,T)$, $P(t) \leq \frac{r_2}{a_2}K$

since we have already shown that solutions are nonnegative. Assume that $\varepsilon_1 > 0$ exists such that $H(t_1+\varepsilon_1) > K$. Let $t_1^* = \inf\{t \geq t_1 | H(t) > K\}$. Since $H(t_1^*) = K$, then $H(t) = K + H'(t_1^*)(t-t_1^*) + o(t-t_1)_{t\to t_1^*}$. Moreover, from the first equation of (1), $H'(t_1^*) = -a_1P(t_1^*)K \leq 0$. Then there exists $\xi > 0$ such that $\forall t_1^* \leq t < t_1^* + \xi$, $H(t) < K$ which is a contradiction. As a result, $\forall t \in [0,T), H(t) \leq K$. Similarly one can prove that if $P(t_1) \leq \frac{r_2}{a_2}K$ then for all $t \in [t_1,T)$, $P(t) \leq \frac{r_2}{a_2}K$.

Now we reach the step that aims to show that the set D is an absorbing set. From the first equation of system (1) one has $\dot{H}(t) \leq r_1\left(1 - \frac{H}{K}\right)H$ which implies that

$$H(t) \leq u(t) \to K \quad \text{as} \quad t \to +\infty,$$

where u is the unique solution of $\dot{u} = r_1\left(1-\frac{u}{K}\right)u$ with $u(0) = H(0)$. Hence for all $\varepsilon > 0$, $\exists T_1 > 0 / H(t) \leq K + \varepsilon$, $\forall t > T_1$. Similarly, from the second equation of system (1) one has $\forall t > T_1$, $\dot{P}(t) \leq r_2\left(1 - \frac{a_2P}{r_2(K+\varepsilon)}\right)P$ which also implies that $P(t) \leq v(t) \to \frac{r_2}{a_2}(K+\varepsilon)$ as $t \to +\infty$, where v is the unique solution of $\dot{v} = r_2\left(1 - \frac{a_2v}{r_2(K+\varepsilon)}\right)v$ with $v(0) = P(0)$. Thus there exists $\exists T_2 > 0 / P(t) \leq \frac{r_2}{a_2}(K + \varepsilon)$. These end the proof of part 2.

To prove part 3, one uses the Dulac function $B(H,P) = \frac{1}{HP}$. Since $-r_1 < 0$ and $r_2 > 0$ are the eigenvalues of the jacobian matrix of system (1) at E_1, it follows that E_1 is a saddle point. Finally, to prove part 5 one can use the Lyapunov function proposed by Korobeinikov (see [5]).

B. Proof of Theorem 3.2

Since $r_2 > 0$ is an eigenvalue of the jacobian matrices of system (3) at E_1 and E_2, it therefore follows that both E_1 and E_2 are unstable.

Since the variable I does not appear in the first and the third equation of system (3) and together we Lemma 3.1 it suffices to compute the eigenvalue of the jacobian matrices

of both E_3 and E_4 in the I-direction. A direct computation leads that the eigenvalue of the jacobian matrix at E_3 is $\eta_{E_3,I} = \lambda H^* \left(1 - \frac{1}{\mathcal{Q}_1}\right)$ while at E_4 it is $\eta_{E_4,I} = -\lambda I_e$. Therefore, E_3 is LAS whenever $\mathcal{Q}_1 < 1$ while E_4 when it exists, i.e. $I_e > 0$, it is LAS. This ends the proof.

C. Proof of Theorem 3.3

Since system (3) is dissipative, that is, its solutions are bounded (see the feasible region D) then one can apply results on triangular systems (see Corollary 4 in [4]). Following Theorem 3.1, we deduce that $\lim_{t\to+\infty}(H,P)(t) = (H^*, P^*)$. Therefore, the limiting equation of variable I is $\dot{I} = \left(\lambda H^* \left(1 - \frac{1}{\mathcal{Q}_1}\right) - \lambda I\right) I$. Finally, it follows that if $\mathcal{Q}_1 \leq 1$ then $I \to 0$ and E_3 is GAS. Similarly, if $\mathcal{Q}_1 > 1$ then $I \to I_e$ and E_4 is GAS. This completes the proof.

D. Proof of Theorem 3.5

Since the endemic equilibrium $e_4 = \left(H^\dagger, H^\dagger \left(1 - \frac{1}{\mathcal{Q}_3}\right), P^\dagger\right)$ exists, that is $\mathcal{R}_2 > 1$ and $\mathcal{Q}_3 > 1$, one has

$$d - \lambda H^\dagger < 0. \qquad [7]$$

For simplicity, in the sequel we note H (resp. I, P) instead of $H^\dagger$ (resp. $I^\dagger$, $P^\dagger$). Moreover, let $A_1 = -b_1 H + d\left(1 - \frac{1}{\mathcal{Q}_3}\right)$; $A_2 = -d$; $A_3 = -a_1 H$;$A_4 = (\lambda - (1-\theta)b_1)I$; $A_5 = -\lambda I$; $A_6 = -a_1 I$; $A_7 = \frac{r_2^2}{a_2}$; $A_8 = -r_2$. $C_0 = A_1A_5A_8 + A_7A_2A_6 - A_7A_3A_5 - A_4A_2A_8$; $C_1 = -A_1A_5 + A_4A_2 - A_1A_8 - A_5A_8 + A_7A_3$; $C_2 = A_1 + A_5 + A_8$. Following Routh-Hurvitz theorem, the endemic equilibrium e_4 is LAS whenever $C_0 < 0$ and $C_2 < 0$ and $C_1C_2 + C_0 > 0$. Straightforward computations lead $C_2 = -b_H + (d - \lambda H)\left(1 - \frac{1}{\mathcal{Q}_3}\right) < 0$; $C_0 = -r_2\left(\lambda b_1 IH + \frac{dI(\sigma+\mu+d+a_1P)}{H}\right) + \frac{r_2^2}{a_2}a_1 I(d - \lambda H) < 0$;

$$\begin{aligned} C_1C_2 + C_0 = \quad & \left(-b_1H + (d-\lambda H)\left(1 - \tfrac{1}{\mathcal{Q}_3}\right)\right)\left(-d(1-\theta)b_1I - \lambda b_1IH - \lambda d\tfrac{I}{\mathcal{Q}_3}\right. \\ & -r_2b_1H + r_2(d-\lambda H)\left(1-\tfrac{1}{\mathcal{Q}_3}\right)\Big) + \tfrac{r_2^2}{a_2}(a_1b_1H^2) \\ & -r_2\left(-a_1H\tfrac{r_2^2}{a_2} + r_2\left(-b_1H + (d-\lambda H)\left(1-\tfrac{1}{\mathcal{Q}_3}\right)\right)\right) \\ & > 0. \end{aligned} \qquad [8]$$

Thus, when the endemic equilibrium, e_4, exists, it is LAS.

Identification of Robin coefficient for Stokes Problem

A. Ben Abda * — F. Khayat **

* LAMSIN-ENIT
BP 37, 1002 Tunis le Belvédère
Tunisia
amel.benabda@enit.rnu.tn

** LAMSIN-ENIT
BP 37, 1002 Tunis le Belvédère
Tunisia
faten.khayat@gmail.com

RÉSUMÉ. Dans ce travail, on s'intéresse à l'identification d'un coefficient de Robin sur une partie non accessible du bord d'un domaine à partir de données **faiblement surdéterminées** sur la partie accessible. Le modèle est régi par les équations de Stokes. Dans un premier temps, nous utilisons une méthode du type décomposition de domaine pour calculer les composantes inconnues de la vitesse et du tenseur des contraintes, puis nous utilisons ces données pour calculer le coefficient recherché. Nous donnons des tests numériques pour valider la méthode utilisée.

ABSTRACT. In this paper, we deal with the inverse problem of identifying a Robin coefficient on some inaccessible part of a boundary of a domain from the knowledge of **partially overdetermined** data on the accessible part. The underlying PDE's system is the Stokes one. We use a domain decomposition-like method to first recover lacking velocity and stress tensor component. Numerical trials highlights the efficiency of the proposed method.

MOTS-CLÉS : Coefficient de Robin, Conditions aux limites défectueuses, Contrainte de cisaillement, Equations de Stokes, Problème inverse

KEYWORDS : Robin coefficient, Defective boundary condition, Shear stress, Stokes equations, Inverse problem

1. Introduction

Consider an incompressible and homogeneous fluid flow governed by Stokes equations into an open bounded and connected domain $\Omega \subset \mathbb{R}^2$. The boundary $\Gamma = \partial\Omega$ is composed of two parts Γ_c and Γ_i having non-vanishing measure and such that $\Gamma_c \cap \Gamma_i$ is empty. Γ_c is the accessible part, Γ_i is the non accessible one. We formulate our problem as follows :

$$(\mathbb{P}) \begin{cases} -\nu\Delta u + \nabla p &= 0 & \text{in} & \Omega \\ \nabla \cdot u &= 0 & \text{in} & \Omega \\ (\sigma(u)\cdot n)\cdot \tau &= g_c & \text{on} & \Gamma_c \\ u \cdot n &= \Phi_c \cdot n & \text{on} & \Gamma_c \\ \sigma(u)\cdot n + R\, u &= 0 & \text{on} & \Gamma_i \end{cases} \tag{1}$$

ν is the viscosity of the fluid that we will assume equal to 1, σ denotes the stress tensor $\sigma(u) = \sigma(u,p) = 2\nu D(u) - pI$, where $D(u)$ is the strain tensor defined by : $D(u) = \frac{1}{2}(\nabla u + \nabla u^T)$. n is the outward normal on $\partial\Omega$ and τ is the tangential vector of $\partial\Omega$. R is the Robin coefficient assumed hereafter to be a positive number.
We want to determine the coefficient R from the knowledge of $u.\tau$ on Γ_c.
The method followed here to recover R lies on the recovery of the velocity and the normal stress on the non accessible part Γ_i.
Notice that the boundary condition on the Γ_c is not the Neumann condition regarding the Stokes operator. Thus, this is a non-trivial situation since on the accessible boundary the information on the normal component of the normal stress is unavailable, and only partially overspecified data are given. Nonetheless, this condition is natural, one may refer to [1, 2], for instance, for the description and the background on this boundary condition.
The Cauchy problem is known since Hadamard to be ill posed in the sense that if a solution exists, it does not depend continuously on the data (Φ_c, g_c). Thus, the lack of complete data on the accessible boundary Γ_c may increase the degree of the ill-posedness, and numerically worst behavior is expected.
Our work is motivated first by the study of airway resistance in pneumology which characterizes the patient's ventilation capability and second by the study of the resistivity of a stent which is a medical device used to prevent rupture of aneurysms [3, 4].
The problem of identifying Robin coefficient has been studied by Chaabane and Jaoua [5] for Laplace equations and by Boulakia, Egloffe and Grandmont [6] for Stokes problem where they consider the full overdetermined problem namely the velocity and the hole stress tensor on Γ_c.
In our case the difficulty is increased as long as the overdetermined data are incomplete. Contrary to the case considered in [6], there is no unique continuation results helping us to prove identifiability results. Neverthless, the authors have studied in [7] the problem of recovering the velocity and the stress tensor on the inaccessible part of the boundary from these incomplete data on the accessible part and made a full study which will be of great help for the present work.

2. Recovering lacking data

Giving a compatible data $(\Phi_c, g_c) \in (H^{\frac{1}{2}}(\Gamma_c))^2 \times H^{-\frac{1}{2}}(\Gamma_c)$, that is a data for which a solution (u, p) exists for the problem :

$$(\mathbb{PI}) \begin{cases} -\nu\Delta u + \nabla p = 0 & \text{in} \quad \Omega \\ \nabla \cdot u = 0 & \text{in} \quad \Omega \\ u = \Phi_c & \text{on} \quad \Gamma_c \\ (\sigma(u) \cdot n) \cdot \tau = g_c & \text{on} \quad \Gamma_c \end{cases} \tag{2}$$

we want to determine the velocity Φ_i together with $G_i = \sigma(u_i) \cdot n$ on the non accessible part Γ_i.
Assume that Φ_i and G_i are recovered, we will have therefore the following partially over-determined boundary conditions system :

$$\begin{cases} -\nu\Delta u + \nabla p = 0 & \text{in} \quad \Omega \\ \nabla \cdot u = 0 & \text{in} \quad \Omega \\ u = \Phi_c, \ (\sigma(u) \cdot n) \cdot \tau = g_c & \text{on} \quad \Gamma_c \\ u = \Phi_i, \ \sigma(u) \cdot n = G_i & \text{on} \quad \Gamma_i \end{cases} \tag{3}$$

In order to solve this problem, we will use a (fictious) domain decompostion-like method [8, 9] which consists on splitting the problem (3) into two direct and well-posed problems using only one data on Γ_c.
Thus, let $(u_D^\lambda, p_D^\lambda)$ and $(u_N^\lambda, p_N^\lambda)$ be respectively the solution of the following Dirichlet and Neumann problems :

$$(\mathbb{P}_D) \begin{cases} -\nu\Delta u_D^\lambda + \nabla p_D^\lambda = 0 & \text{in} \quad \Omega \\ \nabla \cdot u_D^\lambda = 0 & \text{in} \quad \Omega \\ u_D^\lambda = \Phi_c & \text{on} \quad \Gamma_c \\ u_D^\lambda = \lambda & \text{on} \quad \Gamma_i \end{cases} \quad (\mathbb{P}_N) \begin{cases} -\nu\Delta u_N^\lambda + \nabla p_N^\lambda = 0 & \text{in} \quad \Omega \\ \nabla \cdot u_N^\lambda = 0 & \text{in} \quad \Omega \\ (\sigma(u_N^\lambda) \cdot n) \cdot \tau = g_c & \text{on} \quad \Gamma_c \\ u_N^\lambda \cdot n = \Phi_c \cdot n & \text{on} \quad \Gamma_c \\ u_N^\lambda = \lambda & \text{on} \quad \Gamma_i \end{cases}$$

A solution of the problem (2) is recovered if and only if the solutions of the well-posed above problems coïncide. The proposed data-recovering problem therefore amounts to minimizing the gap between u_D^λ and u_N^λ.
Following the study done in [10, 11], we define the cost function E which could be interpreted as an energy-type error functional. E is defined as follows :

$$E(\lambda) = \frac{1}{2} \int_\Omega \sigma(u_D^\lambda - u_N^\lambda) : \nabla(u_D^\lambda - u_N^\lambda) \tag{4}$$

We have proved in [7] the following proposition :

Proposition 1
1. E is a positive quadratic and convex functional on $(H^{\frac{1}{2}}(\Gamma_i))^2$.
2. For a compatible pair (Φ_c, g_c), the solution (Φ_i, G_i) of the partially overdetermined boundary value problem (2) is obtained by the following

$$\Phi_i = u_D^{\lambda_{min}}|_{\Gamma_i},\ G_i = (\sigma(u_N^{\lambda_{min}}) \cdot n)|_{\Gamma_i}$$

where λ_{min} is the solution of the following minimization problem :

$$\lambda_{min} = \arg \min_{\lambda \in (H^{\frac{1}{2}}(\Gamma_i))^2} E(\lambda) \tag{5}$$

2.1. Minimization procedure

We next prove the following result :

Proposition 2
For a compatible pair (Φ_c, g_c), the minimum of E is reached when :

$$\sigma(u_D^\lambda) \cdot n \quad = \sigma(u_N^\lambda) \cdot n \quad \text{on } \Gamma_i \tag{6}$$

Proof :
We derive the first optimality condition. It's easy to prove that for $h \in (H^{\frac{1}{2}}(\Gamma_i))^2$, we have :

$$\frac{\partial E}{\partial \lambda}(h) = \frac{1}{2}\int_\Omega \sigma(u_D^\lambda - u_N^\lambda) : \nabla(r_D^h - r_N^h)$$

where (r_D^h, s_D^h) and (r_N^h, s_N^h) are respectively the solutions of :

$$\begin{cases} -\nu\Delta r_D^h + \nabla s_D^h = 0 & \text{in} \quad \Omega \\ \nabla \cdot r_D^h = 0 & \text{in} \quad \Omega \\ r_D^h = 0 & \text{on} \quad \Gamma_c \\ r_D^h = h & \text{on} \quad \Gamma_i \end{cases}, \qquad \begin{cases} -\nu\Delta r_N^h + \nabla s_N^h = 0 & \text{in} \quad \Omega \\ \nabla \cdot r_N^h = 0 & \text{in} \quad \Omega \\ (\sigma(r_N^h) \cdot n) \cdot \tau = 0 & \text{on} \quad \Gamma_c \\ r_N^h \cdot n = 0 & \text{on} \quad \Gamma_c \\ r_N^h = h & \text{on} \quad \Gamma_i \end{cases} \tag{7}$$

Green Formula gives :

$$\frac{\partial E}{\partial \lambda}(h) = \frac{1}{2}\int_{\partial\Omega} \left(\sigma(u_D^\lambda - u_N^\lambda) \cdot n\right) r_D^h - \frac{1}{2}\int_{\partial\Omega} \left(\sigma(r_N^h) \cdot n\right)(u_D^\lambda - u_N^\lambda)$$

since we have $r_D^h = 0$ on Γ_c and $u_D^\lambda - u_N^\lambda = 0$ on Γ_i, then :

$$\frac{\partial E}{\partial \lambda}(h) = \frac{1}{2}\int_{\Gamma_i} \left(\sigma(u_D^\lambda - u_N^\lambda) \cdot n\right) r_D^h - \frac{1}{2}\int_{\Gamma_c} \left(\sigma(r_N^h) \cdot n\right)(u_D^\lambda - u_N^\lambda)$$

using the boundary condition on $(u_D^\lambda - u_N^\lambda)\cdot n$ and on $(\sigma(r_N^h)\cdot n)\cdot\tau$, we conclude that :

$$\frac{\partial E}{\partial \lambda}(h) = \frac{1}{2}\int_{\Gamma_i} \left(\sigma(u_D^\lambda - u_N^\lambda)\cdot n\right) h, \ \ \forall h \in (H^{\frac{1}{2}}(\Gamma_i))^2.$$

thus our statement follows immediately.

2.2. The interfacial operators

Following the classical framework of the Domain Decomposition Community, we introduce the notations :

$$\left\{ \begin{array}{ll} (u_D^\lambda, p_D^\lambda) & = (u_D^0, p_D^0) + (r_D^\lambda, s_D^\lambda) \\ (u_N^\lambda, p_N^\lambda) & = (u_N^0, p_N^0) + (r_N^\lambda, s_N^\lambda) \end{array} \right.$$

thus, the condition (6) can be written as :

$$\sigma(r_D^\lambda)\cdot n - \sigma(r_N^\lambda)\cdot n = -[\sigma(u_D^0)\cdot n - \sigma(u_N^0)\cdot n]$$

or equivalently by using operator's modelling

$$S(\lambda) = T$$

with

$$T = -[\sigma(u_D^0)\cdot n - \sigma(u_N^0)\cdot n]$$

and $S = S_D - S_N$ is the Steklov-Poincaré operator defined by :

$$S(\lambda) = S_D(\lambda) - S_N(\lambda)$$

and where

$$\begin{array}{lccc} S_D : & H^{1/2}(\Gamma_i)^2 & \to & H^{-1/2}(\Gamma_i)^2 \\ & \lambda & \to & \sigma(r_D^\lambda)\cdot n \end{array} , \qquad \begin{array}{lccc} S_N : & H^{1/2}(\Gamma_i)^2 & \to & H^{-1/2}(\Gamma_i)^2 \\ & \lambda & \to & \sigma(r_N^\lambda)\cdot n \end{array} \tag{8}$$

2.3. Reconstruction of Robin coefficient

From the last equation in (1), we can now determine the value of the real parameter R using the means of the recovered values of u and $\sigma(u).n$ on Γ_i. More precisely, we use the formula :

$$|R| = \left| \frac{\int_{\Gamma_i}[\sigma(u_N).n]_1 + \int_{\Gamma_i}[\sigma(u_N).n]_2}{\int_{\Gamma_i}[u_N]_1 + \int_{\Gamma_i}[u_N]_2} \right| \tag{9}$$

where for a vector u of $\mathbb{R}^2$, $[u]_k$ denotes the k^{th} component of u.
We have not deal in the present work with the case of a spatially dependent R which will be treated later on.

3. Numerical Results

We use a numerical procedure based on the preconditioned gradient algorithm :

$$X_{k+1} = X_k - mP[S(X_k) - T]$$

where P is a preconditioning operator and m is a relaxation parameter. The expressions of S and T are described in the previous section.

3.1. Algorithm

1) Initialization : For $k = 0$ choose $\lambda_0 = 0$

2) Solve $(\mathbb{P}_D)$ and $(\mathbb{P}_N)$ whith $\lambda = \lambda_k$.

3) Compute w_k solution of the following "interface" problem :

$$(\mathbb{P}_I) \begin{cases} -\nu\Delta w_k + \nabla p_k = 0 & \text{in } \Omega \\ \nabla \cdot w_k = 0 & \text{in } \Omega \\ w_k = 0 & \text{on } \Gamma_c \\ \sigma(w_k)\cdot n = \left(\sigma(u_D^k)\cdot n - \sigma(u_N^k)\cdot n\right) & \text{on } \Gamma_i \end{cases} \tag{10}$$

4) Update λ :

$$\lambda_{k+1} = \lambda_k + m\, w_k$$

5) Stopping Criteria : $E(\lambda_k) < \varepsilon$, where ε is the tolerance (selected numerically).

6) Calculate R using formula (9)

3.2. Results and Discussions

We will test our method for two cases corresponding to different choices of the domain Ω. The first choice corresponds to an annular domain and the second to a rectangular one. The overdetermined data are generated from the following test examples given by [12, 9] and refered to by smooth and singular data respectively :

$$u(x,y) = (4y^3 - x^2, 4x^3 + 2xy - 1),\ p(x,y) = 24xy - 2x$$

$$\begin{cases} u(x,y) = \frac{1}{4\pi}\left(log\frac{1}{\sqrt{(x-a)^2+y^2}} + \frac{(x-a)^2}{(x-a)^2+y^2}, \frac{y(x-a)}{(x-a)^2+y^2}\right), \\ p(x,y) = \frac{1}{2\pi}\frac{x-a}{(x-a)^2+y^2}. \end{cases}$$

For each case and for different test values of R, we will compare the components of the velocity and those of the normal stress tensor for the analytical solution u_{exact}, u_D and u_N

on Γ_i. Then we will reconstruct on Γ_c the unkown values $(\sigma(u_D)\cdot n)\cdot n$, $(\sigma(u_N)\cdot n)\cdot n$ and compare them with $(\sigma(u_{exact})\cdot n)\cdot n$.
Moreover, we will compare on Γ_i the normal stress of u_D and u_N with the limit condition $R\,u_{exact}$.
Finally, we will reconstruct the value of the Robin coefficient that we will call ρ and compare it with the exact used value R.
Computations are done under Freefem++ Software environment.

First example : Let Ω be the annular domain with radius $R_1 = 1$ and $R_2 = 2$. Γ_c will be the outer circle and Γ_i the inner one. we mesh with 150 nodes on Γ_c and 100 nodes on Γ_i. $\varepsilon = 6\times 10^{-4}$ (80 iterations were required).
The reconstructed stress tensor on Γ_i from u_D and u_N are compared with the one from the exact solution (figure 1). We give the result for $R = 20$ but the numerical tests are done for several values of R and the results are satisfying.
In table 1 where we compare the exact value of the Robin coefficient R with the identified one by our method ρ, we note that the error rate is interesting it varies between 0.5% and 8.9%.

Second example : In this case, Ω is a rectangular domain with $L = 2$ and $\ell = 1$. $\partial\Omega = \Gamma_c \cup \Gamma_i \cup \Gamma_N$, where $\Gamma_c = [0,2]\times\{1\}$, $\Gamma_i = [0,2]\times\{0\}$, $\Gamma_N = (\{0\}\times[0,1])\cup(\{2\}\times[0,1])$. We mesh with 60 nodes on Γ_c and Γ_i, and with 50 nodes on Γ_N. $\varepsilon = 3\times 10^{-3}$ (50 iterations were required).

In figure 2 we plot the lacking component of the normal stress on Γ_c (left) and compare the normal stress with Ru_{exact} on Γ_i (right). Note that these reconstructed fields are in close agreement with the exact ones. We test for several values of R.

In table 2 we reconstruct the value of the Robin coefficient ρ and compare it with the exact one R. The error rate is varing between 1.2% and 7%.

In order to test the robustness of the used method, we introduce a white noise perturbation to the data with an amplitude ranging from 1 to 15%. We reconstruct the velocity and the stress tensor on Γ_i from these noisy data. We observe in figure3 that the method used is more robust with smooth data (left) than with singular one (right).

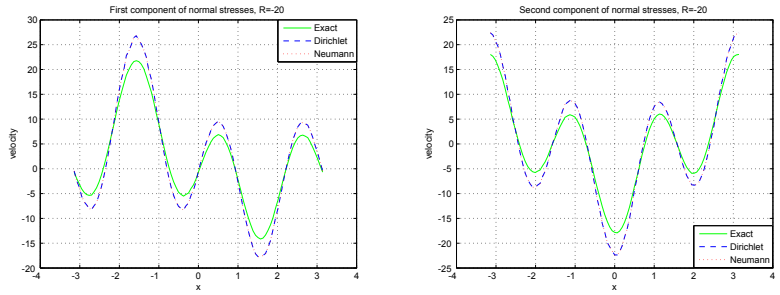


Figure 1. *First example with smooth data, R=20 : the reconstructed stress tensor on Γ_i*

Tableau 1. *First example : Comparaison of ρ and R*

R	5	10	50	70	100
ρ	5.07301	9.94297	45.5175	67.1686	93.8794

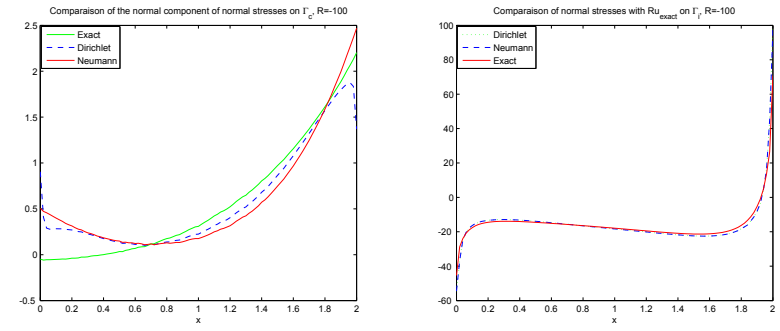


Figure 2. *Second example with smooth data, R=100 : the reconstructed data on Γ_c(left) and comparing normal stress with Ru_{exact} on Γ_i (right)*

Tableau 2. *Comparaison of ρ and R : Rectangular domain*

R	2	5	10	20	50	100
ρ	2.05149	4.93797	9.63617	18.8812	46.4296	92.9558

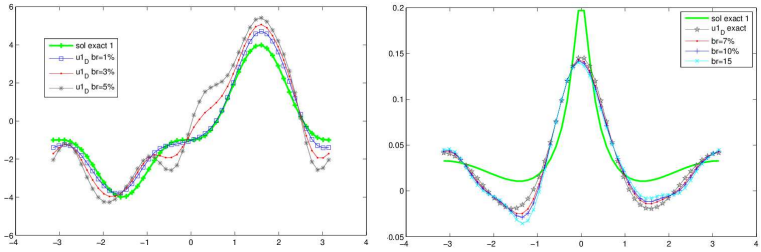


Figure 3. *Comparaison of velocity's first component for noisy data : Smooth data(left), Singular data with a=0.8 (right)*

4. Bibliographie

[1] R. VERFÜRTH, « Finite element approximation on incompressible Navier-Stokes equations with slip boundary condition », *Numerische Mathematik*, n° 50, 1986.

[2] J. M. URQUIZA, A. GARON, M. I. FARINAS, « Weak imposition of the slip boundary condition on curved boundaries for Stokes flow », *Journal of Computational Physics*, n° 256, 2014.

[3] B. MAURY, « The resistance of the respiratory system, from top to bottom », *ESAIM : Proceedings and surveys*, n° 47, 2014.

[4] M.A. FERNANDEZ, J.F. GERBEAU, V. MARTIN, « Numerical simulation of blood flows through a porous interface », *ESAIM : Mathematical Modelling and Numerical Analysis*, n° 42, 2008.

[5] S. CHAABANE, M. JAOUA, « Identification of Robin coefficients by the means of boundary measurements », *Inverse Problems*, n° 15, 1999.

[6] M. BOULAKIA, A.C. EGLOFFE, C. GRANDMONT, « Stability estimates for a robin coefficient in the two- dimensional stokes problem », *Mathematical control and related field*, vol. 3, n° 1, 2013.

[7] A. B. ABDA, F. KHAYAT, « Reconstruction of missing boundary conditions from partially overspecified data : the Stokes system », *Submitted*.

[8] T. MATHEW, « Domain Decomposition Methods for the Numerical Solution of Partial Differentiel Equations », *Springer-Verlag Berlin Heidelberg*, n° 764, 2008.

[9] A. B. ABDA, I. B. SAAD, M. HASSINE, « Recovering boundary data : The Cauchy Stokes system », *Applied Mathematical Modelling*, n° 37, 2013.

[10] S. ANDRIEUX, T. BARANGER, A. B. ABDA, « Solving Cauchy problems by minimizing an energy-like functional », *Inverse problems*, n° 22, 2006 .

[11] F. B. BELGACEM, H. E. FEKIH, « On Cauchy's problem : I. A variational Steklov-Poincaré theory », *Inverse Problems*, vol. 21, 2005.

[12] G. BASTAY, T. JOHANSSON, V. A. KOZLOV, D. LESNIC, « An alternating method for the stationary Stokes system », *Z. Angew. Math. Mech.*, vol. 86, n° 4, 2006.

Rubrique

SCHISTOSOMIA INFECTION

A mathematical analysis of a model with mating structure

Diaby Mouhamadou*, Iggidr Abderrahman **

* LANI, UFR SAT, Université Gaston Berger de Saint-Louis
234, Saint-Louis, Sénégal
diabloss84@yahoo.fr

** INRIA-Nancy Grand-Est & IECL, Université de Lorraine
Metz, France
abderrahman.iggidr@inria.fr

ABSTRACT. Drug treatment, snail control, cercariae control, improved sanitation and health education are the effective strategies which are used to control the schistosomiasis. In this paper, we consider a deterministic model for schistosomiasis transmission dynamics in order to explore the role of the several control strategies. The global stability of a schistosomiasis infection model that involves mating structure including male schistosomes, female schistosomes, paired schistosomes and snails is studied by constructing appropriate Lyapunov functions. We derive the basic reproduction number $\mathcal{R}_0$ for the deterministic model, and establish that the global dynamics are completely determined by the values of $\mathcal{R}_0$. We show that the disease can be eradicated when $\mathcal{R}_0 \leq 1$; otherwise, the system is persistent when $\mathcal{R}_0 > 1$.

RÉSUMÉ. Le traitement médicamenteux, le traitement par les molluscicides, l'amélioration de l'assainissement et l'éducation sanitaire sont les stratégies efficaces qui sont utilisés pour contrôler la schistosomiase. Dans cet article, nous considérons un modèle déterministe pour la dynamique de transmission de la schistosomiase dans le but d'explorer le rôle des différentes stratégies de contrôle. La stabilité globale d'un modèle d'infection de la schistosomiase qui incorpore une structure d'appariement et une dynamique des schistosomes mâles, femelles, pairs et des escargots est étudiée par la construction de fonctions de Lyapunov appropriées. Nous calculons le taux de reproduction de base $\mathcal{R}_0$ pour le modèle déterministe, et établissons que la dynamique globale est complètement déterminée par les valeurs de $\mathcal{R}_0$. Nous montrons que la maladie peut être éradiquée quand $\mathcal{R}_0 \leq 1$; par ailleurs, le système est persistant lorsque $\mathcal{R}_0 > 1$.

KEYWORDS : Epidemic models; Nonlinear dynamical systems; Global stability; Reproduction number; Schistosomiasis.

MOTS-CLÉS : Modèles épidémiologiques; systèmes dynamiques nonlinèaires; Stabilité globale; Taux de reproduction de base; Bilharziose.

1. Introduction

Schistosomiasis (also known as bilharzia, bilharziasis or snail fever) is a vector-borne disease caused by infection of the intestinal or urinary venous system by trematode worms of the genus Schistosoma. More than 207 million people are infected worldwide, with an estimated 700 million people at risk in 74 endemic countries [12]. Schistosomiasis is prevalent in tropical and subtropical areas, especially in poor communities without access to safe drinking water and adequate sanitation. Of the 207 million people with schistosomiasis, 85% live in Africa [12]. Of the tropical diseases, only malaria accounts for a greater global burden than schistosomiasis [11]. Therefore, it is vital to prevent and control the schistosomiasis transmission.

Schistosoma requires the use of two hosts to complete its life cycle: the definitive hosts and the intermediate snail hosts. In definitive hosts, schistosoma has two distinct sexes. Mature male and female worms pair and migrate either to the intestines or the bladder where eggs production occurs. One female worm may lay an average of 200 to $2,000$ eggs per day for up to twenty years. Most eggs leave the blood stream and body through the intestines. Some of the eggs are not excreted, however, and can lodge in the tissues. It is the presence of these eggs, rather than the worms themselves, that causes the disease. These eggs pass in urine or feces into fresh water into miracidia which infect the intermediate snail hosts. In snail hosts, parasites undergo further asexual reproduction, ultimately yielding large numbers of the second free-living stage, the cercaria. Free-swimming cercariae leave the snail host and move through the aquatic or marine environment, often using a whip-like tail, though a tremendous diversity of tail morphology is seen. Cercariae are infective to the second host and turn it into single schistosoma, and infection may occur passively (e.g., a fish consumes a cercaria) or actively (the cercaria penetrates the fish) and terminates the life cycle of the parasite.

Many effective strategies are used in the real world, such as: based on preventive treatment, snail control, cercariae control, improved sanitation and health education. The WHO strategy for schistosomiasis control focuses on reducing disease through periodic, targeted treatment with praziquantel. This involves regular treatment of all people in at-risk groups [12]. Over the past few decades, different mathematical models [3], [5], [13], [10] have been constructed to describe the transmission dynamics involving two-sex problems. In [3], [5], [13], a mathematical model is developed for a schistosomiasis infection that involves pair-formation models and studied the existence, uniqueness and the stabilities of exponential solutions. We note that in [5], [13] authors formulate three forms of pair-formation functions (also known as mating functions) that are the harmonic mean function, the geometric mean function and the minimum function. In [16], Xu et al. have proposed a multi-strain schistosome model with mating structure. Their goal was to study the effect of drug treatment on the maintenance of schistosome genetic diversity. However, in their model they only consider the adult parasite populations. Castillo-Chavez et al. [3] have considered a time delay model but also do not include the snails dynamics. But it is important to take into account the snail dynamics as it is shown in the life cycle of schistosoma. In fact, the parasite offspring is produced directly by infected snails but not by paired parasites as is related in [10].

Recently, Qi et al. [10] have formulated a deterministic mathematical model to study the transmission dynamics of schistosomiasis with a linear mating function incorporating these snail dynamics. This paper gave the expression of a threshold number (and not the basic reproduction number) with a local stability analysis of the disease free equilibrium.

However, no work has been done to investigate the global stability of the equilibria which is more in interest. Here, we take this deterministic schistosomiasis model with mating structure [10] and we propose a complete mathematical analysis. A stability analysis is provided to study the epidemiological consequences of control strategies. We compute the basic reproduction number and we show that when it is less or equal to one then the disease free equilibrium (DFE) is the unique equilibrium of the system and it is globally asymptotically stable, while when the basic reproduction number is greater than one we show that the disease persists. This paper is organized as follows. Model formulation is carried out and the basic properties are shown in the next section. In Section 3, we determine the basic reproductive number $\mathcal{R}_0$ of the model and also establish global stability of the disease-free equilibrium. In the end of this section we show that the disease is uniformly persistent when $\mathcal{R}_0 > 1$. A general conclusion is given in the last section.

2. Mathematical Model

The model that we consider has been presented in [10]. It describes the time evolution of a population divided in three parasites sub-populations and two intermediate snail host sub-populations. The state variables of the model are:

– $X_m(t)$ the male schistosoma population size.

– $X_f(t)$ the female schistosoma population size.

– $X_p(t)$ the pair schistosoma population size.

– $X_s(t)$ the susceptible (uninfected) snail host population size.

– $X_i(t)$ the infected snail host population size.

The time evolution of the different populations is governed by the following system of equations:

$$\begin{cases} \frac{dX_m}{dt} = & k_m\, X_i - (\mu_m + \epsilon_m)\, X_m - \rho\, X_f, \\ \frac{dX_f}{dt} = & k_f\, X_i - (\mu_f + \epsilon_f)\, X_f - \rho\, X_f, \\ \frac{dX_p}{dt} = & \rho\, X_f - (\mu_p + \epsilon_p)\, X_p, \\ \frac{dX_s}{dt} = & \Lambda - (\mu_s + \epsilon_s)\, X_s - \beta\, X_p\, X_s, \\ \frac{dX_i}{dt} = & \beta\, X_p\, X_s - (\mu_s + \epsilon_s + \alpha_s)\, X_i. \end{cases} \tag{1}$$

The different parameters are:

– k_m and k_f are the recruitment rates of male schistosoma and female schistosoma respectively. α_s is the disease-induced death rate of snail hosts.

– μ_m, μ_f, μ_p, and μ_s denote the natural death rate for male, female, pair and snail hosts respectively.

– ρ represents the effective mating rate.

– Λ is the recruitment rate of snail hosts.

– β is the transmission rate from pairs parasite to susceptible snails.

– ϵ_m, ϵ_f, ϵ_p and ϵ_s are the elimination rates of male shistosoma, female schistosoma, paired schistosoma and snails respectively. These elimination rates represent the control strategies.

As it has been done in [10], we shall denote

$$\mu_m + \epsilon_m = \mu_{m\epsilon}, \qquad \mu_f + \epsilon_f = \mu_{f\epsilon},$$
$$\mu_p + \epsilon_p = \mu_{p\epsilon}, \qquad \mu_s + \epsilon_s = \mu_{s\epsilon}.$$

2.1. Basic Properties

In this section, we give some basic results concerning solutions of system (1) that will be subsequently used in the proofs of the stability results.

Proposition 2.1. *The set* $\Gamma = \{M_{sc} \geq F_{sc} \geq 0,\ P_{sc} \geq 0,\ S_{sn} \geq 0,\ I_{sn} \geq 0\}$ *is a positively invariant set for system (1).*

Proof. The vector field given by the right-hand side of system (1) points inward on the boundary of $\mathbb{R}^5_+$. For example, if $X_s = 0$, then, $\dot{X}_s = \Lambda > 0$. In an analogous manner, the same can be shown for the other system components. □

Proposition 2.2. *All solutions of system (1) are forward bounded.*

Proof. Let us define $N_X = X_m + X_f + X_p$ and $N_Y = X_s + X_i$. Using system (1), we have $\dfrac{dN_Y}{dt} = \Lambda - \mu_{s\epsilon}\, N_Y - \alpha_s\, X_i \leq \Lambda - \mu_{s\epsilon}\, N_Y$. This implies that the set $\{N_Y \leq \dfrac{\Lambda}{\mu_{s\epsilon}}\}$ is positively invariant and attracts all the solutions of (1).

We also have:

$$\begin{aligned}\frac{dN_X}{dt} &= (k_m + k_f)\, X_i - \mu_{m\epsilon}\, X_m - (\mu_{f\epsilon} + \rho)\, X_f - \mu_{p\epsilon}\, X_p \\ &\leq (k_m + k_f)\, \frac{\Lambda}{\mu_{s\epsilon}} - \min\{\mu_{m\epsilon}, \mu_{f\epsilon}, \mu_{p\epsilon}\}\, N_X - \rho\, X_f.\end{aligned}$$

Hence, the set $\left\{N_X \leq \dfrac{(k_m + k_f)\Lambda}{\mu_{s\epsilon}\,\gamma}\right\}$, where $\gamma = \min\{\mu_{m\epsilon}, \mu_{f\epsilon}, \mu_{p\epsilon}\}$, is positively invariant set and attracts all the solutions of (1). □

Therefore all feasible solutions of system (1) enter the region

$$\Omega = \left\{(X_m, X_f, X_p, X_s, X_i) \in \mathbb{R}^5_+ : X_s + X_i \leq \frac{\Lambda}{\mu_{s\epsilon}}, \right.$$
$$\left. X_m + X_f + X_p \leq \frac{(k_m + k_f)\Lambda}{\mu_{s\epsilon}\,\gamma}\right\},$$

and the set Ω is a compact positively invariant set for system (1). It is then sufficient to consider solutions in Ω.

3. The basic reproduction number and the disease-free Equilibrium

The disease-free equilibrium of system (1) is $\mathcal{E}^0 = \big(0, 0, 0, X_s^0, 0\big) = \left(0, 0, 0, \dfrac{\Lambda}{\mu_{s\epsilon}}, 0\right)$. Using the notations of [15] for the model system (1), the matrices F and V for the new infection terms and the remaining transfer terms are, respectively, given by

$$F = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \beta\,\dfrac{\Lambda}{\mu_{s\epsilon}} & 0 \end{pmatrix} \quad \text{and} \quad V = \begin{pmatrix} -k_m & \mu_{m\epsilon} & \rho & 0 \\ 0 & \rho + \mu_{f\epsilon} & 0 & -k_f \\ 0 & -\rho & \mu_{p\epsilon} & 0 \\ 0 & 0 & & \mu_{s\epsilon} + \alpha_s \end{pmatrix}$$

The basic reproduction number $\mathcal{R}_0$ is equal to the spectral radius of the matrix $F\,V^{-1}$, a simple computation gives:

$$\mathcal{R}_0 = \frac{\beta\,\rho\,k_f\,\Lambda}{\mu_{s\epsilon}\,\mu_{p\epsilon}\,(\mu_{f\epsilon}+\rho)\,(\mu_{s\epsilon}+\alpha_s)} = \frac{\beta\,\rho\,k_f\,X_s^0}{\mu_{p\epsilon}\,(\mu_{f\epsilon}+\rho)\,(\mu_{s\epsilon}+\alpha_s)}.$$

One can remark that there is a mistake in the formula for $\mathcal{R}_0$ provided in [10].

The basic reproductive number for system (1) measures the average number of new infections generated by a single infected individual in a completely susceptible population.

As it is well known (see, for instance, [15]), the local asymptotic stability of the disease-free equilibrium is completely determined by the value of $\mathcal{R}_0$ compared to unity, i.e., The disease-free equilibrium $\mathcal{E}^0$ of the system (1) is locally asymptotically stable if $\mathcal{R}_0 < 1$ and unstable if $\mathcal{R}_0 > 1$.

Hence $\mathcal{R}_0$ determines whether the disease will be prevalent in the given population or will go extinct.

Next, we discuss the global stability of infection-free equilibrium by using suitable Lyapunov function and LaSalle invariance principle for system (1). In recent years, the method of Lyapunov functions has been a popular technique to study global properties of population models. However, it is often difficult to construct suitable Lyapunov functions.

Theorem 3.1. *The disease-free equilibrium $\mathcal{E}^0$ of system (1) is globally asymptotically stable (GAS) on the nonnegative orthant $\mathbb{R}^5_+$ whenever $\mathcal{R}_0 \leq 1$.*

Proof. See Appendix A. □

Biologically speaking, Theorem 3.1 implies that schistosomiasis may be eliminated from the community if $\mathcal{R}_0 \leq 1$. One can remark that $\mathcal{R}_0$ does not depend on $\mu_{m\epsilon} = \mu_m + \epsilon_m$. Hence it is not helpful to try to control the the male schistosoma population and then one can take $\epsilon_m = 0$. Therefore the only way to eliminate schistosomiasis is to increase the killing rates of female schistosoma (ϵ_f), paired schistosoma (ϵ_p) and snails (ϵ_s) in order to have $\mathcal{R}_0 \leq 1$.

In the rest of this section, we show that the disease persists when $\mathcal{R}_0 > 1$. The disease is endemic if the infected fraction of the population persists above a certain positive level. The endemicity of a disease can be well captured and analyzed through the notion of uniform persistence. System (1) is said to be uniformly persistent in Ω if there exists constant $c > 0$, independent of initial conditions in $\overset{\circ}{\Omega}$ (the interior of Ω), such that all solutions $(X_m(t), X_f(t), X_p(t), X_s(t), X_i(t))$ of system (1) satysfy

$$\liminf_{t\to\infty} X_m(t) \geq c, \quad \liminf_{t\to\infty} X_f(t) \geq c, \quad \liminf_{t\to\infty} X_p(t) \geq c,$$

$$\liminf_{t\to\infty} X_s(t) > c, \quad \liminf_{t\to\infty} X_i(t) \geq c,$$

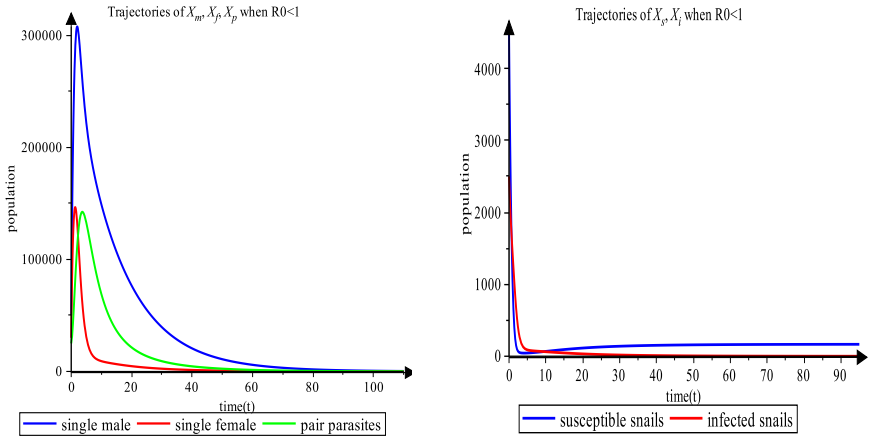
provided $(X_m(0), X_f(0), X_p(0), X_s(0), X_i(0)) \in \overset{\circ}{\Omega}$, (see [14], [2]).

Theorem 3.2. *System (1) is uniformly persistent in* Ω *if and only if* $\mathcal{R}_0 > 1$.

Proof. See Appendix B □

4. Numerical simulation

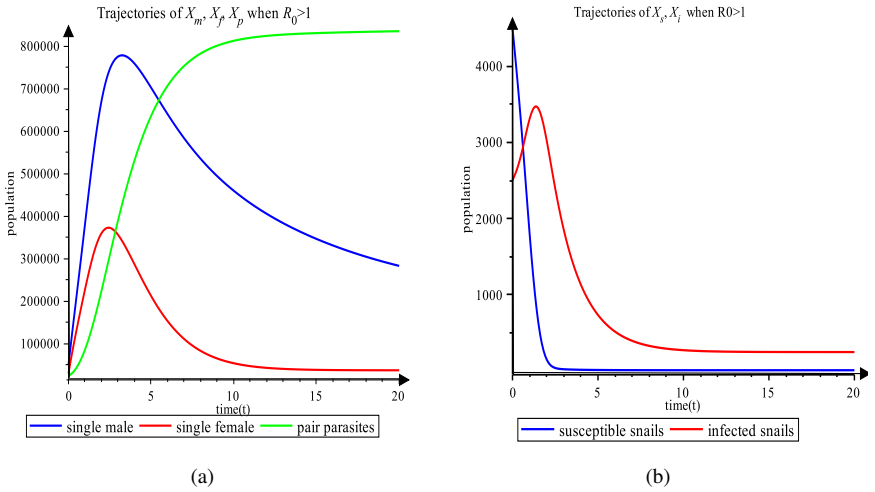
In this section, we use numerical simulations to illustrate the asymptotic stability and persistent results. Parameter values have been chosen in such a way that they are realist and at the same time obey the conditions for stability or persistent. Figure 1 illustrates the convergence of the dynamic of the system to the disease-free point.



(a) (b)

Figure 1. *Solutions of the schistosomiasis model (1) with parameter values defined as follows:* $k_f = 100$, $k_m = 145$, $\Lambda = 150$, $\beta = 0.000018$, $\alpha_s = 0.5$, $\mu_{f\epsilon} = 0.3$, $\mu_{m\epsilon} = 0.1$, $\mu_{p\epsilon} = 0.2$, $\rho = 0.467$, $\mu_{s\epsilon} = 0.9$. *These parameters correspond to* $\mathcal{R}_0 = 0.6$. *The initial condition is* $X_m = 50000$, $X_f = 30000$, $X_p = 25000$, $X_s = 4500$, $X_i = 2500$.

Figure 2 presents how the system persists and approaches the endemic point.



(a) (b)

Figure 2. *Solutions of the schistosomiasis model (1) with parameter values defined as follows:* $k_f = 100$, $k_m = 145$, $\Lambda = 150$, $\beta = 0.000018$, $\alpha_s = 0.5$, $\mu_{f\epsilon} = 0.2$, $\mu_{m\epsilon} = 0.1$, $\mu_{p\epsilon} = 0.02$, $\rho = 0.467$, $\mu_{s\epsilon} = 0.1$. *These parameters correspond to* $\mathcal{R}_0 = 157.5$. *The initial condition is* $X_m = 50000$, $X_f = 30000$, $X_p = 25000$, $X_s = 4500$, $X_i = 2500$.

5. CONCLUSION

In this paper, we have investigated the dynamical properties of a schistosomiasis model with mating structure which incorporates some control strategies and uses the minimum mating function. When the basic reproductive number $\mathcal{R}_0$ is less than 1, we have proved the global asymptotic stability of the disease free equilibrium $\mathcal{E}_0$. When the basic reproductive number $\mathcal{R}_0$ is greater than 1, the persistent of the endemic equilibrium $\mathcal{E}_h$ has been obtained.

6. References

[1] N. P. BHATIA AND G. P. SZEGÖ. *" Dynamical systems: Stability theory and applications "*. Springer Berlin-Heidelberg-New York, 1967.

[2] G. BUTLER, H. FREEDMAN, AND P. WALTMAN. " Uniformly persistent systems. " *Proceedings of the American Mathematical Society*, pages 425–430, 1986.

[3] C. CASTILLO-CHAVEZ, Z. FENG, AND D. XU. " A schistosomiasis model with mating structure and time delay. " *Mathematical biosciences*, 211(2):333–341, 2008.

[4] N. CHITNIS, J. M. HYMAN, AND J. M. CUSHING. " Determining important parameters in the spread of malaria through the sensitivity analysis of a mathematical model. " *Bulletin of mathematical biology*, 70(5):1272–1296, 2008.

[5] K. HADELER, R. WALDSTÄTTER, AND A. WÖRZ-BUSEKROS. " Models for pair formation in bisexual populations. " *Journal of mathematical biology*, 26(6):635–649, 1988.

[6] J. HOFBAUER AND J. W.-H. SO. " Uniform persistence and repellors for maps. " *Proceedings of the American Mathematical society*, 107(4):1137–1142, 1989.

[7] J. LASALLE. " The stability of dynamical systems, regional conference series in applied mathematics. " *SIAM, Philadelphia*, 1976.

[8] X. LIN AND J. W.-H. SO. " Global stability of the endemic equilibrium and uniform persistence in epidemic models with subpopulations. " *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics*, 34(03):282–295, 1993.

[9] A. PERLOWAGORA-SZUMLEWICZ. " The reaction of australorbis glabratus (biomphalaria glabrata) to infection with schistosoma mansoni. " *Rev Inst Med Trop Sao Paulo*, 10:219–228, 1968.

[10] L. QI AND J.-A. CUI. " A schistosomiasis model with mating structure. " In *Abstract and Applied Analysis*, volume 2013. Hindawi Publishing Corporation, 2013.

[11] L. SAVIOLI, S. STANSFIELD, D. A. BUNDY, A. MITCHELL, R. BHATIA, D. ENGELS, A. MONTRESOR, M. NEIRA, AND A. M. SHEIN. " Schistosomiasis and soil-transmitted helminth infections: forging control efforts. " *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 96(6):577–579, 2002.

[12] " Schistosomiasis. " `http://www.who.int/mediacentre/factsheets/fs115/en/index.html` 2010.

[13] S.-F. H. SCHMITZ AND C. CASTILLO-CHAVEZ. " A note on pair-formation functions. " *Mathematical and computer modelling*, 31(4):83–91, 2000.

[14] H. R. THIEME. " Epidemic and demographic interaction in the spread of potentially fatal diseases in growing populations. " *Mathematical biosciences*, 111(1):99–130, 1992.

[15] P. VAN DEN DRIESSCHE AND J. WATMOUGH. " Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. " *Mathematical biosciences*, 180(1):29–48, 2002.

[16] D. XU, J. CURTIS, Z. FENG, AND D. J. MINCHELLA. " On the role of schistosome mating structure in the maintenance of drug resistant strains. " *Bulletin of mathematical biology*, 67(6):1207–1226, 2005.

[17] P. ZHANG, G. J. SANDLAND, Z. FENG, D. XU, AND D. J. MINCHELLA. "Evolutionary implications for interactions between multiple strains of host and parasite. " *Journal of theoretical biology*, 248(2):225–240, 2007.

Appendix A. Proof of Theorem 3.1

Proof. We shall use the following notations: $x = (X_m, X_f, X_p, X_s, X_i)$, and $X_s^0 = \dfrac{\Lambda}{\mu_{s\epsilon}}$. To show the global stability of infection-free equilibrium of system (1), we use the following candidate Lyapunov function:

$$V(x) = \frac{\mu_{s\epsilon} + \alpha_s}{k_f} X_f + \frac{(\mu_{s\epsilon} + \alpha_s)(\mu_{f\epsilon} + \rho)}{k_f \rho} X_p + \int_{X_s^0}^{X_s} \frac{X_\tau - X_s^0}{X_\tau} d X_\tau + X_i \quad (2)$$

This function satisfies: $V(x) \geq 0$ for all $x \in \Omega$, and $V(x) = 0$ if and only if $x = (X_m, 0, 0, X_s^0, 0)$.

Taking the time derivative of the function V (defined by 2), along the solutions of system (1), we obtain

$$\begin{aligned}
\dot{V} &= \left(1 - \frac{X_s^0}{X_s}\right)(\Lambda - \mu_{s\epsilon} X_s - \beta X_s X_p) + (\beta X_s X_p - (\mu_{s\epsilon} + \alpha_s) X_i) \\
&+ \frac{(\mu_{s\epsilon} + \alpha_s)}{k_f}(k_f X_i - (\mu_{f\epsilon} + \rho)) X_f + \frac{(\mu_{s\epsilon} + \alpha_s)(\mu_{f\epsilon} + \rho)}{k_f \rho}(\rho X_f - \mu_{p\epsilon} X_p)
\end{aligned}$$

Using $\Lambda - \mu_{s\epsilon} X_s^0 = 0$, we get

$$\begin{aligned}
\dot{V} &= \left(1 - \frac{X_s^0}{X_s}\right)(-\mu_{s\epsilon} X_s + \mu_{s\epsilon} X_s^0) + \left[\beta X_s^0 X_p - \frac{(\mu_{s\epsilon} + \alpha_s)(\mu_{f\epsilon} + \rho)}{k_f \rho} \mu_{p\epsilon} X_p\right] \\
&= \mu_{s\epsilon} X_s^0 \left(1 - \frac{X_s^0}{X_s}\right)\left(1 - \frac{X_s}{X_s^0}\right) + \frac{\beta \Lambda}{\mu_{s\epsilon}}\left[1 - \frac{(\mu_{s\epsilon} + \alpha_s)(\mu_{f\epsilon} + \rho)\ \mu_{m\epsilon}\mu_{p\epsilon}}{k_f \rho \Lambda \beta}\right] X_p \\
&= \mu_{s\epsilon} X_s^0 \left(1 - \frac{X_s^0}{X_s}\right)\left(1 - \frac{X_s}{X_s^0}\right) + \frac{\beta \Lambda}{\mu_{s\epsilon}}\left[1 - \frac{1}{\mathcal{R}_0}\right] X_p \quad (3) \\
&= -\frac{\mu_{s\epsilon}}{X_s}(X_s^0 - X_s)^2 + \frac{\beta \Lambda}{\mu_{s\epsilon}}\left[1 - \frac{1}{\mathcal{R}_0}\right] X_p
\end{aligned}$$

Hence, $\dot{V} \leq 0$ if $\mathcal{R}_0 \leq 1$, and

$$\Omega \cap \{\dot{V} = 0\} = \begin{cases} \{x \in \Omega : x = (X_m, X_f, 0, X_s^0, X_i)\} \text{ if } \mathcal{R}_0 < 1 \\ \{x \in \Omega : x = (X_m, X_f, X_p, X_s^0, X_i)\} \text{ if } \mathcal{R}_0 = 1 \end{cases}$$

We will show that the largest invariant set $\mathcal{L}$ contained in $\Omega \cap \{\dot{V} = 0\}$ is reduced to the disease-free equilibrium $\mathcal{E}^0$.

Let $x = (X_m, X_f, X_p, X_s, X_i) \in \mathcal{L}$ and $x(t) = (X_m(t), X_f(t), X_p(t), X_s(t), X_i(t))$ the solution of (1) issued from this point. By invariance of $\mathcal{L}$, we have $X_s(t) \equiv X_s^0$ which imlies $\dot{X}_s(t) = 0 = \Lambda - \mu_s\, X_s(t) - \beta\, X_p(t)\, X_s(t) = \Lambda - \mu_s\, X_s^0 - \beta\, X_p(t)\, X_s^0$ and hence $X_p(t) = 0$ for all t. But, $X_p(t) \equiv 0$ implies that $\dot{X}_p(t) = 0$ for all t which implies, using system (1), that $X_f(t) = 0$ for all t. In the same way, it can be proved that $X_i(t) = 0$ for all t. Reporting in the first equation of system (1), one obtains that, in $\mathcal{L}$,

$$\dot{X}_m(t) = -\mu_{m\epsilon}\, X_m(t) \qquad \forall t$$

Thus the solution of (1) issued from $x = (X_m, X_f, X_p, X_s, X_i) \in \mathcal{L}$ is given by $x(t) = (X_m\, e^{-\mu_{m\epsilon}\, t}, 0, 0, X_s^0, 0)$ which clearly leaves Ω and hence $\mathcal{L}$ for $t < 0$ if $X_m \neq 0$. Therefore $\mathcal{L} = \{\mathcal{E}^0\}$ and hence $\mathcal{E}^0$ is a globally asymptotically stable equilibrium state for system (1) on the compact set Ω thanks to LaSalle invariance principle [7], (one can also see [1], Theorem 3.7.11, page 346). Since the set Ω is an attractive set, the DFE is actually GAS on the nonnegative orthant $\mathbb{R}^5_+$. □

Appendix B. Proof of Theorem 3.2

Proof. When $\mathcal{R}_0 \leq 1$, the infection-free equilibrium $\mathcal{E}^0$ is globally asymptotically stable which precludes any sort of persistence and hence $\mathcal{R}_0 > 1$ is a necessary condition for persistence. In order to show that $\mathcal{R}_0 > 1$ is a sufficient condition for uniform persistence, it suffices to verify conditions (1) and (2) of Theorem 4.1 in [6] (one can also see [8], Theorem 3.5).

We use the notations of [6] with $\mathscr{X} = \Omega$ and $\mathscr{Y} = \partial\Omega$. Let M be the largest invariant compact set in $\mathscr{Y}$. We have already seen that $M = \{\mathcal{E}^0\}$, and so M is isolated. To show that $\mathcal{W}^s(M)$ (the stable set of M) is contained in $\mathscr{Y} = \partial\Omega$, we use the following function:

$$\mathcal{F} = \frac{\mu_{s\epsilon} + \alpha_s}{k_f}\, X_f + \frac{(\mu_{s\epsilon} + \alpha_s)\,(\mu_{f\epsilon} + \rho)}{k_f\, \rho}\, X_p + X_i$$

The time derivative of $\mathcal{F}$ along the solutions of system (1) is given by

$$\begin{aligned}
\dot{\mathcal{F}} &= \beta\, X_s\, X_p - \frac{(\mu_{s\epsilon} + \alpha_s)\,(\mu_{f\epsilon} + \rho)}{k_f\, \rho}\, \mu_{p\epsilon}\, X_p \\
&= \left(\beta\, X_s - \frac{(\mu_{s\epsilon} + \alpha_s)\,(\mu_{f\epsilon} + \rho)}{k_f\, \rho}\, \mu_{p\epsilon}\right) X_p \\
&= \frac{\mu_{p\epsilon}\,(\mu_{s\epsilon} + \alpha_s)\,(\mu_{f\epsilon} + \rho)}{k_f\, \rho} \left(\beta\, X_s\, \frac{k_f\, \rho}{\mu_{p\epsilon}\,(\mu_{s\epsilon} + \alpha_s)\,(\mu_{f\epsilon} + \rho)} - 1\right) X_p \\
&= \frac{\mu_{p\epsilon}\,(\mu_{s\epsilon} + \alpha_s)\,(\mu_{f\epsilon} + \rho)}{k_f\, \rho} \left(\mathcal{R}_0\, \frac{X_s}{X_s^0} - 1\right) X_p
\end{aligned}$$

Since $\mathcal{R}_0 > 1$, we have $\dot{\mathcal{F}} > 0$ for $X_p > 0$ and $\dfrac{X_s^0}{\mathcal{R}_0} < X_s \leq X_s^0$. Therefore $\dot{\mathcal{F}} > 0$ in a neighborhood N of $\mathcal{E}^0$ relative to $\Omega \setminus \partial\Omega$. This implies that any solution starting in N must leave N at finite time and hence the stable set of M, $\mathcal{W}^s(M)$ is contained in $\partial\Omega$. □

Analysis a two strain infectious disease

Otto Adamou (1), M'hammed El Kahoui (2), Marie-Françoise Roy (3), Thierry van Effelterre (4)

(1) IREM, Université de Niamey, Niamey, Niger
otto_adamou@yahoo.com
(2) Department of Mathematics, FSSM, Cadi Ayyad University, Marrakesh, Morrocco
elkahoui@uca.ma
(3) IRMAR, Université de Rennes 1, Rennes, France
marie-francoise.roy@univ-rennes1.fr
(4) At the time of the study, Global Epidemiology, GSK Vaccines, Wavre, Belgium

RÉSUMÉ. Un exemple typique de modèle compartimental de transmission avec traitement antibiotique et vaccination, qui peut être étudié algébriquement est présenté. Les méthodes exactes du calcul formel sont utilisées pour déterminer les quatre équilibres du système d'équations différentielles ordinaires représentant le modèle et étudier leur stabilité ainsi que leurs bifurcations.

ABSTRACT. A typical example of a compartmental desease transmission model with antibiotic treatment and vaccination that can be dealt with algebraically is presented. Methods from computer algebra are used to find the four equilibria of the ordinary differential equations characterizing the model and to study their stability as well as their bifurcations.

MOTS-CLÉS : modèle, souche, antibiotique, vaccination, équilibre, stabilité, bifurcation transcritique, taux de reproduction effectif.

KEYWORDS : model, strain, antibiotic,vaccination, equilibrium, stability, transcritical bifurcation, effective reproduction number.

1. Introduction

Compartmental models are a classical tool to model the spread of infectious diseases. Such models have the important feature of being simple enough to allow effective computation but also sufficiently flexible to take into account several behaviors of infectious diseases such as latency, the effect of treatment as well as vaccination [1]. Usually, compartmental models lead to systems of ordinary differential equations (ODE) depending on parameters having a *disease free equilibrium* E_0 characterized by the absence of disease in the population. The most fundamental question is then to find conditions on the parameters so that the disease free equilibrium is globally (or at least locally) asymptotically stable. Many compartmental models have the following behavior : the disease free equilibrium is asymptotically stable if and only if a threshold quantity depending on the parameters, called the *basic reproduction number* and denoted R_0, is < 1. When $R_0 = 1$ a new equilibrium E_1 called the *endemic equilibrium* appears and exchanges stability with the disease free equilibrium through a transcritical bifurcation so that, when $R_0 > 1$, the equilibrium E_1 is asymptotically stable while E_0 is unstable. Such a behavior no longer holds when for example the pathogen agent responsible of the disease transmission has several strains.

In this paper we introduce and study a compartmental model of an infectious disease caused by a two-strain bacterial pathogen. We show how to use methods from real algebraic geometry [6] and computer algebra [5] to find all the equilibria of the ODE system describing the model and to study their stability as well as their bifurcations.

The paper is structured as follows. In Section 2 we present the details of the model. In Section 3 we compute the equilibria of the model by using Groebner bases theory [5]. The stability of these equilibria is then studied in Section 4. A relation between our study and the effective reproduction number is given in Appendix A. We also give details on the bifurcations of the equilibria in Appendix C and a simulation of the vaccination effect in Appendix D.

2. Presentation of the model

The model concerns a host population, a part of its individuals are under antibiotic (Ab) treatment against a two-strain bacterial pathogen. Individuals who are not under Ab treatment can be colonized by an antibiotic-susceptible (Ab-S) strain or by an antibiotic-resistant (Ab-R) strain of a bacterial pathogen, but not by both at the same time (i.e., there is a maximal competition between the two strains), while those under antibiotic treatment can only be colonized by the Ab-R strain. We assume there is a fitness cost for resistance such that the Ab-R strain is somewhat less transmissible than the Ab-S strain.

The host population is divided into seven compartments representing the fractions of the population in each state. There are four states representing individuals not under Ab treatment, namely susceptible individuals (S), colonized individuals by the Ab-S strain (I_1), colonized individuals by the Ab-R strain (I_2) and vaccinated individuals (V). The individuals in V are assumed to have a temporary complete immunity to infection by both strains. There are three states for individuals under Ab treatment, namely susceptible individuals (T), colonized individuals by the Ab-R strain (T_2) and vaccinated individuals who are currently under Ab treatment (V_T). As well as individuals in V, those in V_T are

assumed to have a temporary complete immunity against infection by both strains. The transfer diagram of the model is given in the following figure.

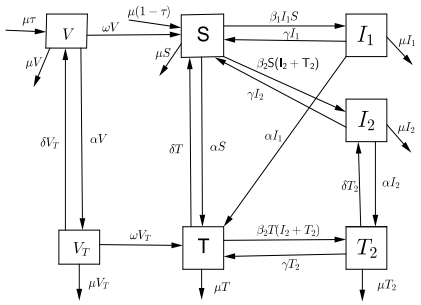


Figure 1. *Transfer diagram*

Mathematically, the model is represented by a parameter-dependent ODE system (E) of the form $\dot{x} = f(x, u)$, where the components of f are polynomials in terms of the states variables $x = (S, I_1, T, T_2, I_2, V, V_T)$ and the parameters $u = (\alpha, \beta_1, \beta_2, \gamma, \delta, \mu, \tau, \omega)$ as well. More precisely, the ODE system writes as

$$\begin{aligned}
\dot{S} &= \mu(1-\tau) + \omega V - \alpha S - \beta_1 S I_1 + \gamma I_1 + \delta T - \mu S - \beta_2 S I_2 - \beta_2 S T_2 + \gamma I_2 \\
\dot{I_1} &= \beta_1 I_1 S - \gamma I_1 - \alpha I_1 - \mu I_1 \\
\dot{T} &= \omega V_T + \alpha S - \delta T - \beta_2 T I_2 - \beta_2 T T_2 + \gamma T_2 + \alpha I_1 - \mu T \\
\dot{T_2} &= \beta_2 T I_2 + \beta_2 T T_2 - \gamma T_2 + \alpha I_2 - \delta T_2 - \mu T_2 \\
\dot{I_2} &= \beta_2 S I_2 + \beta_2 S T_2 - \gamma I_2 - \alpha I_2 + \delta T_2 - \mu I_2 \\
\dot{V} &= \mu\tau + \delta V_T - (\alpha + \mu + \omega) V \\
\dot{V_T} &= \alpha V - (\mu + \omega + \delta) V_T
\end{aligned}$$

where
α is the Ab treatment rate,
δ is the rate at which the effect of Ab treatment ends,
β_1 is the Ab-S strain transmission rate,
β_2 is the Ab-R strain transmission rate,
γ is the clearance rate,
μ is the birth rate which is assumed to be equal to the mortality rate,
τ is the vaccination coverage,
ω is the waning rate of vaccine efficiency.

The time scale here is the year and rates are expressed in terms of $1/t$. For example, $\alpha = 0.5$ means that an antibiotic treatment takes place every 2 years in average for each individual. The parameters $(\alpha, \gamma, \delta, \tau, \omega)$ are nonnegative while β_1, β_2, μ are positive.

One readily checks that $\mathbb{R}^7_+$ est positively invariant under the action of the vector field $f(x, u)$. On the other hand, if we let $P = S + I_1 + T + T_2 + I_2 + V + V_T$ then by summing up the seven equations in (E) we obtain

$$P' = \mu(1 - P)$$

and hence the affine hyperplane $P = 1$ is invariant under the action of the vector field $f(x, u)$. Thus, the set

$$\Omega = \{(S, I_1, T, T_2, I_2, V, V_T) \in \mathbb{R}^7_+ \mid S + I_1 + T + T_2 + I_2 + V + V_T = 1\}$$

is positively invariant under the action of $f(x, u)$. As we assume the host population to be constant we only need to study the dynamics of the ODE system (E) in the compact Ω.

3. Equilibria of the model

As we already have mentioned, the right hand side of the every equation in (E) is a polynomial in terms of the state variables and the parameters. Therefore, to find the equilibria of (E) we can resort to Groebner bases theory, e.g.; [5]. Notice that we are only interested in equilibria whose components are nonnegative and sum up to 1. Thus, to obtain the equilibria of the model we need first to solve the system of polynomial equations formed by the equation $S + I_1 + T + T_2 + I_2 + V + V_T = 1$ together with the seven equations obtained from (E) by putting to 0 the left hand side. The two last equations obtained from (E) form in fact a linear system whose unique solution is

$$(v, v_T) = \left(\frac{\mu\,\tau\,(\delta + \mu + \omega)}{(\mu + \omega)\,(\alpha + \delta + \mu + \omega)}, \frac{\mu\,\tau\alpha}{(\mu + \omega)\,(\alpha + \delta + \mu + \omega)} \right).$$

These will be the two last components of every equilibrium of the model. After respectively substituting v and v_T to V and V_T and then computing a Groebner basis of the obtained system with respect to the lexicographic order $S \prec I_1 \prec T \prec T_2 \prec I_2$ we obtain an equivalent, and much simpler, system (G) consisting of 6 equations. The first one depends only on S and has degree 3. Moreover, its three roots are all nonnegative and are given as

$$\begin{aligned} s_0 &= \frac{S_0}{D_0} \\ s_1 &= \frac{c}{\beta_1} \\ s_2 &= \frac{S_2}{D_2} \end{aligned}$$

with

$$\begin{aligned} c &= \alpha + \gamma + \mu \\ S_0 &= \mu(\alpha + \delta + \mu)(\delta + \mu + \omega)(1 - \tau) + \omega(\alpha\delta + (\delta + \mu)(\delta + \mu + \omega)) \\ D_0 &= (\mu + \omega)\,(\alpha + \delta + \mu)\,(\alpha + \delta + \mu + \omega) \\ S_2 &= \delta(\gamma + \mu)D_0 + (S_{21} + \omega(S_{22}(1 - \tau) + S_{23}) + \omega^2 S_{24})\beta_2 \\ D_2 &= (\alpha + \delta + \mu)(\alpha + \delta + \mu + \omega)\,(\beta_2(\mu(1 - \tau) + \omega) + (\delta + \alpha)(\mu + \omega))\,\beta_2, \end{aligned}$$

where

$$\begin{aligned} S_{21} &= \mu(\alpha + \delta + \mu)(1 - \tau)(\gamma\delta + \mu(\alpha + \gamma + \delta + \mu)) \\ S_{22} &= \mu c(\alpha + \delta + \mu) \\ S_{23} &= (\delta + \mu)^2(\gamma + \mu) + \alpha(\gamma\delta + \mu(\delta + \mu)) \\ S_{24} &= \gamma\delta + \mu(\alpha + \gamma + \delta + \mu) \end{aligned}$$

are positive quantities.

After specializing the variable S to s_0 in the system (G) and then solving for the other variables we obtain a unique solution E_0, whose coordinates are

$$\left(s_0, 0, \frac{T_0}{D_0}, 0, 0, v, v_T\right)$$

with

$$T_0 = \alpha(\mu(\alpha+\delta+\mu)(1-\tau)+\omega(\alpha+\delta+2\mu+\omega).$$

This is the disease free equilibrium of the model. Clearly, its coordinates are non-negative and so it has an epidemiological meaning for all the values of the parameters. Moreover the sum of its coordinates is equal to 1.

By substituting s_2 to S in the system (G) and then solving for the other variables we obtain a unique solution

$$E_2 = \left(s_2, 0, \frac{T_2}{D_2}, \frac{A_2T_{22}}{D_2(\mu+\omega)}, \frac{A_2I_{22}}{D_2(\mu+\omega)}, v, v_T\right)$$

where

$$\begin{array}{rcl} A_2 & = & (\mu(1-\tau)+\omega)\beta_2-(\gamma+\mu)(\mu+\omega) \\ T_2 & = & T_{21}\beta_2+(\gamma+\mu)D_0 \\ T_{22} & = & \mu(\alpha+\delta+\mu)(1-\tau)+\omega(\alpha+\delta+2\mu+\omega)\beta_2+D_0 \\ I_{22} & = & I_{221}\beta_2+I_{222} \end{array}$$

where

$$\begin{array}{rcl} T_{21} & = & \gamma\mu(\alpha+\delta+\mu(1-\tau)+\omega(\gamma(\alpha+\delta+2\mu+\omega)+\mu(\alpha+\delta+\mu)\tau) \\ I_{221} & = & \mu(\alpha+\delta+\mu)(\delta+\mu+\omega)(1-\tau)+\omega(\alpha\delta+(\delta+\mu+\omega)(\delta+\mu)) \\ I_{222} & = & \delta((\alpha+\delta+\mu)\omega(\alpha+\delta 2\mu+\omega)+\mu((\delta+\mu)^2+\alpha(\alpha+2\delta+2\mu))) \end{array}$$

Clearly, $T_2 \geq 0$, $I_{22} \geq 0$ et $I_{22} \geq 0$. Thus, E_2 has an epidemiological meaning if and only if $A_2 \geq 0$. Moreover the sum of its coordinates is equal to 1. This equilibrium, when its exists, corresponds to the absence of the first strain of the bacterial pathogen.

For $S = s_1$, and when solving (G) for the other variables, we obtain two equilibria E_1 and E_3. The coordinates of E_1 are

$$\left(s_1, \frac{A_1}{D_0\beta_1}, \frac{T_0}{D_0}, 0, 0, v, v_T\right)$$

with

$$A_1 = S_0\beta_1 - cD_0$$

This equilibrium has an epidemiological meaning if and only if $A_1 \geq 0$. Moreover the sum of its coordinates is equal to 1.

The coordinates of E_3 are

$$\left(s_1, \frac{A_3}{cD_0\beta_2\,(\beta_1-\beta_2)}, \frac{T_3}{\beta_1\beta_2}, \frac{A_4}{D_0\beta_1\beta_2}, \frac{A_4I_3}{cD_0\beta_1\beta_2\,(\beta_1-\beta_2)}, v, v_T\right)$$

where

$$\begin{array}{rcl} A_3 & = & \delta(\gamma+\mu)D_0\beta_1+A_{31}\beta_2+(A_{32}(1-\tau)+\omega(A_{33}\omega+A_{34})\beta_1\beta_2 \\ T_3 & = & (\gamma+\mu)\beta_1-c\beta_2 \\ A_4 & = & -(\gamma+\mu)D_0\beta_1+cD_0\beta_2+A_{41}\beta_1\beta_2 \\ I_3 & = & \delta\beta_1+c\beta_2 \end{array}$$

with

$$\begin{aligned}
A_{31} &= -c(\alpha+\delta+\mu)(\alpha+\delta+\mu+\omega)(((1-\tau)\mu+\omega)\beta_2+(\alpha+\delta)(\mu+\omega)) \\
A_{32} &= \mu(\alpha+\delta+\mu)(\delta(\gamma-\omega)+(\mu+\omega)(\alpha+\gamma+\delta+\mu)) \\
A_{33} &= \gamma\delta+\mu(\alpha+\gamma+\delta+\mu) \\
A_{34} &= (\delta+\mu)^2(\gamma+\mu)+\alpha(\delta(\gamma+\mu)+\mu^2) \\
A_{41} &= \alpha((\alpha+\delta+2\mu)\omega+\mu(\alpha+\delta+\mu)(1-\tau)+\omega^2)
\end{aligned}$$

The coordinates of E_3 are nonnegative if and only if $A_3 \geq 0$, $T_3 \geq 0$ and $A_4 \geq 0$. The fact that $T_3 \geq 0$ follows from the fact that $A_3 \geq 0$ (details are given in Appendix B). Moreover the sum of its coordinates is equal to 1. Thus, E_3 has an epidemiological meaning if and only if $A_3 \geq 0$ and $A_4 \geq 0$.

4. Stability of equilibria

In this section we study the local asymptotic stability of the four equilibria of the model. To this aim, we use the classical technique which consists in linearizing the system around the given equilibrium.

In the rest of this paper we let

$$Q_0 = (Z+\mu)(Z+\mu+\omega)(Z+\alpha+\delta+\mu)(Z+\alpha+\delta+\mu+\omega).$$

This polynomial is a common factor of the characteristic polynomials of all the four equilibria.

The characteristic polynomial P_0 of the Jacobian matrix $\partial_x f(u, E_0)$ factorizes as follows [7].

$$P_0 = (Z+c')\left(Z-\frac{A_1}{D_0}\right)\left(Z-\frac{A_2}{\mu+\omega}\right)Q_0,$$

with $c' = \alpha+\gamma+\delta+\mu$ Hence, the equilibrium E_0 is hyperbolic and locally asymptotically stable if and only if $A_1 < 0$ and $A_2 < 0$.

Concerning the equilibrium E_1, we have the following factorization of the characteristic polynomial P_1 of the Jacobian matrix $\partial_x f(u, E_1)$.

$$P_1 = (Z+c')\left(Z+\frac{A_1}{D_0}\right)\left(Z-\frac{A_4}{D_0\beta_1}\right)Q_0.$$

This shows that E_1 is hyperbolic and locally asymptotically stable if and only if $A_1 > 0$ and $A_4 < 0$.

For the equilibrium E_2, the characteristic polynomial P_2 of the Jacobian matrix $\partial_x f(u, E_2)$ factorizes as follows.

$$P_2 = (Z+c')\left(Z+\frac{A_2}{D_0}\right)\left(Z-\frac{A_3}{D_0\beta_1}\right)Q_0.$$

This shows that E_2 is hyperbolic and locally asymptotically stable if and only if $A_2 > 0$ and $A_3 < 0$.

The characteristic polynomial P_3 of the Jacobian matrix $\partial_x f(u, E_3)$ at the equilibrium E_3 does not completely factorize. We have in fact

$$P_3 = (Z^3+q_2Z^2+q_1Z+q_0)Q_0,$$

where q_0, q_1, q_2 are polynomials in terms of the parameters. We apply for this case the classical Liénard-Chipart criterion, e.g. ; [6], to the polynomial $Z^3 + q_2 Z^2 + q_1 Z + q_0$.

When we respectively substitute s_3 and t_3 to S and T we obtain

$$\begin{array}{rcl} q_0 & = & c\, I_1\, (I_2 + T_2)\, \beta_2\, (\beta_1 - \beta_2) \\ q_1 & = & (c' + (T_2 + I_2)\beta_2)\, ((T_2 + I_2)\beta_2 + I_1\beta_1) \\ q_2 & = & c + 2\, (I_2 + T_2)\, \beta_2 + I_1\, \beta_1 \end{array}$$

All three quantities are positive provided that E_3 has positive coordinates, that is $A_3 > 0, A_4 > 0$. The quantity that remains to check is $q_0(q_2 q_1 - q_0)$. After simplification we obtain $q_2 q_1 - q_0$ equal to

$$\begin{array}{l} 2(T_2 + I_2)^3 {\beta_2}^3 + (3(T_2 + I_2)^2 I_1\beta_1 + (T_2 + I_2)(cI_1 + 3c'(T_2 + I_2)))\beta_2{}^2 \\ +({I_1}^2(T_2 + I_2)\beta_1{}^2 + (3c' + \delta)I_1(T_2 + I_2)\beta_1 + {c'}^2(T_2 + I_2))\beta_2 \\ +c' I_1(I_1\beta_1 + c')\beta_1 \end{array}$$

which is positive at E_3 if $A_3 > 0, A_4 > 0$. Thus q_2, q_1, q_0 and $q_0(q_2 q_1 - q_0)$ are all positive at E_3 if $A_3 > 0, A_4 > 0$. The equilibrium E_3 is therefore hyperbolic and locally asymptotically stable if and only if $A_3 > 0$ and $A_4 > 0$.

We have thus the following result.

Theorem 1 *The model represented by the system (E) has four equilibria.*

1) A disease free equilibrium E_0 which exists for all values of the parameters. It is hyperbolic and locally asymptotically stable if and only if $A_1 < 0$ et $A_2 < 0$.

2) An equilibrium E_1 which exists if and only if $A_1 > 0$ and is hyperbolic and locally esymptotically stable if and only if $A_1 > 0$ and $A_4 < 0$.

3) An equilibrium E_2 which exists if and only if $A_2 \geq 0$ and is hyperbolic and locally asymptotically stable if and only if $A_2 > 0$ and $A_3 < 0$.

4) An equilibrium E_3 which exists if and only if $A_3 \geq 0$ and $A_4 \geq 0$, and is hyperbolic and locally asymptotically stable if and only if $A_3 > 0$ and $A_4 > 0$.

All the local codimension-one bifurcations of the system (E) are transcritical (details are given in Appendix C). To illustrate the results, we represent the curves $A_1 = 0$, $A_2 = 0$, $A_3 = 0$ et $A_4 = 0$ $\mathbb{R}_+ \times \mathbb{R}_+$ in termes of the parameters $0 < \beta_2 < \beta_1$. The other parameters $(\alpha, \gamma, \delta, \mu, \tau, \omega)$ are given fixed values.

The figure corresponds to $\alpha = 0.4, \gamma = 0.6, \delta = 0.3, \mu = 0.2, \tau = 0.35, \omega = 0.19$. These values have been chosen to make visible the stability domains. More realistic values, from the epidemiological point of view, could be $\alpha = 0.4, \gamma = 15, \delta = 60, \mu = 0.0125, \tau = 0.70, \omega = 0.19$.

Conclusion

In this paper we studied a two-strain compartmental model with vaccination and antibiotic treatment. All the equilibria and codimension-one local bifurcations of the model have been exactly characterized using computer algebra.

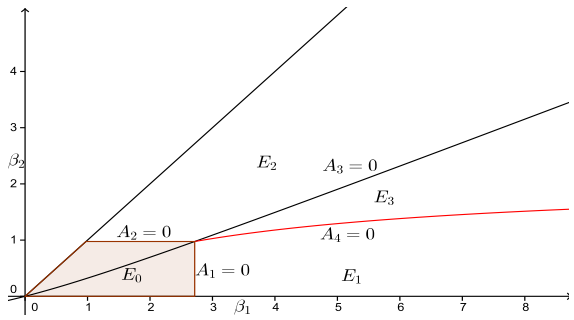


Figure 2. *Equilibria in the β_1, β_2 plane*

5. Bibliographie

[1] H. W HETHCOTE. « The mathematic of infectiouse disease. », *SIAM Rev.* , vol. 42, n° 4 : 599-653 (electronic), 2000.

[2] S. WIGGINS. « Introduction to applied nonlinear dynamical systemsand chaos », *Texts in Applied Mathematics. Springer-Verlag*, vol. 2, n° 2, 2003.

[3] N. FERGUSON, R. ANDERSON, S. GUPTA. « The effect of antibody-dependant enhancement on the transmissiondynamics and persistence of multiple strain pathogens. », *Proc. Natl. Acad. Sci. USA*, vol. 96, n° , 1999.

[4] L. BILLINGS , A. FIORILLO , I. B. SCHWARTZ. « Vaccinations in disease models with antibody-dependent enhancement. », *Math. Biosci., 2008.*, vol. 211, n° 2, 2008.

[5] D. COX , J. LITTLE., D. O'SHEA. « Ideals, varieties, and algorithms. », *Undergraduate Texts in Mathematics. Springer, New York*, vol. , n° 2, 2007.

[6] S. BASU, R. POLLACK , M.-F. ROY. « Algorithms in real algebraic geometry », *Algorithms and Computation in Mathematics. Springer-Verlag, Berlin*, vol. 10 , n° 2 , 2006.

[7] C. W. BROWN , M. EL KAHOUI , D. NOVOTNI, , A. WEBER.« Algorithmic methods for investigating equilibria in epidemic modeling. », *J. Symbolic Comput.*,vol. 41 , n° 11, 2006.

[8] P VAN DEN. DRIESSCHE , J. WATMOUGH. « Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. », *Math. Biosci*,vol. 180 , 2002.

A. Relation to the effective reproduction number

We check the results of Section 4 on the stability of the disease free equilibrium by using the notion of effective reproduction number (see [8]).

It is easy to see that the effective reproduction number of the first strain in the absence of the second one, i.e. ; $\beta_2 = 0, \beta_1 > 0$, is

$$R_{\text{eff}1} = \frac{s_0 \beta_1}{c}.$$

Also, one easily checks that the effective reproduction number of the second strain in the absence of the first one, i.e. ; $\beta_1 = 0, \beta_2 > 0$, is

$$R_{\text{eff}2} = \frac{(\mu(1-\tau)+\omega)\beta_2}{(\gamma+\mu)(\mu+\omega)}.$$

From the variations of the infectious compartments, namely

$$\begin{array}{rcl} \dot{I_1} & = & I_1 S\beta_1 - (\alpha+\gamma+\mu)I_1 \\ \dot{T_2} & = & T(I_2+T_2)\beta_2 - ((\delta+\gamma+\mu)T_2 - \alpha I_2) \\ \dot{I_2} & = & S(I_2+T_2)\beta_2 - (cI_2 - \delta T_2) \end{array}$$

and by letting $w = (I_1, T_2, I_2)$ we define

$$\mathcal{F}(w) = \begin{pmatrix} I_1 S\beta_1 \\ T(I_2+T_2)\beta_2 \\ S(I_2+T_2)\beta_2 \end{pmatrix}$$

This vector captures the rates at which new infected individuals, per infectious compartment, appear. We also define

$$\mathcal{V}(w) = \begin{pmatrix} cI_1 \\ (\gamma+\delta+\mu)T_2 - \alpha I_2 \\ cI_2 - \delta T_2 \end{pmatrix}$$

the vector whose components are the differences between the rate of individuals leaving an infectious compartment and the rate of those arriving at the same compartment. We then compute the matrices

$$F = \partial_w \mathcal{F}(w) = \begin{pmatrix} s_0\beta_1 & 0 & 0 \\ 0 & t_0\beta_2 & t_0\beta_2 \\ 0 & s_0\beta_2 & s_0\beta_2 \end{pmatrix}$$

and

$$V = \partial_w \mathcal{V}(w) = \begin{pmatrix} c & 0 & 0 \\ 0 & \gamma+\delta+\mu & -\alpha \\ 0 & -\delta & c \end{pmatrix}.$$

The matrix $F \cdot V^{-1}$ is called *the next generation matrix*, and its spectral radius is the effective reproduction number of the model.

$$F \cdot V^{-1} = \begin{pmatrix} \dfrac{s_0\beta_1}{c} & 0 & 0 \\ 0 & \dfrac{t_0\beta_2}{\gamma+\mu} & \dfrac{t_0\beta_2}{\gamma+\mu} \\ 0 & \dfrac{s_0\beta_2}{\gamma+\mu} & \dfrac{s_0\beta_2}{\gamma+\mu} \end{pmatrix}.$$

Clearly, $R_{\text{eff}1}$ is an eigenvalue of $F \cdot V^{-1}$. On the other hand, the determinant $|F \cdot V^{-1}|$ is zero, and hence 0 is an eigenvalue of $F \cdot V^{-1}$. The third eigenvalue of $F \cdot V^{-1}$ is the trace of the second block of $F \cdot V^{-1}$ and it is equal to

$$\frac{(s_0+t_0)\beta_2}{\gamma+\mu} = \frac{(\mu(1-\tau)+\omega)\beta_2}{(\gamma+\mu)(\mu+\omega)} = R_{\text{eff}2}.$$

Thus, the effective reproduction number of the model in question is $R_{\text{eff}} = \max(R_{\text{eff}1}, R_{\text{eff}2})$. This shows that E_0 hyperbolic and locally asymptotically stable if and only if $R_{\text{eff}} < 1$ [8]. This is clearly equivalent to the condition $A_1 < 0$ et $A_2 < 0$.

B. Conditions of the existence of the equilibrium E_3

As we have seen in Section 3, the equilibrium E_3 has nonnegative coordinates if and only if $A_3, T_3, A_4 \geq 0$. Here we show that $A_3 \geq 0$ implies $T_3 \geq 0$, and so E_3 has an epidemiological meaning if and only if $A_3 \geq 0$ and $A_4 \geq 0$. Let

$$\begin{aligned} D &= \mu\,(\alpha+\delta+\mu)\,(1-\tau)\,\left((\mu+\omega)c+\delta(\gamma+\mu)\right)\beta_2 + R, \\ R &= \omega(\alpha\delta\gamma + (\delta+\mu)(\delta+\mu+\omega) + \mu(\alpha+\delta+\mu)(\delta+\mu+\omega))\beta_2 + \delta(\gamma+\mu)D_0, \end{aligned}$$

Then we have

$$T_3 = \frac{\gamma+\mu}{D}A_3 + \frac{N}{D}$$

with

$$\begin{aligned} N &= \gamma\mu(\alpha+\delta+\mu)((1-\tau)\beta_2 + (\alpha+\delta+\mu)(\gamma+\mu)) + \omega L + \omega^2 M \\ M &= \mu c' + \gamma(\alpha+\delta+\beta_2) \\ L &= (\gamma+\mu)\alpha^2 + ((2\delta+3\mu)(\gamma+\mu) + \alpha(\gamma+\mu\tau)\beta_2) + K \\ K &= (\gamma+\mu)\delta^2 + \delta(3\mu(\gamma+\mu) + (\gamma+\mu\tau)\beta_2) + \mu(2\mu^2 + 2\mu\gamma + 2(\gamma+\mu\tau)\beta_2) \end{aligned}$$

which shows that $T_3 \geq 0$ whenever $A_3 \geq 0$.

C. Codimension-one bifurcations of equilibria

In this section we study the local codimension-one bifurcations of the system (E) when the parameters change. It turns out that all such bifurcations have a transcritical nature.

Stability exchange between E_0 and E_1. As we have seen, the equilibrium E_0 is hyperbolic and locally asymptotically stable if and only if $A_1 < 0$ and $A_2 < 0$. Moreover, E_1 has an epidemiological meaning if and only if $A_1 \geq 0$. When $A_1 = 0$, E_0 and E_1 become the same and their common characteristic polynomial

$$P_{01} = Z(Z+c')\left(Z - \frac{A_2}{\mu+\omega}\right)Q_0$$

has 0 as a simple root, while the other ones are negative. Thus, when A_1 moves from negative values to positive ones, E_0 becomes unstable while E_1 gains stability as long as $A_2 < 0$.

Stability exchange between E_0 and E_2. When $A_2 = 0$ and $A_1 < 0$ the two equilibria E_0 and E_2 become the same and their common characteristic polynomial

$$P_{02} = Z(Z+c')\left(Z - \frac{A_1}{D_0}\right)Q_0$$

has 0 as a simple root, while the others are negative. When A_2 moves from negative values to positive ones the equilibrium E_0 becomes unstable, while E_2 gains stability as long as $A_1 < 0$.

Stability exchange between E_2 and E_3. When $A_3 = 0$ and $A_2 > 0$ the two equilibria E_2 and E_3 are the same and their common characteristic polynomial

$$P_{23} = Z\left(Z + \frac{(\mu(1-\tau)+\omega)\beta_2 + (\alpha+\delta)(\mu+\omega)}{\mu+\omega}\right)\left(Z + \frac{A_2}{\mu+\omega}\right)Q_0$$

has 0 as simple root while the other ones are negative. Thus, when A_3 moves from negative values to positive ones the equilibrium E_2 becomes unstable while E_3 gains stability as long as $A_2 > 0$.

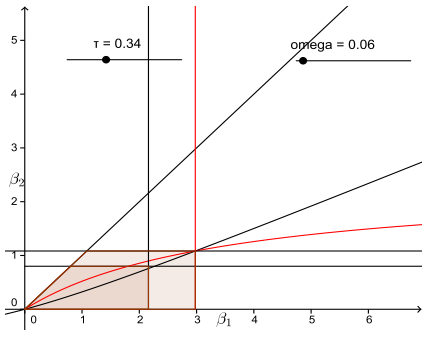
Stability exchange between E_1 and E_3. When $A_4 = 0$ and $A_1 > 0$ the two equilibria E_1 and E_3 become the same and their common characteristic polynomial

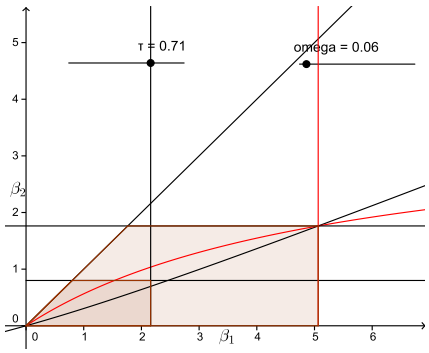
$$P_{13} = Z(Z + c')\left(Z + \frac{A_1}{D_0}\right)Q_0$$

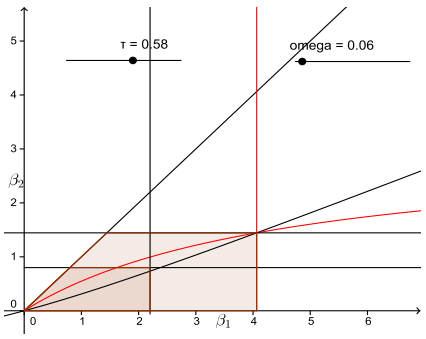
has 0 as simple root while the other are negative. Thus, when A_4 moves from negative values to positive ones E_1 becomes unstable while E_3 becomes stable as long as $A_1 > 0$

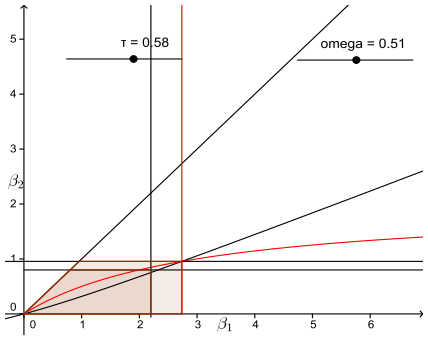
D. Simulation of the vaccination effect

In this section the parameters μ, γ, α and δ are given fixed values. For several values of τ and ω (the vaccination parameters) we represent the curves $A_1 = 0$, $A_2 = 0$, $A_3 = 0$ and $A_4 = 0$ as functions of $(\beta_1, \beta_2) \in \mathbb{R}^2_+$. The light-colored regions represent the domain of stability of the disease free equilibrium in the presence of vaccination, while the dark-colored ones correspond to the absence of vaccination. The figures show that the stability domain of E_0 increases in terms of τ but decreases in terms of ω.




τ = 0.71
omega = 0.06
β_2
β_1


τ = 0.58
omega = 0.06
β_2
β_1


τ = 0.58
omega = 0.51
β_2
β_1

Rubrique

Sensitivity of the electrocardiographic forward problem to the heart potential measurement noise and conductivity uncertainties

Rajae Aboulaich[1*] — Najib Fikal[1] — El Mahdi El Guarmah[1,2] — Nejib Zemzemi[3,4]

[1] Mohammed V University of Rabat, Mohammadia school of Engineering. LERMA and LIRIMA Laboratories. Av. Ibn Sina Agdal, Rabat Morocco
[2] Royal Air School, Informatics and Mathematics Department. DFST, BEFRA, POB40002, Marrakech, Morocco
[3] INRIA Bordeaux Sud-Ouest, Carmen project 200 rue de la vieille tour 33405 Talence Cedex, France
[4] IHU Liryc, Electrophysiology and heart modeling institute. Avenue du Haut-Lévêque, 33604 Pessac, France
(*) aboulaich@gmail.com

ABSTRACT. In this work we are interested in quantifying the conductivity and epicardial potential boundary data uncertainties for the forward problem of electrocardiography (ECG). Indeed these input data are very important for the computation of the torso potential and consequently for the computation of the ECG. We use a stochastic approach for two dimensional torso geometry. We attribute probability density functions for the various source of randomness, and apply stochastic finite elements based on generalized polynomial chaos method. This work is the first step in order to quantify the uncertainties in inverse problem, which the goal is to complete the epicardial data. The efficiency of this approach to solve the forward ECG problem and the usability to quantify the effect of organs conductivity and epicardial boundary data uncertainties in the torso are demonstrated through a number of numerical simulations on a 2D computational mesh of the torso geometry.

KEYWORDS : electrocardiographic forward problem, stochastic finite elements, polynomial chaos, uncertainty quantification, stochastic processes, stochastic Galerkin

1. Introduction

Many studies have been performed on the forward problem of electrocardiography, in order to create more accurate methods allowing to find the electrical potential on the heart surface. However the data required by the mathematical electrocardiographic model, is in practice subject to uncertainties due to measurement errors or modeling assumptions and the resulting lack of knowledge. Therefore the idea of uncertainties quantification has attracted much interest in the last few years [5, 4]. The goal is to propagate information on the uncertainty of input data to the solution of a PDE [6]. Moreover the electrical potential in the torso depends on some physical parameters and on the geometry of the patient. In this work we are interested in studying the effect of the conductivity uncertainties, and also epicardial boundary data, in the ECG forward problem solved via stochastic finite element method (SFEM). For this aim we consider a stochastic approach in which the parameters of the model will be viewed as having statistical distributions, then as result the solutions of the stochastic system obtained have statistical characteristics, and we can determine the mean and the standard deviation of the electrical potential in the torso.

2. Stochastic forward problem of electrocardiography

2.1. Function spaces and notation

We give in the following a short overview of the notations, and definition of the stochastic Sobolev space used throughout this paper. Let D be the spatial domain. Ω is sample space that belongs to a probability space (Ω, A, P), A denotes the σ-algebra of subsets of Ω, and let P be the probability measure. Following the theory of Wiener [7] , as well as Xiu and Karniadakis [6] , we can represent any general second-order random process $X(\omega)$, $\omega \in \Omega$, in terms of a collection of finite number of random variables. We represent this random process by a vector $\boldsymbol{\xi} = \boldsymbol{\xi}(\omega) = (\xi_1(\omega), ..., \xi_N(\omega)) \in \mathbb{R}^N$, where N is the dimension of the approximated stochastic space. We assume that each random variable is independent, its image space is given by $\Theta_i \equiv \xi_i(\Omega) \subset \mathbb{R}$. Each random variable is characterised by a probability density function (PDF) $\rho_i : \Theta_i \longrightarrow \mathbb{R}^+$, for $i = 1, \ldots, N$. Then, we define the joint PDF of the random vector $\boldsymbol{\xi}$

$$\rho(\boldsymbol{\xi}) = \prod_{i=1}^{N} \rho_i(\xi_i) \qquad \forall \boldsymbol{\xi} \in \Theta,$$

where the support of ρ is $\Theta = \prod_{i=1}^{N} \Theta_i$. The probability measure on Θ is $\rho(\boldsymbol{\xi})d\boldsymbol{\xi}$. As commented in [6], this allows us to conduct numerical formulations in the finite dimensional (N-dimensional) random space Θ.

In this paper we treat a stochastic problem of electrocardiography, we suppose that the conductivity parameter and the epicardial boundary data acts like two different and independent sources of uncertainties, which will be represented by two random process. For the conductivity parameter we define the probability space (respectively the vector of random variables, PDF, the PDF support) with (Ω_0, A_0, P_0), (respectively $\boldsymbol{\xi_0}, \rho_0, \Theta_0$) and with (Ω_1, A_1, P_1), (respectively $\boldsymbol{\xi_1}, \rho_1, \Theta_1$) for the epicardial data.

Let us denote $\Theta = \Theta_0 \times \Theta_1$ and $L^2(\Theta) = L^2(\Theta_0) \times L^2(\Theta_1)$ the space of random variables X with finite second moments :

$$\mathbb{E}[X^2(\boldsymbol{\xi_0}, \boldsymbol{\xi_1})] = \int_{\Theta_1} \left(\int_{\Theta_0} X^2(\boldsymbol{\xi_0}, \boldsymbol{\xi_1}) \rho(\boldsymbol{\xi_0}) d\boldsymbol{\xi_0} \right) \rho(\boldsymbol{\xi_1}) d\boldsymbol{\xi_1} < +\infty,$$

where $\mathbb{E}[.]$ denotes the mathematical expectation operator. This space is a Hilbert space with respect to the inner product :

$$\langle X, Y \rangle_{L^2(\Theta)} = \mathbb{E}[XY] = \int_{\Theta_1} \left(\int_{\Theta_0} XY(\boldsymbol{\xi_0}, \boldsymbol{\xi_1}) \rho(\boldsymbol{\xi_0}) d\boldsymbol{\xi_0} \right) \rho(\boldsymbol{\xi_1}) d\boldsymbol{\xi_1}$$

Additionally, we consider a spatial domain D and we define the tensor product Hilbert space $H^1(D) \otimes L^2(\Theta)$ of second-order random fields as :

$$L^2(D) \otimes L^2(\Theta) = \left\{ u : D \otimes \Theta \longrightarrow \mathbb{R}, \int_{\Theta_1} \left(\int_{\Theta_0} \int_D |u(x, \boldsymbol{\xi_0}, \boldsymbol{\xi_1})|^2 \, dx \rho(\boldsymbol{\xi_0}) d\boldsymbol{\xi_0} \right) \rho(\boldsymbol{\xi_1}) d\boldsymbol{\xi_1} \right\}$$

Analogously, the tensor product spaces $H^1(D) \otimes L^2(\Theta)$ and $H_0^1(D) \otimes L^2(\Theta)$ can be defined.

2.2. Stochastic formulation of the forward problem

Under our assumption the conductivity uncertainties and epicardial boundary data uncertainties do not interact, and they are supposed to be independent each other, consequently we represent the stochastic forward solution of the Laplace equation as random field depending to the both kinds of uncertainties. For the space domain we use a 2D computational mesh of the torso geometry (see Figure 1)

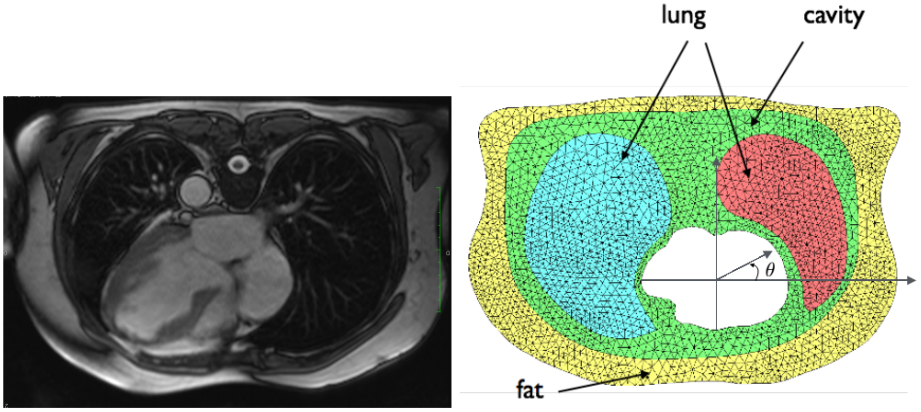


Figure 1 – MRI 2D slice of the torso (left), 2D computational mesh of the torso geometry showing the different regions of the torso considered in this study : fat, lungs and torso cavity, (right). The angle θ is the second polar coordinate.

Since we suppose that the conductivity parameter (σ) depends on the space (x) and on the stochastic variable (ξ_0), and the boundary epicardial data (f) depends on the space (x) and on a stochastic variable (ξ_1). Thus, the solution of the Laplace equation will depend on space and the both stochastic variables $u(x, \xi_0, \xi_1)$. The stochastic forward problem of electrocardiography can be written as follows

$$\begin{cases} \bigtriangledown.(\sigma(x, \xi_0) \bigtriangledown u(x, \xi_0, \xi_1)) = 0 & \text{in } D \times \Omega, \\ u(x, \xi_0, \xi_1) = f(x, \xi_1) & \text{on } \Gamma_{\text{i}} \times \Omega, \\ \sigma(x, \xi_0) \frac{\partial u(x, \xi_0, \xi_1)}{\partial n} = 0 & \text{on } \Gamma_{\text{c}} \times \Omega, \end{cases} \tag{1}$$

Where, Γ_i and Γ_c are the epicardial and torso boundaries respectively.
The weak formulation of SPDEs is based on an extension of the deterministic theory [3], test function become random fields and an integration over stochastic space is done with respect to the corresponding measure. Thus, the weak form involves expectations of the weak problem formulation in the physical space. Then, denoting by u_f the extension of f to the whole domain, we look for $\tilde{u} \in H_0^1(D) \otimes L^2(\Theta)$, where $\tilde{u} = u - u_f$ is the weak solution of (1), if for all $v \in H_0^1(D) \otimes L^2(\Theta)$, we have :

$$\mathbb{E}\left[\int_D \sigma(x,\xi_1)\nabla\tilde{u}(x,\xi_0,\xi_1).\nabla v(x,\xi_0,\xi_1)dx\right] + \mathbb{E}\left[\int_D \sigma(x,\xi_1)\nabla u_f(x,\xi_0).\nabla v(x,\xi_0,\xi_1)dx\right] = 0. \quad (2)$$

3. Descretization of the stochastic forward problem

A stochastic process $X(\xi)$ of a parameter or a variable X is represented by weighted sum of orthogonal polynomials $\{\Psi_i(\xi)\}$ denoting the generalized chaos polynomial. More details about the different choices of PDFs could be found in [6].
We have

$$X(\xi) = \sum_{i=0}^{p} \hat{X}_i \Psi_i(\xi),$$

where $\hat{X}_i$ are the projections of the random process on the stochastic basis $\{\Psi_i(\xi)\}_{i=1}^{p}$ with respect to the joint PDF ρ.

$$\hat{X}_i = \int_\Omega X(\xi)\Psi_i(\xi)d\rho = \langle X(\xi).\Psi_i(\xi)\rangle_\rho.$$

In order to solve the equation (2) we use the stochastic Galerkin (SG) method to compute the approximate solutions. To develop this method, we denote $Y_\sigma^p \subset L^2(\Theta_0)$ and $Y_{u_f}^p \subset L^2(\Theta_1)$ the stochastic approximation spaces, and we have $Y_\sigma^p \times Y_{u_f}^q \subset L^2(\Theta)$.
In our case we suppose that the conductivity parameter varies uniformly like in [4, 2] and we use the Legendre chaos polynomials which are more suitable for uniform probability density, in other hand we assigned Gaussian probability density to the epicardial boundary data, the corresponding stochastic orthogonal basis to Gaussian random field is Hermite chaos polynomials [6].

$$Y_\sigma^P = span\{\ L_0, ...,\ L_p\}.$$
$$Y_{u_f}^P = span\{H_0, ..., H_p\}.$$

In this study we have targeted to evaluate in the same time, two different source of uncertainties on the electrical potential, then σ, u_f and u are now expressed in the Galerkin space $Y_\sigma^p \times Y_{u_f}^q$ as follows :

$$\sigma(x,\xi_0) = \sum_{l=1}^{r} \hat{\sigma}_l(x)L_l(\xi_0). \quad (3)$$

$$u_f(x,\xi_1) = \sum_{k=1}^{q} (\tilde{u}_f)_k(x)H_k(\xi_1). \quad (4)$$

$$u(x,\xi_0,\xi_1) = \sum_{i=1}^{p}\sum_{j=1}^{q} \hat{u}_{ij}(x) L_i(\xi_0) H_j(\xi_1) \tag{5}$$

By substituting (4),(3),(5) into the stochastic diffusion equation (1) and by projecting the result on the polynomial basis $\{L_m(\xi_0)H_n(\xi_1)\}_{m,n=1}^{(p,q)}$:
For $m = 1, ..., q$ et $n = 1, ..., p$,

$$\begin{aligned}
\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{l=1}^{r} D_{jn} C_{iml} \nabla.(\hat{\sigma}_l(x)\nabla)\hat{u}_{ij}(x)) = 0 \qquad & \text{in } D, \\
\hat{u}_{11}(x) = (\tilde{u}_f)_1(x) \quad & \text{on } \Gamma_{\mathrm{i}}, \forall i = 1, ...p, \\
\hat{u}_{12}(x) = (\tilde{u_f})_2(x) \quad & \text{on } \Gamma_{\mathrm{i}} \forall i, = 1, ...p, \\
\hat{u}_{ij}(x) = 0 \quad & \text{on } \Gamma_{\mathrm{i}} \forall i, = 2, ...p, j = 3, ...q, \\
\hat{\sigma}_l(x)\frac{\partial \hat{u}_{ij}(x)}{\partial n} = 0 \quad & \text{on } \Gamma_{\mathrm{c}} \forall i = 1, ...p, j = 1, ...q,
\end{aligned} \tag{6}$$

Where $C_{iml} = \mathbb{E}[L_i(\xi_0), L_m(\xi_0), L_l(\xi_0)]$ et $D_{jn} = \mathbb{E}[H_j(\xi_1), H_n(\xi_1)]$.
For the spatial domain, we define a subspace $V_h \subset H_0^1(D)$ of standard Lagrange finite element functions on a triangulation of the domain D.

$$V_h := span\,\{\phi_1, \phi_2, ..., \phi_{Nx}\}$$

Obviously this ordering induces the following block structure of the linear system of equations :

$$\begin{bmatrix}
A^{(1,1;1,1)} & A^{(1,1;1,2)} & \dots & A^{(1,1;1,q)} & A^{(1,1;2,1)} & \dots & A^{(1,1;p,q)} \\
A^{(1,2;1,1)} & A^{(1,2;1,2)} & \dots & A^{(1,2;1,q)} & A^{(1,2;2,1)} & \dots & A^{(1,2;p,q)} \\
\vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\
A^{(1,q;1,1)} & A^{(1,q;1,2)} & \dots & A^{(1,q;1,q)} & A^{(1,q;2,1)} & \dots & A^{(1,q;p,q)} \\
A^{(2,1;1,1)} & A^{(2,1;1,2)} & \dots & A^{(2,1;1,q)} & A^{(2,1;2,1)} & \dots & A^{(2,1;p,q)} \\
\vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\
A^{(p,q;1,1)} & A^{(p,q;1,2)} & \dots & A^{(p,q;1,q)} & A^{(p,q;2,1)} & \dots & A^{(p,q;p,q)}
\end{bmatrix}
\begin{bmatrix} \hat{U}_{11} \\ \hat{U}_{12} \\ \vdots \\ \hat{U}_{1q} \\ \hat{U}_{21} \\ \vdots \\ \hat{U}_{pq} \end{bmatrix}
=
\begin{bmatrix} B^{11} \\ B^{12} \\ \vdots \\ B^{1q} \\ B^{21} \\ \vdots \\ B^{pq} \end{bmatrix}$$

where every matrix $A^{(i,j;m,n)} \in \mathbb{R}^{Nx} \times \mathbb{R}^{Nx}$ is a linear combination of finite element stiffness matrices

$$A^{(i,j;m,n)} = D_{j,n} \sum_{l=1}^{r} C_{iml} K_l \quad \forall i, m = 1, ...p; j, n = 1, ...q, \tag{7}$$

$$K_l = [K_l]_{h,t} = (\sigma_l \nabla \phi_h . \nabla \phi_t) \quad \forall l = 1, ...r, \tag{8}$$

h denotes the degrees of freedom of the nodes of the mesh in which the electrical potential values is unknown.
Similarly, every vector $B^{ij} \in R^{Nx}$ is a linear combination of finite element load vectors :

$$B^{ij} = \sum_{l=1}^{r} C_{iml} f_l \quad \forall i = 1, ...p, j = 1, ...q, \tag{9}$$

$$f_l = \sum_{x_h \in \Gamma_i} \hat{u}_{ij} \left(\sigma_l \nabla \phi_h . \nabla \phi_t\right) \quad \forall l = 1, ...r, \tag{10}$$

with h denoting the degrees of freedom of the (known) Dirichlet boundary conditions of the solution.

4. Results

In this section we conduct the numerical simulation obtained in order to show the influence of the conductivity variabilities and the epicardial potential data uncertainties on the electrical potential in the torso. For instance we suppose that the electrical potential in the heart boundary is equal to U_{ex}.

$$U_{ex} = sin(y).$$

Since we assume that the uncertainty of the conductivity value follows a uniform probability density, as probability density functions $\rho_0$we use the Legendre polynomials defined on the interval $\Omega = [-1, 1]$. We also suppose that the true conductivity uncertainty interval is centered by σ_T, the true conductivity see Table 1. In other hand U_{ex} will represent the mean of the Gaussian random field representing the epicardial boundary data uncertainty, we denote its stdev by (ν).

organ category	conductivity (σ_T :S/m)
lungs	0.096
torso cavity	0.200
fat	0.045

Tableau 1 – Conductivity values corresponding to the organs that are considered in the model.

In the following we present four cases, in the first case we only study the effect of epicardial boundary data uncertainties where we gradually increase the stdev ν from zero to 50%. In the second (respectively, third, fourth) case we add the effect of fat (respectively, cavity, lung) with $\pm 50\%$ of uncertainties. Figure 2 summarize the obtained results for all cases. First we see that the forward solution after adding epicardial boundary data uncertainties is more sensitive to the torso cavity and lung conductivities than it is for fat. This result is in line with the numerical results obtained in [4, 2]. Second we remark that the influence of organs conductivity uncertainties disappear when $\nu \geq 10^{-1}$ and all curves take the same values as the case with only epicardial boundary data uncertainty. Figure 3 displays an example for obtained results with respect to lung with $\pm 50\%$ of conductivity uncertainties and epicardial boundary data uncertainty with different values of ν. Figure 3(a) shows the mean value of $u(x, \xi_0, \xi_1)$. Figure3 (b) (respectively Figure 3(c), Figure3(d), and Figure3(e)) shows $u(x, \xi_0, \xi_1)$ stdev with respect to $\pm 50\%$ lung uncertainties and $\nu = 0.03$ (respectively $\nu = 0.05, \nu = 0.1, \nu = 0.5$), finally Figure3(f) represents the case supposing that there is no conductivity uncertainties.

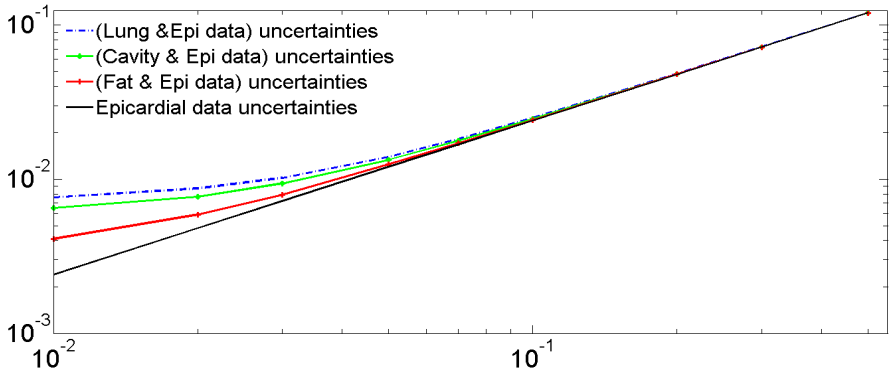


Figure 2 – The effects of $\pm 50\%$ uncertainty to each organ conductivity from it's reference conductivity, and different levels of uncertainty on the the epicardial boundary data. X-axis denote the different stdev value (ν) of the Gaussian epicardial data boundary field. Y-axis the mean square of the stdev value of $u(x, \xi_0, \xi_1)$.

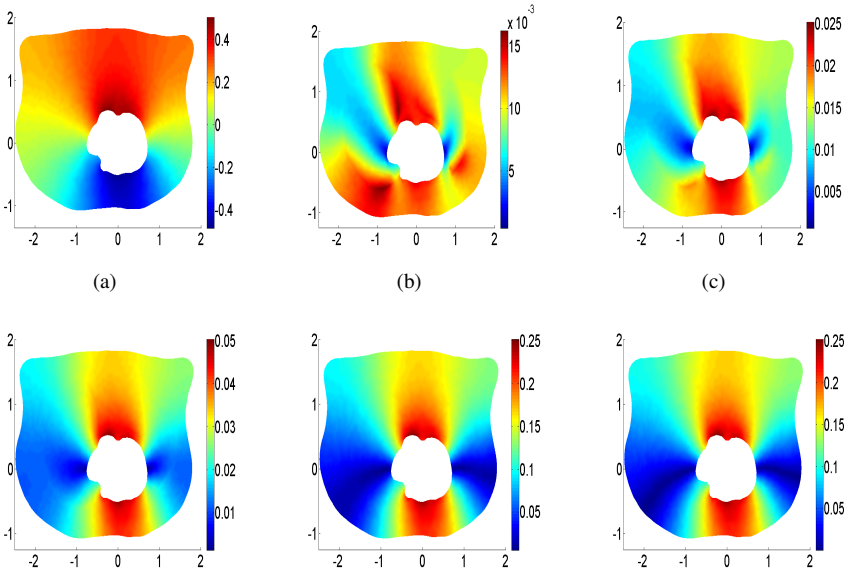


Figure 3 – Mean value of the SFE panel (a). Standard deviation of the SFE solution for $\pm 50\%$ of uncerainty for lung and epirdial data uncertainty for $\nu = 0.03$ panel(b) (respectively $\nu = 0.05$ panel(c), $\nu = 0.1$ panel(d) , $\nu = 0.5$ panel(e)). Panel(f) shows the Standard deviation of the SFE solution with only epirdial data uncertainty for $\nu = 0.5$

5. Conclusion :

This work is a novel approach allowing to study the sensitivity of forward problem of electrocardiography, taking into account two sources of uncertainty having different kinds of randomness, using for this the chaos polynomial and SFE method. Compared to [4] in which the authors study the sensitivity of forward problem with respect to a single source of uncertainty (organs conductivities), this study leads to a different computational framework of SFEM. The obtained results permit to classify the influence of each input parameter. We conclude that epicardial potential boundary data uncertainty have a strong effect on forward problem solution errors, compared to the organs conductivity, which at some level of boundary data uncertainty becomes insignificant. This finding suggests that the precise determination of the epicardial boundary data is very important. In a next work we will solve the inverse problem following the formulation presented in [1, 2], using stochastic approach developed in this work, and we will study the uncertainties in the case of the inverse problem.

Références

[1] R. ABOULAICH, A. BEN ABDA, M. KALLEL, « missing boundary data reconstruction via an approximate optimal control », *Inverse Problems and Imaging*, vol. 2, n° 4, 2008.

[2] R. ABOULAICH, N. FIKAL, E. EL GUARMAH, N. ZEMZEMI, « Stochastic Finite Element Method for torso conductivity uncertainties quantification in electrocardiography inverse problem », *Accepted in Math. Model. Nat. Phenom.*, vol. Jan. (2016).

[3] I. BABUSKA, R. TEMPONE, G.E. ZOURARIS, « Galerkin finite element approximations of stochastic elliptic partial differential equations », *SIAM Journal on Numerical Analysis*, vol. 42, n° 2, 2005.

[4] S.E. GENESER, R.M. KIRBY, R.S. MACLEOD, RM. KIRBY, « Application of stochastic finite element methods to study the sensitivity of ECG forward modeling to organ conductivity », *Biomedical Engineering, IEEE Transactions*, vol. 55, n° 1, 2008.

[5] A. OOSTEROM, G.J. HUISKAMP, « The effect of torso inhomogeneities on body surface potentials quantified using "tailored" geometry », *Journal of electrocardiology*, vol. 22, n° 1, 1989.

[6] D. XIU, G.E. KARNIADAKIS, « Modeling uncertainty in flow simulations via generalized polynomial chaos », *Elsevier, J.Comput.Phys.*, vol. 194, 2003.

[7] S. WIENER, « The homogeneous chaos », *Am. J. Math.*, vol. 60, 1998.

Hopf bifurcation properties of a delayed Predator-Prey model with threshold prey harvesting

Israël Tankam[a, d, *]- M. P. Tchinda[b, d]- JJ. Tewa[c, d]

a,* Department of Mathematics, University of Yaoundé I, PO Box 812 Yaoundé, Cameroon, israeltankam@gmail.com, Corresponding author, Tel.+(237) 698 74 58 64
b Department of Mathematics, University of Yaoundé I, PO Box 812 Yaoundé, Cameroon, tchindaplaire@yahoo.fr
c National Advanced School of Engineering University of Yaoundé I, Department of Mathematics and Physics P.O. Box 8390 Yaoundé, Cameroon, tewajules@gmail.com
d UMI 209 IRD/UPMC UMMISCO, University of Yaoundé I, Faculty of Science, CETIC Project team GRIMCAPE, University of Yaoundé I, Faculty of Science P.O. Box 812, Yaoundé, Cameroon

RÉSUMÉ. Dans cet article, nous étudions les propriétés de la bifurcation de Hopf pour un modèle prédateur-proie à retard avec deux seuils de collecte des proies et la stabilité des solutions périodiques obtenues via la bifurcation de Hopf en utilisant la théorie des formes normales et la réduction sur la variété centrale pour les équations différentielles fonctionnelles retardées (EDFr). Le long de cet article, nous supposerons toujours que les équations subissent une bifurcation de Hopf à l'équilibre positif $G(x^*, y^*)$ pour $\tau = \tau_0^j$,$(j = 0, 1, 2, ...)$ et les $\pm i\omega_0$ correspondent aux racines imaginaires pures de l'équation caractéristique.

ABSTRACT. In this paper,we shall study the properties of the Hopf bifurcations obtained for a delayed predator-prey model with threshold prey harvesting and the stability of bifurcated periodic solutions occurring through Hopf bifurcation by using the normal form theory and the center manifold reduction for retarded functional dfferential equations(RFDEs). Throughout this paper,we always assume that the equations undergoes Hopf bifurcation at the positive equilibrium $G(x^*, y^*)$ for $\tau = \tau_0^j$,$(j = 0, 1, 2, ...)$ and then $\pm i\omega_0$ is corresponding purely imaginary roots of the characteristic equation.

MOTS-CLÉS : Retard; prédateur-proie; bifurcation de hopf; bifurcations locales.

KEYWORDS : Delay; predator-prey; Hopf bifurcation; local bifurcations.

1. Introduction

In this paper, we consider a system of delayed differential equations modelling the predator-prey dynamic with a continuous double thresholds harvesting and a Holling response function of type III. Recently, Tankam & al. [3] considered the following model :

$$\left\{\begin{array}{lcl} \dot{x}(t) & = & \varphi(x(t)) - my(t)p(x(t)) - H(x(t)), \\ \dot{y}(t) & = & [-d + cmp(x(t-\tau))]y(t). \end{array}\right. \quad [1]$$

where x and y represent the population of preys and predators respectively. d is the natural mortality rate of the predators. c and m are positive constants. The function

$$\varphi(x) = rx\left(1 - \frac{x}{K}\right) \ , \quad [2]$$

models the dynamics of preys in absence of predators, r is the growth rate of preys for small values of x, while K is the capacity of the environment to support the preys. The function $p(x)$ is the Holling response function of type III given by :

$$p(x) = \frac{x^2}{a\,x^2 + b\,x + 1}, \quad [3]$$

(where $a > 0$ is constant and b is nonnegative constant)
and $H(x)$ is the double thresholds harvesting function given by :

$$H(x) = \left\{\begin{array}{cl} 0 & \text{if } x < T_1, \\ \dfrac{h(x - T_1)}{T_2 - T_1} & \text{if } T_1 \leq x \leq T_2, \\ h & \text{if } x \geq T_2, \end{array}\right. \quad [4]$$

This piecewise linear operator policy harvesting has been introduced in [1] in a predator-prey model without delay, where a Holling response function of type II was considered. In 2015, Tankam & al. have prooved that a Hopf bifurcation occurs. The following Theorem was given :

Theorem 1 *(Tankam & al., 2015) Suppose that a positive equilibrium E exists and is locally asymptotically stable for (1) with $\tau = 0$. Also let $\eta_0 = w_0^2$ be a positive root of $\eta^2 + \left[\varphi'(x^\star) - H'(x^\star) - mp'(x^\star)y^\star\right]^2 \eta - dmp'(x^\star){y^\star}^2 = 0$. Then there exists a $\tau = \tau^0$ such that E is locally asymptotically stable for $\tau \in (0, \tau^0]$ and unstable for $\tau > \tau^0$. Furthermore, the system undergoes a Hopf bifurcation at E when $\tau = \tau^0$.*

The aim of the following section is to study the properties of the Hopf bifurcation obtained by Theorem 1 and stability of bifurcated periodic solutions occuring through the Hopf bifurcation.

2. Properties of Hopf Bifurcation

In this section, we analyse the properties of the Hopf bifurcation using normal forms theory as in Hassard et al.[2]. The main result is given in Theorem 2 after having been

proved by pre-calculations.
Considering the equations (1) and $x_1(t) = x(t) - x^\star$ and $x_2(t) = y(t) - y^\star$; then system (1) is equivalent to the following two dimensional system :

$$\left\{\begin{array}{rcl} \dot{x}_1(t) &=& \Big[\varphi'(x^\star) - my^\star p'(x^\star) - H'(x^\star)\Big]\, x_1(t) - mp(x^\star)\, x_2(t) + f_1\big(x_1(t), x_2(t)\big), \\ \dot{x}_2(t) &=& cmy^\star p'(x^\star)\, x_1(t-\tau) + f_2\big(x_1(t), x_2(t), x_1(t-\tau)\big). \end{array}\right. \quad [5]$$

where

$$\begin{array}{rcl} f_1\big(x_1(t), x_2(t)\big) &=& \varphi\big(x_1(t) + x^\star\big) - m(x_2(t) + y^\star)p\big(x_1(t) + x^\star\big) - H\big(x_1 + x^\star\big) \\ &-& \Big[\varphi'(x^\star) - my^\star(t)p'(x^\star) - H'(x^\star)\Big]\, x_1(t) + mp(x^\star)\, x_2(t) \end{array}$$

and

$$\begin{array}{rcl} f_2\big(x_1(t), x_2(t), x_1(t-\tau)\big) &=& \Big[-d + cmp\big(x_1(t-\tau) + x^\star\big)\Big]\big(x_2(t) + y^\star\big) \\ &-& y^\star cmp\big(x^\star\big) x_1(t-\tau) \end{array}$$

let $\tau = \tau_j^0 + \mu$; then $\mu = 0$ is the Hopf bifurcation value of system (1) at the positive equilibrium $G(x^\star, y^\star)$. Since system (1) is equivalent to system (5), in the following discussion we shall consider mainly system (5).

In system (5), let $\bar{x}_k(t) = x_k(\tau t)$ and drop the bars for simplicity of notation. Then system (5) can be rewritten as a system of RFDEs in $\mathcal{C}\big([-1, 0], R^2\big)$ of the form :

$$\left\{\begin{array}{rcl} \dot{x}_1(t) &=& (\tau_j^0 + \mu)\Big[\varphi'(x^\star) - my^\star p'(x^\star) - H'(x^\star)\Big]\, x_1(t) - (\tau_j^0 + \mu)mp(x^\star)\, x_2(t) \\ &+& (\tau_j^0 + \mu) f_1\big(x_1(t), x_2(t)\big), \\ \dot{x}_2(t) &=& (\tau_j^0 + \mu)cmy^\star p'(x^\star)\, x_1(t-\tau) + (\tau_j^0 + \mu) f_2\big(x_1(t), x_2(t), x_1(t-\tau)\big). \end{array}\right. \quad [6]$$

Let us consider the following lemma proved in annex.

Lemma 1 *The system [6] is equivalent to*

$$\dot{x}(t) = A(\mu)x_t + R(\mu)x_t, \quad [7]$$

where $A(\mu)$ is linear. Besides, there exists an inner product $< \bullet, \bullet >$ and eigenvectors $q(\theta)$ and $q^\star(s)$ respectively of $A(0)$ and $A^\star$ such as $< q^\star(s), q(\theta) >= 1$, where $A^\star$ is the associate operator of A.

Using the same notations as in [2], we first compute the coordinates to describe the center manifold $\mathcal{C}_0$ at $\mu = 0$. Let x_t be the solution of Equation (5) when $\mu = 0$. Define

$$\begin{array}{rcl} z(t) &=& \langle q^* \ , \ x_t \rangle \\ W(t, \theta) &=& x_t(\theta) - 2\mathcal{R}_e\Big(z(t)q(\theta)\Big) \\ &=& x_t(\theta) - \Big(z(t)q(\theta) + \bar{z}(t)\bar{q}(\theta)\Big) \end{array} \quad [8]$$

On the center manifold $\mathcal{C}_0$ we have

$$W(t, \theta) = W(z, \bar{z}, \theta) \quad [9]$$

where

$$W(z,\bar{z},\theta) = W_{20}(\theta)\frac{z^2}{2} + W_{11}(\theta)z\bar{z} + W_{02}\frac{\bar{z}^2}{2} + W_{30}(\theta)\frac{z^3}{6} + \cdots \qquad [10]$$

z and $\bar{z}$ are local coordinates for center manifold C_0 in the direction of q^* and $\bar{q}^*$. Note that W is real if x_t is real. We only consider real solutions. For solution $x_t \in \mathcal{C}_0$ of (5), since $\mu = 0$, we have

$$\dot{z}(t) = iw_0\tau_j^0 z + \bar{q}^*(0)f\Big(0, W(z,\bar{z},0) + 2\mathcal{R}_e\big(z(t)q(\theta)\big)\Big) \equiv iw_0\tau_j^0 z + \bar{q}^*(0)f_0(z,\bar{z})$$

We rewrite this equation as

$$\dot{z}(t) = iw_0\tau_j^0 z + g\big(z,\bar{z}\big) \qquad [11]$$

where

$$g\big(z,\bar{z}\big) = g_{20}(\theta)\frac{z^2}{2} + g_{11}(\theta)z\bar{z} + g_{02}\frac{\bar{z}^2}{2} + g_{21}(\theta)\frac{z^2\bar{z}}{2} + \cdots \qquad [12]$$

The following lemma gives the values of the coefficients of $g(z,\bar{z})$.

Lemma 2

$$\begin{aligned}
g_{20} &= 2\tau_j^0\bar{D}\Big[-\Big(\tfrac{r}{K} + mp'(x^\star)\nu_1 + \tfrac{my^\star p''(x^\star)}{2}\Big)\\
&+ \bar{\nu}_1\Big(\tfrac{y^\star cmp''(x^\star)e^{-2iw_0\tau_j^0}}{2} + cmp'(x^\star)\nu_1 e^{-iw_0\tau_j^0}\Big)\\
g_{02} &= 2\tau_j^0\bar{D}\Big[-\Big(\tfrac{r}{K} + mp'(x^\star)\bar{\nu}_1 + \tfrac{my^\star p''(x^\star)}{2}\Big)\\
&+ \bar{\nu}_1\Big(\tfrac{y^\star cmp''(x^\star)e^{2iw_0\tau_j^0}}{2} + cmp'(x^\star)\bar{\nu}_1 e^{iw_0\tau_j^0}\Big)\\
g_{11} &= 2\tau_j^0\bar{D}\Big[-\Big(\tfrac{r}{K} + mp'(x^\star)\mathcal{R}_e\{\nu_1\} + \tfrac{my^\star p''(x^\star)}{2}\Big)\\
&+ \bar{\nu}_1\Big(\tfrac{y^\star cmp''(x^\star)}{2} + cmp'(x^\star)\mathcal{R}_e\Big\{\nu_1 e^{iw_0\tau_j^0}\Big\}\Big)\\
g_{21} &= \tau_j^0\bar{D}\Big[-\tfrac{r}{K}\Big(4W_{11}^{(1)}(0) + 2W_{20}^{(1)}(0)\Big)\\
&- mp'(x^\star)\Big(2W_{11}^{(2)}(0) + W_{20}^{(2)}(0) + \bar{\nu}_1 W_{20}^{(1)}(0) + 2\nu_1 W_{11}^{(1)}(0)\Big)\\
&- \tfrac{mp''(x^\star)}{2}(2\bar{\nu}_1 + 4\nu_1) - \tfrac{my^\star p''(x^\star)}{2}\Big(4W_{11}^{(1)}(0) + 2W_{20}^{(1)}(0)\Big)\\
&+ \bar{\nu}_1 my^\star p''(x^\star)\Big(2W_{11}^{(1)}(-1) + W_{20}^{(1)}(-1)e^{iw_0\tau_j^0}\Big)\\
&+ \bar{\nu}_1 cmp'(x^\star)\Big(\bar{\nu}_1 W_{20}^{(1)}(-1) + W_{20}^{(2)}(0)e^{iw_0\tau_j^0} + 2W_{11}^{(2)}(0)e^{-iw_0\tau_j^0} + 2\nu_1 W_{11}^{(1)}(-1)\Big)\\
&+ \tfrac{cmp''(x^\star)}{2}\Big(4\nu_1 + 2\bar{\nu}_1 e^{-2iw_0\tau_j^0}\Big)\Big]
\end{aligned} \qquad [13]$$

Since there are $W_{20}(\theta)$ and $W_{11}(\theta)$ in g_{21}, we still need to compute them. From (35) (*cf Annex 1*) and (8), we have :

$$\begin{aligned} \dot{W} &= \dot{x}_t - \dot{z}q - \dot{\bar{z}}\bar{q} \\ &= \begin{cases} AW - 2\mathcal{R}_e\Big\{\bar{q}^*(0)f_0q(\theta)\Big\}, & \theta \in [-1;0)\,; \\ AW - 2\mathcal{R}_e\Big\{\bar{q}^*(0)f_0q(\theta)\Big\} + f_0, & \theta = 0. \end{cases} \\ &\equiv^{\text{def}} AW + \mathcal{H}\big(z,\bar{z},\theta\big) \end{aligned} \quad [14]$$

where

$$\mathcal{H}\big(z,\bar{z},\theta\big) = \mathcal{H}_{20}(\theta)\frac{z^2}{2} + \mathcal{H}_{11}(\theta)z\bar{z} + \mathcal{H}_{02}(\theta)\frac{\bar{z}^2}{2} + \cdots \quad [15]$$

Substituting the corresponding series into (14) and comparing the coefficients, we obtain

$$\begin{aligned} (A - 2iw_0\tau_j^0)W_{20}(\theta) &= -\mathcal{H}_{20}(\theta) \\ AW_{11}(\theta) &= -\mathcal{H}_{11}(\theta) \end{aligned} \quad [16]$$

From (14), we know that for $\theta \in [-1, 0)$,

$$\mathcal{H}\big(z,\bar{z},\theta\big) = -\bar{q}^*(0)f_0q(\theta) - q^*(0)\bar{f}_0\bar{q}(\theta) = -g(z,\bar{z})q(\theta) - \bar{g}(z,\bar{z})\bar{q}(\theta) \quad [17]$$

Comparing the coefficient with (15), we get :

$$-g_{20}q(\theta) - \bar{g}_{02}\bar{q}(\theta) = H_{20}(\theta) \quad [18]$$

$$-g_{11}q(\theta) - \bar{g}_{11}\bar{q}(\theta) = H_{11}(\theta) \quad [19]$$

From (16) and (18) and the definition of A, it follows that

$$\dot{W}(\theta) = 2iw_0\tau_j^0W_{20} + g_{20}q(\theta) + \bar{g}_{02}\bar{q}(\theta) \quad [20]$$

Notice that $q(\theta) = \Big(1, \nu_1\Big)^T e^{iw_0\tau_j^0\theta}$. Hence,

$$W_{20}(\theta) = \frac{ig_{20}}{w_0\tau_j^0}q(0)e^{iw_0\tau_j^0\theta} + \frac{i\bar{g}_{02}}{3w_0\tau_j^0}\bar{q}(0)e^{-iw_0\tau_j^0\theta} + E_1e^{2iw_0\tau_j^0\theta} \quad [21]$$

where $E_1 = \Big(E_1^{(1)}, E_1^{(2)}\Big) \in R^2$ is a constant vector. Similarly, from (16) and (19), we obtain

$$W_{11}(\theta) = -\frac{ig_{11}}{w_0\tau_j^0}q(0)e^{iw_0\tau_j^0\theta} + \frac{i\bar{g}_{11}}{w_0\tau_j^0}\bar{q}(0)e^{-iw_0\tau_j^0\theta} + E_2 \quad [22]$$

where $E_2 = \Big(E_2^{(1)}, E_2^{(2)}\Big) \in R^2$ is also a constant vector.

In what follows, we will seek appropriate E_1 and E_2. From the definition of A and (16), we obtain

$$\int_{-1}^{0} d\eta(\theta)W_{20}(\theta) = 2iw_0\tau_jW_{20}(0) - H_{20}(0) \quad [23]$$

$$\int_{-1}^{0} d\eta(\theta) W_{11}(\theta) = -H_{11}(0) \qquad [24]$$

where $\eta(\theta) = \eta(0, \theta)$. By (14), we have

$$H_{20}(0) = -g_{20}q(0) - \bar{g}_{02}\bar{q}(0) + 2\tau_j^0 \begin{pmatrix} -\frac{r}{K} - mp'(x^\star)\nu_1 - \frac{my^\star p''(x^\star)}{2} \\ \frac{y^\star cmp''(x^\star)}{2} e^{-2iw_0\tau_j^0} + cmp'(x^\star)\nu_1 e^{-iw_0\tau_j^0} \end{pmatrix} \qquad [25]$$

$$H_{11}(0) = -g_{11}q(0) - \bar{g}_{11}\bar{q}(0) + 2\tau_j^0 \begin{pmatrix} -\frac{r}{K} - mp'(x^\star)\mathcal{R}_e\{\nu_1\} - \frac{my^\star p''(x^\star)}{2} \\ \frac{y^\star cmp''(x^\star)}{2} + cmp'(x^\star)\mathcal{R}_e\left\{\nu_1 e^{iw_0\tau_j^0}\right\} \end{pmatrix} \qquad [26]$$

Substituting (21) and (25) into (23) and noticing that

$$\begin{array}{rcl} \left(iw_0\tau_j^0 I - \int_{-1}^{0} e^{iw_0\tau_j^0\theta} d\eta(\theta)\right) q(0) & = & 0 \\ \left(-iw_0\tau_j^0 I - \int_{-1}^{0} e^{-iw_0\tau_j^0\theta} d\eta(\theta)\right) \bar{q}(0) & = & 0 \end{array} \qquad [27]$$

we obtain

$$\left(2iw_0\tau_j^0 I - \int_{-1}^{0} e^{2iw_0\tau_j^0\theta} d\eta(\theta)\right) E_1 =$$

$$2\tau_j^0 \begin{pmatrix} -\frac{r}{K} - mp'(x^\star)\nu_1 - \frac{my^\star p''(x^\star)}{2} \\ \frac{y^\star cmp''(x^\star)}{2} e^{-2iw_0\tau_j^0} + cmp'(x^\star)\nu_1 e^{-iw_0\tau_j^0} \end{pmatrix}$$

This leads to

$$\begin{pmatrix} 2iw_0 - \varphi'(x^\star) + my^\star p'(x^\star) + H'(x^\star) & mp(x^\star) \\ y^\star cmp'(x^\star) e^{-2iw_0\tau_j^0} & 2iw_0 \end{pmatrix} E_1$$

$$= 2 \begin{pmatrix} -\frac{r}{K} - mp'(x^\star)\nu_1 - \frac{my^\star p''(x^\star)}{2} \\ \frac{y^\star cmp''(x^\star)}{2} e^{-2iw_0\tau_j^0} + cmp'(x^\star)\nu_1 e^{-iw_0\tau_j^0} \end{pmatrix}$$

Solving this system for E_1, we obtain

$$E_1^{(1)} = \frac{2}{\sigma} \begin{vmatrix} -\frac{r}{K} - mp'(x^\star)\nu_1 - \frac{my^\star p''(x^\star)}{2} & mp(x^\star) \\ \frac{y^\star cmp''(x^\star)}{2} e^{-2iw_0\tau_j^0} + cmp'(x^\star)\nu_1 e^{-iw_0\tau_j^0} & 2iw_0 \end{vmatrix}$$

$$E_1^{(2)} =$$

$$\frac{2}{\sigma} \begin{vmatrix} 2iw_0 - \varphi'(x^\star) + my^\star p'(x^\star) + H'(x^\star) & -\frac{r}{K} - mp'(x^\star)\nu_1 - \frac{my^\star p''(x^\star)}{2} \\ y^\star cmp'(x^\star) e^{-2iw_0\tau_j^0} & \frac{y^\star cmp''(x^\star)}{2} e^{-2iw_0\tau_j^0} + cmp'(x^\star)\nu_1 e^{-iw_0\tau_j^0} \end{vmatrix}$$

where

$$\sigma = \begin{vmatrix} 2iw_0 - \varphi'(x^\star) + my^\star p'(x^\star) + H'(x^\star) & mp(x^\star) \\ y^\star cmp'(x^\star)e^{-2iw_0\tau_j^0} & 2iw_0 \end{vmatrix}$$

Similarly, substituting (22) and (26) into (24), we get

$$\begin{pmatrix} \varphi'(x^\star) - my^\star p'(x^\star) - H'(x^\star) & -mp(x^\star) \\ -y^\star cmp'(x^\star) & 0 \end{pmatrix} E_2$$

$$= 2 \begin{pmatrix} -\frac{r}{K} - mp'(x^\star)\mathcal{R}_e\{\nu_1\} - \frac{my^\star p''(x^\star)}{2} \\ \frac{y^\star cmp''(x^\star)}{2} + cmp'(x^\star)\mathcal{R}_e\left\{\nu_1 e^{iw_0\tau_j^0}\right\} \end{pmatrix}$$

and hence

$$E_2^{(1)} = \frac{2}{\varrho} \begin{vmatrix} -\frac{r}{K} - mp'(x^\star)\mathcal{R}_e\{\nu_1\} - \frac{my^\star p''(x^\star)}{2} & -mp(x^\star) \\ \frac{y^\star cmp''(x^\star)}{2} + cmp'(x^\star)\mathcal{R}_e\left\{\nu_1 e^{iw_0\tau_j^0}\right\} & 0 \end{vmatrix}$$

$$E_2^{(2)} = \frac{2}{\varrho} \begin{vmatrix} \varphi'(x^\star) - my^\star p'(x^\star) - H'(x^\star) & -\frac{r}{K} - mp'(x^\star)\mathcal{R}_e\{\nu_1\} - \frac{my^\star p''(x^\star)}{2} \\ -y^\star cmp'(x^\star) & \frac{y^\star cmp''(x^\star)}{2} + cmp'(x^\star)\mathcal{R}_e\left\{\nu_1 e^{iw_0\tau_j^0}\right\} \end{vmatrix}$$

where

$$\varrho = \begin{vmatrix} \varphi'(x^\star) - my^\star p'(x^\star) - H'(x^\star) & -mp(x^\star) \\ -y^\star cmp'(x^\star) & 0 \end{vmatrix}$$

Thus, we can determine W_{20} and W_{11} from (21) and (22). Furthermore, g_{21} in (13) can be expressed by the parameters and delay. Thus, we can compute the following values :

$$\begin{aligned} C_1(0) &= \frac{i}{2w_0\tau_j^0}\left(g_{20}g_{11} - 2|g_{11}|^2 - \frac{|g_{02}|^2}{3}\right) + \frac{g_{21}}{2} \\ \nu_2 &= -\frac{\mathcal{R}_e\{C_1(0)\}}{\mathcal{R}_e\left\{\lambda'(\tau_j^0)\right\}} \\ \beta_2 &= 2\mathcal{R}_e\{C_1(0)\} \\ P_2 &= -\frac{\mathcal{I}_m\{C_1(0)\} + \nu_2\mathcal{I}_m\left\{\lambda'(\tau_j^0)\right\}}{w_0\tau_j^0} \end{aligned} \qquad [28]$$

which determine the qualities of bifurcating periodic solution in the center manifold at the critical value τ_j^0.

Theorem 2 *: In Eq. (28), the sign of ν_2 determines the direction of the Hopf bifurcation. Thus, if $\nu_2 > 0$, then the Hopf bifurcation is supercritical and the bifurcating periodic*

solution exists for $\tau_1 > \tau_1^0$. If $\nu_2 < 0$, then the Hopf bifurcation is subcritical and the bifurcating periodic solution exists for $\tau_1 < \tau_1^0$. β_2 determines the stability of the bifurcating periodic solution : The bifurcating periodic solutions are stable if $\beta_2 < 0$ and unstable if $\beta_2 > 0$. T_2 determines the period of the bifurcating periodic solutions : the period increases if $P_2 > 0$ and decreases if $P_2 < 0$.

3. Bibliographie

[1] BOHN J., REBAZA J., SPEER K., « Continuous Threshold Prey Harvesting in Predator-Prey Models », *World Academy of Science, Engineering and Technology*, n° 79, 2011.

[2] HASSARD B.D., KAZARINOFF N.D., WAN Y.H., « Theory and Applications of Hopf Bifurcation », *Cambridge University, Cambridge*, 2011.

[3] TANKAM I., TCHINDA M. P., MENDY A. , Lam M. , Tewa J.J. , BOWONG S., « Local Bifurcations and Optimal Theory in a Delayed Predator-Prey Model with Threshold Prey Harvesting », *International Journal of Bifurcation and Chaos*, vol. 25, n° 07, 2015.

Annex 1 : Proof of the Lemma 1

Define the linear operator $L(\mu) : \mathcal{C} \to R^2$ and the nonlinear operator $f(\cdot, \mu) : \mathcal{C} \to R^2$ by :

$$\begin{aligned} L_\mu(\phi) &= (\tau_j^0 + \mu) \begin{pmatrix} \varphi'(x^\star) - my^\star p'(x^\star) - H'(x^\star) & -mp(x^\star) \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \phi_1(0) \\ \phi_2(0) \end{pmatrix} \\ &+ (\tau_j^0 + \mu) \begin{pmatrix} 0 & 0 \\ y^\star cmp'(x^\star) & 0 \end{pmatrix} \begin{pmatrix} \phi_1(-1) \\ \phi_2(-1) \end{pmatrix} \end{aligned} \quad [29]$$

and

$$f(\phi, \mu) = (\tau_j^0 + \mu) \begin{pmatrix} f_1\big(\phi_1(0), \phi_2(0)\big) \\ f_2\big(\phi_1(0), \phi_2(0), \phi_1(-1)\big) \end{pmatrix} \quad [30]$$

respectively, where $\phi = (\phi_1, \phi_2)^T \in \mathcal{C}$.

By the Riesz representation theorem, there exists a 2×2 matrix function $\eta(\theta, \mu)$, $-1 \leq \theta \leq 0$ whose elements are of bounded variation such that

$$L_\mu(\phi) = \int_{-1}^{0} d\eta(\theta, \mu)\phi(\theta) \quad \text{for} \quad \phi \in \mathcal{C}\big([-1, 0], R^2\big). \quad [31]$$

In fact, we can choose

$$\begin{aligned} \eta(\theta, \mu) &= (\tau_j^0 + \mu) \begin{pmatrix} \varphi'(x^\star) - my^\star p'(x^\star) - H'(x^\star) & -mp(x^\star) \\ 0 & 0 \end{pmatrix} \delta\big(\theta\big) \\ &+ (\tau_j^0 + \mu) \begin{pmatrix} 0 & 0 \\ y^\star cmp'(x^\star) & 0 \end{pmatrix} \delta\big(\theta + 1\big) \end{aligned} \quad [32]$$

where δ is the Dirac delta function

For $\phi \in \mathcal{C}\big([-1, 0], R^2\big)$, define

$$A(\mu)\phi = \begin{cases} \frac{d\phi(\theta)}{d\theta}, & \theta \in [-1, 0)\,; \\ \int_1^0 d\eta(\mu, s)\phi(s), & \theta = 0. \end{cases} \quad [33]$$

and

$$R(\mu)\phi = \begin{cases} 0, & \theta \in [-1, 0)\,; \\ f(\mu, \phi), & \theta = 0. \end{cases} \quad [34]$$

Then, the system (6) is equivalent to

$$\dot{x}(t) = A(\mu)x_t + R(\mu)x_t \quad [35]$$

where $x_t(\theta) = x(t + \theta)$, $\theta \in [-1, 0]$.

For $\psi \in \mathcal{C}^1\big([0, 1], R^2\big)$, define

$$A^*\psi = \begin{cases} -\frac{d\psi(s)}{ds}, & s \in (0,1]\,; \\ \int_1^0 d\eta(t,0)\phi(-t), & s = 0. \end{cases} \quad [36]$$

and a bilinear inner product

$$\langle \psi(s), \phi(\theta)\rangle = \bar{\psi}(0)\phi(0) - \int_{-1}^{0}\int_{\xi=0}^{\theta} \bar{\psi}(\xi-\theta)d\eta(\theta)\phi(\xi)d\xi \quad [37]$$

where $\eta(\theta) = \eta(\theta, 0)$. In addition, by Theorem 1 we know that $\pm iw_0\tau_j^0$ are eigenvalues of $A(0)$. Thus, they are also eigenvalues of A^*. Let $q(\theta)$ be the eigenvector of $A(0)$ corresponding to $iw_0\tau_j^0$ and $q^*(s)$ be the eigenvector of A^* corresponding to $-iw_0\tau_j^0$.

Let $q(\theta) = \begin{pmatrix} 1, & \nu_1 \end{pmatrix}^T e^{iw_0\tau_j^0\theta}$ and $q^*(s) = D\begin{pmatrix} 1, & \nu_1^* \end{pmatrix}^T e^{iw_0\tau_j^0 s}$. From the above discussion, it is easy to know that $A(0)q(0) = iw_0\tau_j^0 q(0)$ and $A^*(0)q^*(0) = -iw_0\tau_j^0 q^*(0)$. That is

$$\tau_j^0 \begin{pmatrix} \varphi'(x^\star) - my^\star p'(x^\star) - H'(x^\star) & -mp(x^\star) \\ 0 & 0 \end{pmatrix} q(0)$$
$$+\tau_j^0 \begin{pmatrix} 0 & 0 \\ y^\star cmp'(x^\star) & 0 \end{pmatrix} q(-1) = iw_0\tau_j^0 q(0)$$

and

$$\tau_j^0 \begin{pmatrix} \varphi'(x^\star) - my^\star p'(x^\star) - H'(x^\star) & 0 \\ -mp(x^\star) & 0 \end{pmatrix} q^*(0)$$
$$+\tau_j^0 \begin{pmatrix} 0 & y^\star cmp'(x^\star) \\ 0 & 0 \end{pmatrix} q^*(-1) = -iw_0\tau_j^0 q^*(0)$$

Thus, we can easily obtain

$$q(\theta) = \begin{pmatrix} 1\,, & \frac{y^\star cmp'(x^\star)e^{-iw_0\tau_j^0}}{iw_0} \end{pmatrix}^T e^{iw_0\tau_j^0\theta} \quad [38]$$

$$q^*(s) = D\begin{pmatrix} 1\,, & \frac{mp(x^\star)}{iw_0} \end{pmatrix}^T e^{iw_0\tau_j^0 s} \quad [39]$$

In order to assure $\langle \bar{q}^*(s), q(\theta)\rangle = 1$, we need to determine the value of D. From (37), we have

$$\begin{aligned} \langle q^*(s), q(\theta)\rangle &= \bar{q}^*(0)q(0) - \textstyle\int_{-1}^{0}\int_{\xi=0}^{\theta} \bar{q}^*(\xi-\theta)d\eta(\theta)q(\xi)d\xi \\ &= \bar{q}^*(0)q(0) - \textstyle\int_{-1}^{0}\int_{\xi=0}^{\theta} \bar{D}\Big(1,\ \bar{\nu}_1^*\Big)e^{-iw_0\tau_j^0(\xi-\theta)}d\eta(\theta)\Big(1,\ \nu_1\Big)^T e^{iw_0\tau_j^0\xi}d\xi \\ &= \bar{q}^*(0)q(0) - \bar{q}^*(0)\textstyle\int_{-1}^{0}\theta e^{iw_0\tau_j^0\theta}d\eta(\theta)q(0) \\ &= \bar{q}^*(0)q(0) \\ &\quad -\bar{q}^*(0)\tau_j^0 \begin{pmatrix} \varphi'(x^\star) - my^\star p'(x^\star) - H'(x^\star) & -mp(x^\star) \\ 0 & 0 \end{pmatrix} \\ &\quad \Big(-e^{-iw_0\tau_j^0}\Big)q(0) \\ &= \bar{D}\Big[1 + \nu_1\bar{\nu}_1^* + \tau_j^0 e^{-iw_0\tau_j^0\bar{\nu}_1^*}y^\star cmp'(x^\star)\Big] \end{aligned}$$

So, we have

$$\begin{aligned}\bar{D} &= \frac{1}{1+\nu_1\bar{\nu}_1^*+\tau_j^0 e^{-iw_0\tau_j^0\bar{\nu}_1^*}y^\star cmp'(x^\star)}\\ D &= \frac{1}{1+\bar{\nu}_1\nu_1^*+\tau_j^0 e^{iw_0\tau_j^0\nu_1^*}y^\star cmp'(x^\star)}\end{aligned} \quad [40]$$

That ends our proof.

Annex 2 : Proof of the Lemma 2

We have $x_t(\theta) = \big(x_{1t}(\theta), x_{2t}(\theta)\big)$ and $q(\theta) = \big(1,\nu_1\big)^T e^{iw_0\tau_j^0\theta}$. So, from (8) and (10), it follows that

$$\begin{aligned}x_t(\theta) &= W(t,\theta) + 2\mathcal{R}_e\Big(z(t)q(\theta)\Big)\\ &= W_{20}(\theta)\tfrac{z^2}{2} + W_{11}(\theta)z\bar{z} + W_{02}\tfrac{\bar{z}^2}{2} + \big(1,\nu_1\big)^T e^{iw_0\tau_j^0\theta}z(t) + \big(1,\bar{\nu}_1\big)^T e^{-iw_0\tau_j^0\theta}\bar{z}(t) + \cdots\end{aligned} \quad [41]$$

and then we have

$$\begin{aligned}x_{1t}(0) &= z + \bar{z} + W_{20}^{(1)}(0)\tfrac{z^2}{2} + W_{11}^{(1)}(0)z\bar{z} + W_{02}^{(1)}(0)\tfrac{\bar{z}^2}{2} + \cdots\\ x_{2t}(0) &= \nu_1 z + \bar{\nu}_1\bar{z} + W_{20}^{(2)}(0)\tfrac{z^2}{2} + W_{11}^{(2)}(0)z\bar{z} + W_{02}^{(2)}(0)\tfrac{\bar{z}^2}{2} + \cdots\\ x_{1t}(-1) &= ze^{-iw_0\tau_j^0} + \bar{z}e^{iw_0\tau_j^0} + W_{20}^{(1)}(-1)\tfrac{z^2}{2} + W_{11}^{(1)}(-1)z\bar{z} + W_{02}^{(1)}(-1)\tfrac{\bar{z}^2}{2} + \cdots\\ x_{2t}(-1) &= \nu_1 ze^{-iw_0\tau_j^0} + \bar{\nu}_1\bar{z}e^{iw_0\tau_j^0} + W_{20}^{(2)}(-1)\tfrac{z^2}{2} + W_{11}^{(2)}(-1)z\bar{z} + W_{02}^{(2)}(-1)\tfrac{\bar{z}^2}{2} + \cdots\end{aligned} \quad [42]$$

It follows together with (30) that

$$\begin{aligned}g(z,\bar{z}) &= \tfrac{z^2}{2}\Big\{2\tau_j^0\bar{D}\Big[-\Big(\tfrac{r}{K} + mp'(x^\star)\nu_1 + \tfrac{my^\star p''(x^\star)}{2}\Big)\\ &+ \bar{\nu}_1\Big(\tfrac{y^\star cmp''(x^\star)e^{-2iw_0\tau_j^0}}{2} + cmp'(x^\star)\nu_1 e^{-iw_0\tau_j^0}\Big)\Big]\Big\}\\ &+ \tfrac{\bar{z}^2}{2}\Big\{2\tau_j^0\bar{D}\Big[-\Big(\tfrac{r}{K} + mp'(x^\star)\bar{\nu}_1 + \tfrac{my^\star p''(x^\star)}{2}\Big)\\ &+ \bar{\nu}_1\Big(\tfrac{y^\star cmp''(x^\star)e^{2iw_0\tau_j^0}}{2} + cmp'(x^\star)\bar{\nu}_1 e^{iw_0\tau_j^0}\Big)\Big]\Big\}\\ &+ z\bar{z}\Big\{2\tau_j^0\bar{D}\Big[-\Big(\tfrac{r}{K} + mp'(x^\star)\mathcal{R}_e\{\nu_1\} + \tfrac{my^\star p''(x^\star)}{2}\Big)\\ &+ \bar{\nu}_1\Big(\tfrac{y^\star cmp''(x^\star)}{2} + cmp'(x^\star)\mathcal{R}_e\Big\{\nu_1 e^{iw_0\tau_j^0}\Big\}\Big)\Big]\Big\}\\ &+ \tfrac{z^2\bar{z}}{2}\Big\{\tau_j^0\bar{D}\Big[-\tfrac{r}{K}\Big(4W_{11}^{(1)}(0) + 2W_{20}^{(1)}(0)\Big)\end{aligned}$$

$$
\begin{aligned}
- \quad & mp'(x^\star)\Big(2W_{11}^{(2)}(0) + W_{20}^{(2)}(0) + \bar{\nu}_1 W_{20}^{(1)}(0) + 2\nu_1 W_{11}^{(1)}(0)\Big) \\
- \quad & \tfrac{mp''(x^\star)}{2}(2\bar{\nu}_1 + 4\nu_1) - \tfrac{my^\star p''(x^\star)}{2}\Big(4W_{11}^{(1)}(0) + 2W_{20}^{(1)}(0)\Big) \\
+ \quad & \bar{\nu}_1 my^\star p''(x^\star)\Big(2W_{11}^{(1)}(-1) + W_{20}^{(1)}(-1)e^{iw_0\tau_j^0}\Big) \\
+ \quad & \bar{\nu}_1 cmp'(x^\star)\Big(\bar{\nu}_1 W_{20}^{(1)}(-1) + W_{20}^{(2)}(0)e^{iw_0\tau_j^0} \\
& +2W_{11}^{(2)}(0)e^{-iw_0\tau_j^0} + 2\nu_1 W_{11}^{(1)}(-1)\Big) \\
+ \quad & \tfrac{cmp''(x^\star)}{2}\Big(4\nu_1 + 2\bar{\nu}_1 e^{-2iw_0\tau_j^0}\Big)\Big]\Big\}
\end{aligned}
$$

Where f and D are given in the proof of the lemma 1 respectively by (30) and (40).

Comparing the coefficients with (12), we obtain the coefficients of $g(z, \bar{z})$.
That ends our proof.

Optimal Control of Arboviral Diseases

ABBOUBAKAR Hamadjam[⋆,‡] & KAMGANG Jean Claude[†]

[⋆]UIT-Department of Computer science
University of Ngaoundere, Cameroon
abboubakarhamadjam@yahoo.fr
[†]ENSAI-Department of Mathematics and Computer science
University of Ngaoundere, Cameroon
jckamgang@yahoo.fr
[‡] Corresponding author.

ABSTRACT. In this paper, we derive and analyse a model for the control of arboviral diseases which takes into account an imperfect vaccine combined with some other mechanisms of control already studied in the literature. We use f ve time dependent controls, to assess the impact of vaccination combined with treatment, individual protection and vector control strategies such as killing adult vectors, reduction of eggs and larvae. By using optimal control theory, we establish optimal conditions under which the disease can be eradicated and we examine the impact of a possible combined control tools on the disease transmission. The Pontryagin's maximum principle is used to characterize the optimal control. Numerical simulationsshow that, vaccination combined with other control mechanisms, would reduce the spread of the disease appreciably.

RÉSUMÉ. Dans cet article, nous dérivons et analysons un modèle, pour le contrôle des arboviroses, qui prend en compte un vaccin imparfait combiné avec d'autres mécanismes de contrôle déjà étudiés dans la littérature. Nous utilisons cinq contrôles dépendant du temps, pour évaluer l'impact de la vaccination combiné avec le traitement, la protection individuelle et les stratégies de lutte anti-vectoriel telles que l'utilisation des adulticides et des larvicides. En utilisant la théorie du contrôle optimal, nous établissons des conditions optimales dans lesquelles la maladie peut être éradiquée et nous examinons l'impact d'une éventuelle combinaison de contrôle sur la transmission de la maladie. Le principe du maximum de Pontryagin est utilisé pour caractériser le contrôle optimal. Des simulations numériques montrent que la vaccination combinée avec d'autres mécanismes de contrôle, permettrait de réduire de façon considérable la propagation de la maladie.

KEYWORDS : Arboviral diseases; Optimal control; Pontryagin's Maximum Principle.

MOTS-CLÉS : Arboviroses, Contrôle optimal, Maximum de Pontryagin.

1. Introduction

Arboviral diseases are affections transmitted by hematophagous arthropods. There are currently 534 viruses registered in the International Catalog of Arboviruses and 25% of them have caused documented illness in human populations [6, 11]. Examples of those kinds of diseases are Dengue, Yellow fever, Saint Louis fever, Encephalitis, West Nile fever and Chikungunya. A wide range of arboviral diseases are transmitted by mosquito bites and constitute a public health emergency of international concern. For example, Dengue, caused by any of four closely-related virus serotypes (DEN-1-4) of the genus Flavivirus, causes 50–100 million infections worldwide every year, and the majority of patients worldwide are children aged 9 to 16 years [19, 22].

For all the diseases mentioned above, only yellow fever has a licensed vaccine. Nevertheless, considerable efforts are made to obtain vaccines for other diseases. In the case of dengue, for example, tests carried out in Asia and Latin America, have shown that the future dengue vaccine will have a efficacy between 30.2% and 77.7%, and this, depending on the serotype [18, 21]. Also, the future dengue vaccine will have an overall efficacy of 60.8% against all forms of the disease in children and adolescents aged 9-16 years who received three doses of the vaccine[20].

As the future vaccines (e.g., dengue vaccine) will be imperfect, it is therefore necessary to combine such vaccines with some control mechanisms (individual protection, treatment, chemical control) [1, 2, 15], to find the best sufficient combination, which permit to decrease the expansion of these kind of diseases in human communities.

A number of studies have been conducted to study host-vector models for arboviral diseases transmission. Some of these works have been conducted to explore optimal control theory for arboviral disease models (see [3, 4, 7, 14, 17]).

None of the above mentioned models takes into account the combination of optimal control mechanisms such as vaccination, individual protection, treatment and vector control strategies. In our effort, we investigate such optimal strategies for vaccination combined with individual protection, treatment and two vector controls (adulticiding–killing of adult vectors, and larviciding–killing eggs and larvae), using two systems of ODEs which consist of a complete stage structured model Eggs-Larvae-Pupae for the vectors, and a SEI/SEIR type model for the vector/host population. This provides a new different mathematical perspective to the subject.

The rest of the paper is organized as follows. In Section 2 we present the optimal control problem and its mathematical analysis. Section 3 is devoted to numerical simulations. A conclusion round up the paper.

2. A Model for Optimal Control

There are several possible interventions in order to reduce or limit the proliferation of mosquitoes and the explosion of the number of infected humans and mosquitoes. In addition of controls used in [14], we add vaccination and the control of adult vectors as control variables to reduce or even eradicate the disease. So we introduce five time dependent controls:

1) The first control $0 \leq u_1(t) \leq 1$ denotes the percentage of susceptible individuals that one decides to vaccinate at time t. A parameter ω associated to the control $u_1(t)$ represents the waning immunity process [17].

2) The second control $0 \leq u_2(t) \leq 1$ represents efforts made to protect human from mosquito bites. It mainly consists to the use of mosquito nets or wearing appropriate clothes [14]. Thus we modify the infection term as follows:

$$\lambda_h^c = (1 - \alpha_1 u_2(t))\lambda_h, \quad , \lambda_v^c = (1 - \alpha_1 u_2(t))\lambda_v \tag{1}$$

where α_1 measures the effectiveness of the prevention measurements against mosquito bites.

3) The third control $0 \leq u_3(t) \leq 1$ represents efforts made for treatment. It mainly consists in isolating infected patients in hospitals, installing an anti-mosquito electric diffuser in the hospital room, or symptomatic treatments [14]. Thus we modify the recovery rate such that $\sigma_h^c := \sigma_h + \alpha_2 u_3$. α_2 is the effectiveness of the anti-arboviral diseases drugs with $\alpha_2 = 0.3$ [14]. Note that this control also permit to reduce the disease-induced death.

4) The fourth control $0 \leq u_4(t) \leq 1$ represents mosquitoes adulticiding effort with killing efficacy c_m. Thus the mosquito natural mortality rate becomes $\mu_v^c = \mu_v + c_m u_4(t)$.

5) The fifth control $0 \leq u_5(t) \leq 1$ represents the effect of interventions used for the vector control. It mainly consists in the reduction of breeding sites with chemical application methods, for instance using larvicides like BTI (*Bacillus Thuringensis Israelensis*) which is a biological larvicide, or by introducing larvivore fish. This control focuses on the reduction of the number of larvae, and thus eggs, of any natural or artificial water-filled container [14]. Thus the eggs and Larvae natural mortality rate become $\mu_E^c = \mu_E + \eta_1 u_5(t)$ and $\mu_L^c = \mu_L + \eta_2 u_5(t)$ where η_1, η_2, represent the chemical eggs and larvae mortality rate, respectively [14].

Note that $0 \leq u_i \leq 1$, for $i = 1, \ldots, 5$, means that when the control is zero there is no any effort invested (i.e. no control) and when it is one, the maximum control effort is invested.

Therefore, our optimal control model of arboviral diseases reads as

$$\begin{cases} \dot{S}_h &= \Lambda_h - \left[(1 - \alpha_1 u_2(t))\lambda_h + \mu_h + u_1(t)\right] S_h + \omega u_1(t) R_h \\ \dot{E}_h &= (1 - \alpha_1 u_2(t))\lambda_h S_h - (\mu_h + \gamma_h) E_h \\ \dot{I}_h &= \gamma_h E_h - \left[\mu_h + (1 - \alpha_2 u_3(t))\delta + \sigma + \alpha_2 u_3(t)\right] I_h \\ \dot{R}_h &= (\sigma + \alpha_2 u_3(t)) I_h + u_1 S_h - (\mu_h + \omega u_1) R_h \\ \dot{S}_v &= \theta P - (1 - \alpha_1 u_2(t))\lambda_v S_v - (\mu_v + c_m u_4(t)) S_v \\ \dot{E}_v &= (1 - \alpha_1 u_2(t))\lambda_v S_v - (\mu_v + \gamma_v + c_m u_4(t)) E_v \\ \dot{I}_v &= \gamma_v E_v - (\mu_v + c_m u_4(t)) I_v \\ \dot{E} &= \mu_b \left(1 - \dfrac{E}{\Gamma_E}\right)(S_v + E_v + I_v) - (s + \mu_E + \eta_1 u_5(t)) E \\ \dot{L} &= sE \left(1 - \dfrac{L}{\Gamma_L}\right) - (l + \mu_L + \eta_2 u_5(t)) L \\ \dot{P} &= lL - (\theta + \mu_P) P \end{cases} \tag{2}$$

with initial conditions given at $t = 0$.

The states variables and parameters of model (2) are described in Table 1 and 2.

For the non-autonomous system (2), the rate of change of the total populations of humans and adults vectors is given, respectively, by

$$\begin{cases} \dot{N}_h &= \Lambda_h - \mu_h N_h - (1 - \alpha_2 u_3(t))\delta I_h \\ \dot{N}_v &= \theta P - (\mu_v + c_m u_4(t)) N_v \end{cases} \tag{3}$$

For bounded Lebesgue measurable controls and non-negative initial conditions, non-negative bounded solutions to the state system exist [12].

Table 1: The state variables of model (2).

	Humans		Aquatic Vectors		Adult Vectors
S_h:	Susceptible	E:	Eggs	S_v:	Susceptible
E_h:	Infected in latent stage	L:	Larvae	E_v	Infected in latent stage
I_h:	Infectious	P:	Pupae	I_v	Infectious
R_h:	Resistant (immune)				

Table 2: Description and baseline values/range of parameters of model 2. The baseline values refer to dengue fever transmission.

Parameter	Description	Baseline value/range	Sources
Λ_h	Recruitment rate of humans	2.5 day^{-1}	[10]
μ_h	Natural mortality rate in humans	$\frac{1}{(67\times365)}$ day^{-1}	[10]
a	Average number of bites	1 day^{-1}	[3, 10]
β_{hv}	Probability of transmission of infection from an infected vector to a susceptible human	0.1, 0.75 day^{-1}	[3, 10]
γ_h	Progression rate from E_h to I_h	$\left[\frac{1}{15}, \frac{1}{3}\right]$ day^{-1}	[8]
δ	Disease–induced death rate	10^{-3} day^{-1}	[10]
σ	Recovery rate for humans	0.1428 day^{-1}	[3, 10]
η_h,η_v	Modifications parameter	$[0, 1)$	[10]
μ_v	Natural mortality rate of vectors	$\left[\frac{1}{30}, \frac{1}{14}\right]$ day^{-1}	[3, 10]
γ_v	Progression rate from E_v to I_v	$\left[\frac{1}{21}, \frac{1}{2}\right]$ day^{-1}	[8]
β_{vh}	Probability of transmission of infection from an infected human to a susceptible vector	0.1, 0.75 day^{-1}	[3, 10]
θ	Maturation rate from pupae to adult	0.08 day^{-1}	[8, 14]
μ_b	Number of eggs at each deposit	6 day^{-1}	[8]
Γ_E	Carrying capacity for eggs	$10^3, 10^6$	[3]
Γ_L	Carrying capacity for larvae	$5\times10^2, 5\times10^5$	[3]
μ_E	Eggs death rate	0.2 or 0.4	[14]
μ_L	Larvae death rate	0.2 or 0.4	[14]
μ_P	Pupae death rate	0.4	Assumed
s	Transfer rate from eggs to larvae	0.7 day^{-1}	[14]
l	Transfer rate from larvae to pupae	0.5 day^{-1}	[13]

The objective of control is to minimize: the number of symptomatic humans infected with arboviruses (that is, to reduce sub-population I_h), the number of vector (N_v) and the number of eggs and larvae (that is, to reduce sub-population E and L, respectively), while keeping the costs of the control as low as possible.

To achieve this objective we must incorporate the relative costs associated with each policy (control) or combination of policies directed towards controlling the spread of arboviral diseases. We define the objective function as

$$J(u_1,u_2,u_3,u_4,u_5)=\int_0^{t_f}\left[D_1I_h(t)+D_2N_v(t)+D_3E(t)+D_4L(t)+\sum_{i=1}^{5}B_iu_i^2(t)\right]dt \tag{4}$$

and the control set

$$\Delta=\{(u_1,u_2,u_3,u_4,u_5)|u_i(t)\text{ is Lebesgue measurable on }[0,t_f],\ 0\le u_i(t)\le 1, i=1,\dots,5\}.$$

The first fourth terms in the integrand J represent benefit of I_h, N_v, E and L populations, describing the comparative importance of the terms in the functional. A high value of D_1 for example, means that it is more important to reduce the burden of disease as reduce the costs related to all control strategies [5]. Positive constants B_i, $i=1,\dots,5$ are weight for vaccination, individual protection (human), treatment and vector control effort respectively, which regularize the optimal control. In line with the authors of some studies on the optimal control (see [7, 14, 17]), we choose a linear function for the cost on infection, D_1I_h, D_2N_v, D_3E, D_4L, and quadratic forms for the cost on the controls $B_1u_1^2$, $B_2u_2^2$, $B_3u_3^2$, $B_4u_4^2$, and $B_5u_5^2$. This choice can be justified by the following arguments:

1) An epidemiological control can be likened to an expenditure of energy, by bringing to the applications of physics in control theory;

2) In a certain sense, minimize u_i is like minimize u_i^2, because $u_i\ge 0$, $i=1,\dots,5$.

3) Among the nonlinear representation of intervention costs, the quadratic approximation is the simplest and most widely used, contrary to the linear controls that usually lead to the bang-bang controls.

We solve the problem using optimal control theory.

Theorem 1. *Let $X=(S_h,E_h,I_h,R_h,S_v,E_v,I_v,E,L,P)$. The following set*

$$\Omega=\left\{X\in\mathbf{R}^{10}:N_h\le\frac{\Lambda_h}{\mu_h};E\le\Gamma_E;L\le\Gamma_L;P\le\frac{l\Gamma_L}{k_7};N_v\le\frac{\theta l\Gamma_L}{k_7k_8}\right\}$$

is positively invariant under system (2).

Proof. On the one hand, one can easily see that it is possible to get,

$$\begin{cases}\dot{S}_h & \ge-(\lambda_h+\mu_h)S_h,\ \dot{E}_h\ge-(\mu_h+\gamma_h)E_h,\ \dot{I}_h\ge-(\mu_h+\delta+\sigma)I_h,\ \dot{R}_h\ge-\mu_hR_h\\ \dot{E} & \ge-(\frac{\mu_b}{K_E}+s+\mu_E+\eta_1)E,\ \dot{L}\ge-(\frac{s}{K_L}+l+\mu_L+\eta_2)L,\ \dot{P}\ge-(\theta+\mu_P+\eta_3)P\\ \dot{S}_v & \ge-(\lambda_v+\mu_v)S_v,\ \dot{E}_v\ge-(\mu_v+\gamma_v)E_v,\ \dot{I}_v\ge-\mu_vI_v,\end{cases} \tag{5}$$

for $(S_h(0),E_h(0),I_h(0),R_h(0),E(0),A(0),P(0),S_v(0),E_v(0),I_v(0))\ge 0$. Thus, solutions with initial value in Ω remain nonnegative for all $t\ge 0$. On the other hand, we have

$$\begin{cases} \dot{N}_h & \leq \Lambda_h - \mu_h N_h \\ \dot{N}_v & \leq \theta P - \mu_v N_v \\ \dot{E} & \leq \mu_b \left(1 - \dfrac{E}{K_E}\right)(S_v + E_v + I_v) - (s + \mu_E)E \\ \dot{L} & \leq sE\left(1 - \dfrac{L}{K_L}\right) - (l + \mu_L)L \\ \dot{P} & \leq lL - (\theta + \mu_P)P \end{cases} \tag{6}$$

The right hand side of the inequalities correspond to the transmission model without control, and it is easy to show that solutions remain in Ω. Then using Gronwall's inequality, we deduce that solutions of (2) are bounded. □

2.1. Existence of an optimal control

The existence of an optimal control can be obtained by using a result of Fleming and Rishel [9].

Theorem 2. *Consider the control problem with system* (2)*.*
There exists $u^\star = (u_1^\star, u_2^\star, u_3^\star, u_4^\star, u_5^\star)$ *such that*

$$\min_{(u_1,u_2,u_3,u_4,u_5)\in\Delta} J(u_1, u_2, u_3, u_4, u_5) = J(u_1^\star, u_2^\star, u_3^\star, u_4^\star, u_5^\star)$$

Proof. To use an existence result, Theorem III.4.1 from [9], we must check if the following properties are satisfied:

1) the set of controls and corresponding state variables is non empty;

2) the control set Δ is convex and closed;

3) the right hand side of the state system is bounded by a linear function in the state and control;

4) the integrand of the objective functional is convex;

5) there exist constants $c_1 > 0$, $c_2 > 0$, and $\beta > 1$ such that the integrand of the objective functional is bounded below by $c_1 \left(\sum_{i=1}^{5} |u_i|^2\right)^{\frac{\beta}{2}} - c_2$.

In order to verify these properties, we use a result from Lukes [12] to give the existence of solutions for the state system (2) with bounded coefficients, which gives condition 1. Since by definition, the control set Δ is bounded , then condition 2 is satisfied. The right hand side of the state system (2) satisfies condition 3 since the state solutions are bounded. The integrand of our objective functional is clearly convex on Δ, which gives condition 4. There are $c_1 > 0$, $c_2 > 0$ and $\beta > 1$ satisfying $D_1 I_h + D_2 N_v + D_3 E + D_4 L + \sum_{i=1}^{5} B_i u_i^2 \geq c_1 \left(\sum_{i=1}^{5} |u_i|^2\right)^{\frac{\beta}{2}} - c_2$, because the states variables are bounded. Thus condition 5 is satisfied. We conclude that there exists an optimal control $u^* = (u_1^\star, u_2^\star, u_3^\star, u_4^\star, u_5^\star)$ that minimizes the objective functional $J(u_1, u_2, u_3, u_4, u_5)$. □

2.2. Characterization of an optimal control

The necessary conditions that an optimal control must satisfy come from the Pontryagin's Maximum Principle (PMP) [16]. This principle converts (2)-(4) into a problem of minimizing point wise a Hamiltonian $\mathbb{H}$, with respect to $(u_1, u_2, u_3, u_4, u_5)$:

$$\begin{aligned}
\mathbb{H} &= D_1 I_h + D_2 N_v + D_3 E + D_4 L + \sum_{i=1}^{5} B_i u_i^2 \\
&+ \lambda_{S_h} \left\{ \Lambda_h - \left[(1-\alpha_1 u_2)\lambda_h + \mu_h + u_1 \right] S_h + \omega u_1 R_h \right\} \\
&+ \lambda_{E_h} \left\{ \left[1 - \alpha_1 u_2 \right] \lambda_h S_h - (\mu_h + \gamma_h) E_h \right\} \\
&+ \lambda_{I_h} \left\{ \gamma_h E_h - (\mu_h + (1-\alpha_2 u_3)\delta + \sigma + \alpha_2 u_3) I_h \right\} \\
&+ \lambda_{R_h} \left\{ (\sigma + \alpha_2 u_3) I_h + u_1 S_h - (\mu_h + \omega u_1) R_h \right\} \\
&+ \lambda_{S_v} \left\{ \theta P - \left[1 - \alpha_1 u_2 \right] \lambda_v S_v - (\mu_v + c_m u_4) S_v \right\} \\
&+ \lambda_{E_v} \left\{ (1-\alpha_1 u_2)\, \lambda_v S_v - (\mu_v + \gamma_v + c_m u_4) E_v \right\} + \lambda_{I_v} \left\{ \gamma_v E_v - (\mu_v + c_m u_4) I_v \right\} \\
&+ \lambda_E \left\{ \mu_b \left(1 - \frac{E}{\Gamma_E} \right) (S_v + E_v + I_v) - (s + \mu_E + \eta_1 u_5) E \right\} \\
&+ \lambda_L \left\{ sE \left(1 - \frac{L}{\Gamma_L} \right) - (l + \mu_L + \eta_2 u_5) L \right\} \\
&+ \lambda_P \left\{ lL - (\theta + \mu_P) P \right\}
\end{aligned} \tag{7}$$

where the λ_i, $i = S_h, E_h, I_h, R_h, S_v, E_v, I_v, E, L, P$ are the adjoint variables or co-state variables. Applying Pontryagin's Maximum Principle [16], we obtain the following result.

Theorem 3. *Given an optimal control $u^\star = (u_1^\star, u_2^\star, u_3^\star, u_4^\star, u_5^\star)$ and solutions $(S_h^\star, E_h^\star, I_h^\star, R_h^\star, S_v^\star, E_v^\star, I_v^\star, E^\star, A^\star, P^\star)$ of the corresponding state system (2), there exist adjoint variables $\Pi = (\lambda_{S_h}, \lambda_{E_h}, \lambda_{I_h}, \lambda_{R_h}, \lambda_{S_v}, \lambda_{E_v}, \lambda_{I_v}, \lambda_E, \lambda_L, \lambda_P)$ satisfying,*

$$\begin{aligned}
\frac{d\lambda_{S_h}}{dt} &= \mu_h \lambda_{S_h} + u_1(\lambda_{S_h} - \lambda_{R_h}) + (1-\alpha_1 u_2)\lambda_h \left(1 - \frac{S_h}{N_h} \right) (\lambda_{S_h} - \lambda_{E_h}) \\
&\quad + (1-\alpha_1 u_2) \frac{S_v \lambda_v}{N_h} (\lambda_{E_v} - \lambda_{S_v})
\end{aligned} \tag{8}$$

$$\begin{aligned}
\frac{d\lambda_{E_h}}{dt} &= \mu_h \lambda_{E_h} + \gamma_h (\lambda_{E_h} - \lambda_{I_h}) + (1-\alpha_1 u_2) \frac{S_h \lambda_h}{N_h} (\lambda_{E_h} - \lambda_{S_h}) \\
&\quad + (1-\alpha_1 u_2) \frac{S_v}{N_h} (a\beta_{vh}\eta_h - \lambda_v) (\lambda_{S_v} - \lambda_{E_v})
\end{aligned} \tag{9}$$

$$\begin{aligned}
\frac{d\lambda_{I_h}}{dt} &= -D_1 + \left[\mu_h + (1-\alpha_2 u_3)\delta \right] \lambda_{I_h} + (\sigma + \alpha_2 u_3)(\lambda_{I_h} - \lambda_{R_h}) \\
&\quad + (1-\alpha_1 u_2) \frac{S_h \lambda_h}{N_h} (\lambda_{E_h} - \lambda_{S_h}) + (1-\alpha_1 u_2) \frac{S_v}{N_h} (a\beta_{vh} - \lambda_v) (\lambda_{S_v} - \lambda_{E_v})
\end{aligned} \tag{10}$$

$$\frac{d\lambda_{R_h}}{dt} = \mu_h\lambda_{R_h} + \omega u_1(\lambda_{R_h} - \lambda_{S_h}) + (1-\alpha_1 u_2)\frac{S_h\lambda_h}{N_h}(\lambda_{E_h} - \lambda_{S_h}) + (1-\alpha_1 u_2)\frac{S_v\lambda_v}{N_h}(\lambda_{E_v} - \lambda_{S_v}) \tag{11}$$

$$\frac{d\lambda_{S_v}}{dt} = -D_2 + (\mu_v + c_m u_4)\lambda_{S_v} + (1-\alpha_1 u_2)\lambda_v(\lambda_{S_v} - \lambda_{E_v}) - \mu_b\left(1 - \frac{E}{\Gamma_E}\right)\lambda_E \tag{12}$$

$$\frac{d\lambda_{E_v}}{dt} = -D_2 + (\mu_v + c_m u_4)\lambda_{E_v} + \gamma_v(\lambda_{E_v} - \lambda_{I_v}) + a\eta_v\beta_{hv}(1-\alpha_1 u_2)(\lambda_{S_h} - \lambda_{E_h})\frac{S_h}{N_h} - \mu_b\left(1 - \frac{E}{\Gamma_E}\right)\lambda_E \tag{13}$$

$$\frac{d\lambda_{I_v}}{dt} = -D_2 + (\mu_v + c_m u_4)\lambda_{I_v} + a\beta_{hv}(1-\alpha_1 u_2)\frac{S_h}{N_h}(\lambda_{S_h} - \lambda_{E_h}) - \mu_b\left(1 - \frac{E}{\Gamma_E}\right)\lambda_E \tag{14}$$

$$\frac{d\lambda_E}{dt} = -D_3 + \left[\frac{\mu_b}{\Gamma_E}N_v + s + \mu_E + \eta_1 u_5\right]\lambda_E - s\left(1 - \frac{L}{\Gamma_L}\right)\lambda_L \tag{15}$$

$$\frac{d\lambda_L}{dt} = -D_4 - l\lambda_P + \left[\frac{s}{\Gamma_L}E + \mu_L + l + \eta_2 u_5\right]\lambda_L \tag{16}$$

$$\frac{d\lambda_P}{dt} = (\mu_P + \theta)\lambda_P - \theta\lambda_{S_v} \tag{17}$$

and the transversality conditions

$$\lambda_i^*(t_f) = 0, \qquad i = 1, \ldots 10. \tag{18}$$

Furthermore,

$$\begin{aligned}
u_1^\star &= \min\left\{1, \max\left(0, \frac{(S_h - \omega R_h)(\lambda_{S_h} - \lambda_{R_h})}{2B_1}\right)\right\},\\
u_2^\star &= \min\left\{1, \max\left(0, \frac{\alpha_1\left[\lambda_h S_h(\lambda_{E_h} - \lambda_{S_h}) + \lambda_v S_v(\lambda_{E_v} - \lambda_{S_v})\right]}{2B_2}\right)\right\},\\
u_3^\star &= \min\left\{1, \max\left(0, \frac{\alpha_2\left[(1-\delta)\lambda_{I_h} - \lambda_{R_h}\right]I_h}{2B_3}\right)\right\},\\
u_4^\star &= \min\left\{1, \max\left(0, \frac{c_m\left[S_v\lambda_{S_v} + E_v\lambda_{E_v} + I_v\lambda_{I_v}\right]}{2B_4}\right)\right\},\\
u_5^\star &= \min\left\{1, \max\left(0, \frac{\eta_1 E\lambda_E + \eta_2 L\lambda_L}{2B_5}\right)\right\}.
\end{aligned} \tag{19}$$

Proof. The differential equations governing the adjoint variables are obtained by differentiation of the Hamiltonian function, evaluated at the optimal control. Then the adjoint system can be written as

$$\frac{d\lambda_{S_h}}{dt} = -\frac{\partial \mathbb{H}}{\partial S_h}, \ \frac{d\lambda_{E_h}}{dt} = -\frac{\partial \mathbb{H}}{\partial E_h}, \ \frac{d\lambda_{I_h}}{dt} = -\frac{\partial \mathbb{H}}{\partial I_h}, \ \frac{d\lambda_{R_h}}{dt} = -\frac{\partial \mathbb{H}}{\partial R_h}, \ \frac{d\lambda_{S_v}}{dt} = -\frac{\partial \mathbb{H}}{\partial S_v}$$
$$\frac{d\lambda_{E_v}}{dt} = -\frac{\partial \mathbb{H}}{\partial E_v},, \ \frac{d\lambda_{I_v}}{dt} = -\frac{\partial \mathbb{H}}{\partial I_v}, \ \frac{d\lambda_E}{dt} = -\frac{\partial \mathbb{H}}{\partial E}, \ \frac{d\lambda_L}{dt} = -\frac{\partial \mathbb{H}}{\partial L}, \ \frac{d\lambda_P}{dt} = -\frac{\partial \mathbb{H}}{\partial P},$$

with zero final time conditions (transversality).

To get the characterization of the optimal control given by (19), we follow [14, 17] and solve the equations on the interior of the control set,

$$\frac{\partial \mathbb{H}}{\partial u_i} = 0, \ i = 1, \ldots, 5.$$

Using the bounds on the controls, we obtain the desired characterization. This ends the proof. □

3. Numerical simulations and discussion

The simulations were carried out using the values of Table 3. We use an iterative scheme to solve the optimality system.

The optimality system for our problem is derived (see Appendix) and numerically solved by using the so called forward–backward sweep method (FBSM).The process begin with an initial guess on the control variable. Then, the state equations are solved simultaneously forward in time, and next the adjoint equations (8)– (17) are simultaneously solved backward in time. The control is updated by inserting the new values of states and adjoints into its characterization, and the process is repeated until convergence occurs (see e.g. [5, 14]).

The values chosen for the weights in the objective functional J (see Eq. (4)) are given in Table 4. Table 5 gives the initial conditions of state variables. We simulated the system (2) in a period of twenty days ($t_f = 20$).

Table 3: Value of parameters used in numerical simulations.

Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
μ_v	$\frac{1}{30}$	l	0.5	α_2	0.5	γ_h	$\frac{1}{14}$
a	1	μ_E	0.2	μ_h	$\frac{1}{67*365}$	γ_v	$\frac{1}{21}$
Λ_h	2.5	μ_b	6	θ	0.08	μ_P	0.4
β_{hv}	0.75			σ	0.1428	η_v	0.35
β_{vh}	0.75	ω	0.05	μ_L	0.4	δ	10^{-3}
Γ_E	10000	s	0.7	η_1	0.001	η_2	0.3
Γ_L	5000	η_h	0.35	c_m	0.2	α_1	0.5

3.1. Vaccination combined with individual protection, adulticide and larvicide

With this strategy, only the combination of the control u_1 on vaccination, the control u_2 on individual protection, the control u_4 on adulticide and the control u_5 on larvicide, is used to minimise the objective function J (4), while the other control u_3 are set to zero.

Table 4: Numerical values for the cost functional parameters.

Parameters	Value	Source	Parameters	Value	Source
D_1:	10,000	[14]	B_1	10	Assumed
D_2:	10,000	[14]	B_2:	10	[14]
D_3:	5000	Assumed	B_3:	10	[14]
D_4:	1	[14]	B_4:	10	Assumed
			B_5	10	[14]

Table 5: Initial conditions.

Human states	Initial value	Adult Vector states	Initial value	Aquatic states	Initial value
S_{h_0}:	700	S_{v_0}	3000	E_0	10000
E_{h_0}:	220	E_{v_0}	400	L_0	5000
I_{h_0}:	100	I_{v_0}	120	P_0	3000
R_{h_0}:	60				

On figure 1, we observed that the control strategy resulted in a decrease in the number of infected humans (I_h) while an increase is observed in the number of infected humans (I_h) in strategy without control. The use of this combination have a great impact on the decreasing total vector population (N_v), as well as aquatic vector populations (E and L).

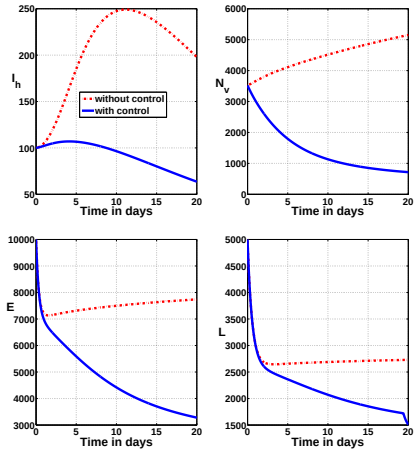


Figure 1: Simulation results of optimal control model (2) showing the effect of using optimal vaccination combined with individual protection, adulticide and larvicide ($u_1 \neq 0$, $u_2 \neq 0$, $u_4 \neq 0$, $u_5 \neq 0$).

3.2. The combination of all the five controls

In this strategy, the combination of all the five controls is applied. On figure 2, we observed that combining all the five controls gives a better result in a decrease in the

number of infected humans (I_h), as well as, the total number of vector population (N_v), and the aquatic vector populations (E and L).

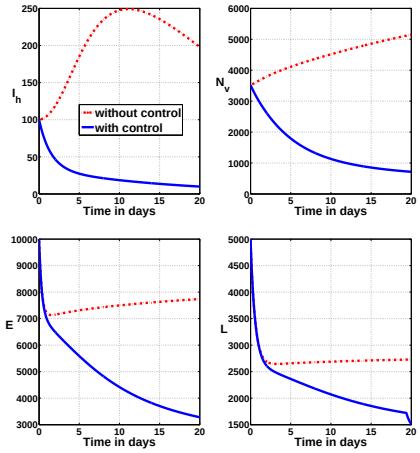


Figure 2: Simulation results of optimal control model (2) showing the effect of using the combination of all the five controls ($u_i \neq 0$, $i = 1, \ldots, 5$).

4. Conclusion

In this paper, we derived and analysed a model for the control of arboviral diseases with non linear form of infection and complete stage structured model for vectors, and which takes into account a vaccination with waning immunity, treatment, individual protection and vector control strategies (adult vectors, eggs and larvae reduction strategies).

We performed optimal control analysis of the model. In this light, we addressed the optimal control by deriving and analysing the conditions for optimal eradication of the disease and in a situation where eradication is impossible or of less benefit compared with the cost of intervention, we also derived and analysed the necessary conditions for optimal control of the disease.

From the numerical results, we concluded that the optimal strategy to effectively control arboviral diseases is the combination of vaccination, individual protection, (with or without treatment), and other mechanisms of vector control (by chemical intervention). However this conclusion must be taken with caution because of the uncertainties around the parameter values and to the budget/resource limitation.

5. References

[1] H. ABBOUBAKAR, J. C. KAMGANG, D. TIEUDJO, "Backward bifurcation and control in transmission dynamics of arboviral diseases", To appear in Mathematical Biosciences, Doi: 10.1016/j.mbs.2016.06.002.

[2] H. ABBOUBAKAR, J.C. KAMGANG, L.N. NKAMBA, D. TIEUDJO, L. EMINI, "Modeling the dynamics of arboviral diseases with vaccination perspective", *Biomath*, vol. 4, 2015, pp. 1–30.

[3] D. ALDILA, T. GÖTZ, E. SOEWONO, "An optimal control problem arising from a dengue disease transmission model", *Mathematical Biosciences* vol. 242, 2013, pp. 9–16.

[4] K. W. BLAYNEHA, A. B. GUMEL, S. LENHART, T. CLAYTON, "Backward bifurcation and optimal control in transmission dynamics of west nile virus", *Bulletin of Mathematical Biology*, vol. 72, 2010, pp. 1006–1028. doi:10.1007/s11538-009-9480-0.

[5] B. BUONOMO, "A simple analysis of vaccination strategies for rubella", *Mathematical Biosciences and Engineering* vol. 8, num. 3, 2011, pp. 677–687.

[6] A. CHIPPAUX, "Généralités sur arbovirus et arboviroses–overview of arbovirus and arbovirosis", *Med. Maladies Infect.*, vol. 33, 2003, pp. 377–384.

[7] W. O. DIAS, E. F. WANNER, R.T.N. CARDOSO, "A multiobjective optimization approach for combating aedes aegypti using chemical and biological alternated step-size control", *Mathematical Biosciences* vol. 269, 2015, pp. 37–47.

[8] Y. DUMONT, F. CHIROLEU, "Vector control for the chikungunya disease", *Math. Biosci. Eng.*, vol. 7, 2010, pp. 313–345.

[9] W. H. FLEMING, R. W. RISHEL, "Deterministic and Stochastic Optimal Control", *Springer Verlag*, 1975.

[10] S. M. GARBA, A. B. GUMEL, M. R. A. BAKAR, "Backward bifurcations in dengue transmission dynamics", *Math. Biosci.*, vol. 215, 2008 pp. 11–25.

[11] N. KARABATSOS, "International Catalogue of Arboviruses, including certain other viruses of vertebrates", *American Society of Tropical Medicine and Hygiene, San Antonio, TX.*, 1985, 2001 update.

[12] D. L. LUKES, "Differential equations: classical to controlled", *Academic Press, New York*, 1982.

[13] D. MOULAY, M. A. AZIZ-ALAOUI, M. CADIVEL, "The chikungunya disease: Modeling, vector and transmission global dynamics", *Math. Biosci.* vol. 229, 2011, pp. 50–63.

[14] D. MOULAY, M. A. AZIZ-ALAOUI, K. HEE-DAE, "Optimal control of chikungunya disease: larvae reduction,treatment and prevention", *Mathemtical Biosciences and Engineering* vol. 9, num. 2, April 2012, pp. 369–393.

[15] H. NISHIURA, "Mathematical and statistical analyses of the spread of dengue", *Dengue Bulletin*, vol. 30, 2006, pp. 51–67.

[16] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, E. F. MISHCHENKO, "The mathematical theory of optimal processes", *Wiley, New York*, 1962.

[17] H. S. RODRIGUES, M. T. T. MONTEIRO, D. F. M. TORRES, "Vaccination models and optimal control strategies to dengue", *Mathematical Biosciences*, vol. 247, 2014, pp. 1–12.

[18] A. SABCHAREON, D. WALLACE, C. SIRIVICHAYAKUL, K. LIMKITTIKUL ET AL., "Protective efficacy of the recombinant, live-attenuated, cyd tetravalent dengue vaccine in thai schoolchildren: a randomised, controlled phase 2b trial", *Lancet*, vol. 380, 2012, pp. 1559–1567.

[19] SANOFI PASTEUR, "Dengue vaccine, a priority for global health", 2013.

[20] SANOFI PASTEUR, "Communiqué de presse: The new england journal of medicine publie les résultats de l'étude clinique d'efficacité de phase 3 du candidat vaccin dengue de sanofi pasteur", 2014.

[21] L. VILLAR, G. H. DAYAN, J. L. ARREDONDO-GARCIA ET AL., "Efficacy of a tetravalent dengue vaccine in children in latin america", *The New England Journal of Medicine* vol. 372, num. (2), 2015, pp. 113–123.

[22] WORLD HEALTH ORGANIZATION, "Dengue and dengue haemorhagic fever", www.who.int/mediacentre/factsheets/fs117/en, 2009.

Rubrique

Identification of self-heating effects on the behaviour of HEMA-EGDMA hydrogels biomaterials using non-linear thermo-mechanical modeling

N. Santatriniaina†, M. Nassajian Moghadam††, D. Pioletti††, L. Rakotomanana†

† Mathematical Research Institute of Rennes
University of Rennes, France.
†† Laboratory of Biomechanical Orthopedics Lausanne
Federal Polytechnic School of Lausanne, Switzerland. nirinasantatriniaina@gmail.com

RÉSUMÉ. Ce papier est dédié à la quantification de la production de chaleur dans l'hydrogel de type HEMA-EGDMA sous chargement dynamique. On s'intéresse à la modélisation du phénomène de self-heating dans les polymères, les hydrogels et les tissus biologiques. On compare les résultats théoriques avec les résultats expérimentaux combinés avec une proposition d'optimisation numérique pour identifier les paramètres influençant le phénomène de self-heating. D'abord, nous nous sommes focalisés sur la modélisation de la loi constitutive de l'hydrogel de type HEMA-EGDMA. Nous avons utilisé la théorie des invariants polynomiaux pour définir la loi constitutive du matériau. Ensuite, nous avons mis en place un modèle théorique en thermomécanique couplée d'un milieu continu classique pour analyser la production de chaleur dans ce matériau. Deux potentiels thermodynamiques ont été proposés et identifiés avec les mesures expérimentales. Une nouvelle forme d'équation du mouvement non-linéaire et couplée a été obtenue. Enfin, une méthode numérique des équations thermo-mécaniques pour les modèles a été utilisée. Cette étape nous a permis, entre autres, de résoudre ce système couplé. La méthode numérique est basée sur la méthode des éléments finis.

ABSTRACT. This paper is dedicated to the quantification of the heat production in the HEMA-EGDMA hydrogel under dynamic loading. We focus on modeling of the self-heating phenomenon in polymers, hydrogels and biological tissues. We compare the theoretical and experimental results combined with numerical optimization proposal to identify the influencing parameters on the self-heating phenomenon. We develop constitutive law of the HEMA-EGDMA hydrogel, focusing on the heat effects in this material. We set up a theoretical model of coupled thermo-mechanical classic continuum for a better understanding of the heat production in this media. We use polynomial invariants theory to define the constitutive law of the media. Two thermodynamic potentials are proposed and are identified with the experimental measurements. New form of non-linear and coupled governing equations were obtained. Numerical methods were used to solve thermo-mechanical formalism for the model. Then, this step allows us, among other things, to propose an appropriate numerical methods to solve this system. The numerical methods is based on the finite element methods.

MOTS-CLÉS : Hydrogel, self-heating, thermomécanique, méthodes numériques, EDPs.

KEYWORDS : Hydrogel, self-heating, thermomechanics, numerical methods, PDEs.

1. Introduction

Hydrogels have been widely employed in biomedical areas [1], [2], [3]. The thermomechanical response of these materials depends strongly on temperature, cross-link density and frequency if the hydrogel is under cyclic loading [4]. Particular hydrogel possessing high dissipation properties may induce a heat production under cyclic loading [5]. Due to the heat production, an increase of the local temperature can be observed in the material, a phenomenon also known as self-heating [4], [5]. In turn, the increase in temperature has an effect on its properties and on the thermomechanical behavior [5], [6], [7]. Modeling and simulation methods are one of the strong characterization methods of the physical phenomena in this kind of material [8], [9]. When the sample is simultaneously subjected to mechanical and thermal loads, we need to develop experimental tool and coupled formulation to investigate and to measure simultaneously the mechanical response and the heat production in the sample [9], [10], [11]. The goal of this work is to identify a constitutive law based on generalized standard materials in correlation with the experimental measurements. Numerical methods for a coupled partial differential equation with dynamic boundary conditions are developed with the conservation laws [12], [13], [14]. Nonlinear constitutive law for viscoelastic material without heat effect has been established by Pioletti, Rakotomanana *et al.* for biological tissues in large deformation [9]. The present work extends this model to nonlinear constitutive law for thermo-viscoelastic model with heat effect in the particular case of matrix HEMA-EGDMA hydrogel. In this work, a general continuum thermomechanical framework describing the effect is adapted to the description of the self-heating phenomenon. Numerical studies are then carried out to examine the ability of the model to predict the heat production and to define the nature of the coupling as well as to evaluate the influence of the main parameters such as cross-link density and frequency of loading. In parallel, microcalorimetric experimental measurements are performed to quantify the heat production and the mechanical response in the HEMA-EGDMA hydrogel sample.

2. Microcalorimetric test

In order to characterize the heat production in the hydrogel samples, an adiabatic deformation microcalorimeter is used [4]. The hydrogel sample consists of cylindrical samples 5 mm of diameter and 8 mm of height are subjected to cyclic mechanical load at various frequencies $f = 0.5, ..., 1Hz$. For the mechanical boundary conditions, on the top of the cylinder we apply the cyclic load, while the bottom is fixed. For the thermal boundary condition, we have an adiabatic condition (non inward and outward flux). The initial conditions are : initial stress null and initial temperature θ_0. The heat production is measured with a specific sensor inserted within the sample and the data acquisition is directly obtained with a computer. For a more detailed description, the reader is reffered to [4]. The displacement is prescibed on the top of the sample to 20% of the sample height. The sample loading is done in three parts including preload, cyclic loading and relaxation. And the bottom of the sample is "fixed". We chose 30 s of preload, 5 mn cyclic loading and 5 mn relaxation. For the sample we use the composition is given by : HEMA+40%w+ϕ% EGDMA with 8.93 mm diameter, 5.33 mm of height, 40% of water, 6% and 8% of crosslink density.

3. Mathematical settings

The self-heating phenomena are governed by a nonlinear-coupled partial differential equation system deduced from two conservation equations of classical continuum thermomechanics. We assume the postulate of the existence of two thermodynamic potentials the strain energy function and the dissipation potential defined per unit of the reference volume. The model is obtained by constructing with the free energy method, new non-negative convex energy functions given by the equation (1). For physical and mathematical considerations, convexity/polyconvexity of the strain energy and dissipation functions are an essential point since the common methods in computer simulation depend on gradient methods.

$$\psi(\mathbf{E},\theta) = \frac{\lambda}{2}\mathrm{tr}^2\mathbf{E} + \mu\mathrm{tr}\mathbf{E}^2 - (3\lambda+2\mu)\alpha\mathrm{tr}\mathbf{E}(\theta-\theta_0) - \frac{c_v}{2\theta_0}(\theta-\theta_0)^2$$
$$\chi(\dot{\mathbf{E}},\nabla\theta) = \frac{\lambda'}{2}\mathrm{tr}^2\dot{\mathbf{E}} + \mu'\mathrm{tr}\dot{\mathbf{E}}^2 + \frac{\kappa}{2}||\nabla\theta||^2 \qquad (1)$$

where λ, μ, α, c_v, λ', μ' and κ are respectively the Lamé constants, the thermal expansion coefficient, specific heat capacity coefficient, viscosity coefficient and heat conduction coefficient. The reference temperature is denoted by θ_0. Parameters α, c_v and κ are considered as constants.

Hypothesis 3.1. *For the thermodynamic potentials given by the relations* (1)*, the Lamé's constants λ, μ are known for the hydrogel HEMA-EGDMA, the specific heat capacity coefficient is estimated by microcalorimetric test. The remaining constant are unknowns ($\alpha[1/K]$, $\lambda'[MPa.s]$, $\mu'[MPa.s]$ and $\kappa[W/(m.K)]$). We assume the following mechanical properties for the sample :*

Samples	E[MPa]	ν	$\lambda[MPa]$	$\mu[MPa]$	$c_v[J/(kg.K)]$
Sample 1	10-30	0.45	3.10-9.3	0.34-1.02	2900-3200
Sample 2	20-50	0.40	2.86-7.15	0.71-1.78	2900-3200

The balance of linear momentum and the energy conservation allow us to express the governing equations of the hydrogel sample and can be formulated as :

$$\begin{cases} \mathbf{DivFS^e} + \mathbf{DivFS^v} + \rho\mathbf{B} = \rho\dfrac{\partial^2\mathbf{u}}{\partial t^2} & \text{in } (\mathcal{B}\times[0,T]) \\ \rho\dot{e} = (\mathbf{S^e}+\mathbf{S^v}) : \dot{\mathbf{E}} - \mathbf{DivQ} + \rho r & \text{in } (\mathcal{B}\times[0,T]) \end{cases} \qquad (2)$$

where $\mathbf{S^e}(\mathbf{E},\theta) = \rho\partial\psi/\partial\mathbf{E}(\mathbf{E},\theta)$ and $\mathbf{S^v}(\dot{\mathbf{E}},\nabla\theta) = \partial\chi/\partial\dot{\mathbf{E}}(\dot{\mathbf{E}},\nabla\theta)$ are the elastic and viscous parts of the second Piola-Kirchhoff stress tensor $\mathbf{Q}/\theta = -\partial\chi/\partial\nabla\theta(\dot{\mathbf{E}},\nabla\theta)$ is the heat flux, $e = \psi(\mathbf{E},\theta) + s\theta$ the internal energy, $s = -\partial\psi/\partial\theta(\mathbf{E},\theta)$ the entropy density and $\mathbf{E} = \nabla\mathbf{u} + \nabla^{\mathbf{T}}\mathbf{u} + \nabla\mathbf{u}\nabla^{\mathbf{T}}\mathbf{u}/2$ is the Green-Lagrange strain tensor.

Equations of the three-dimensional continuum, developed avove, define the initial boundary value problem of thermomechanics. In detail, these were the description of deformation in the context of kinematics, the formulation of the force equilibrium based on kinetic considerations, the constitutive equation as well as the initial and boundary conditions. We assume the following mechanical boundary conditions which include three parts, preloading, cyclic loading and relaxation ($\mathbf{S}^t\mathbf{DN}$).

$$\begin{cases} \mathbf{u}\cdot\mathbf{n} = -\begin{cases} u_p\left(\dfrac{t}{\tau}\right) & \text{if } t < t_p \\ u_p\left(\dfrac{t_p}{\tau}\right) + \mathbf{u}_0\cos(2\pi f t) & \text{if } t_p \le t \le t_c \end{cases} & \text{on } (\Gamma_t \times [0,T]) \\ \mathbf{P}\cdot\mathbf{n} = \mathbf{0} & \text{if } t > t_c \quad \text{on } (\Gamma_t \times [0,T]) \\ \mathbf{P}\cdot\mathbf{n} = \mathbf{0} & \text{on } (\Gamma_l \times [0,T]) \\ \mathbf{u}\cdot\mathbf{n} = u_0 & \text{on } (\Gamma_u \times [0,T]) \\ \mathbf{P} = \mathbf{F}(\mathbf{S^e} + \mathbf{S^v}) \text{ in } (\mathcal{B}\times[0,T]) & \\ \textbf{I.C } \mathbf{u}(t=0,\cdot) := 0,\ \mathbf{P}(t=0,\cdot) := \mathbf{0} & \text{in } (\mathcal{B}\times\{0\}) \end{cases} \tag{3}$$

where $\tau \in \mathcal{R}_+$ is a time constant. $u_p \in \mathcal{R}$ denotes the prescribed displacement during the preloading and the relaxation. $u_0 \in \mathcal{R}$ denotes the prescribed displacement during the cyclic loading. We consider two time characteristics $t_p \in \mathcal{R}_+$ the preloading time and $t_c \in \mathcal{R}_+$ the time during which the cyclic load is applied. Experimentally, we apply the preload as a ramp form during the preload time t_p. Then we apply the mechanical cyclic loading during the load time t_c. Finally, after $t_c + t_p$, the discharge and relaxation time are beginning for a new t_p. For the heat boundary condition, we use the same continuous

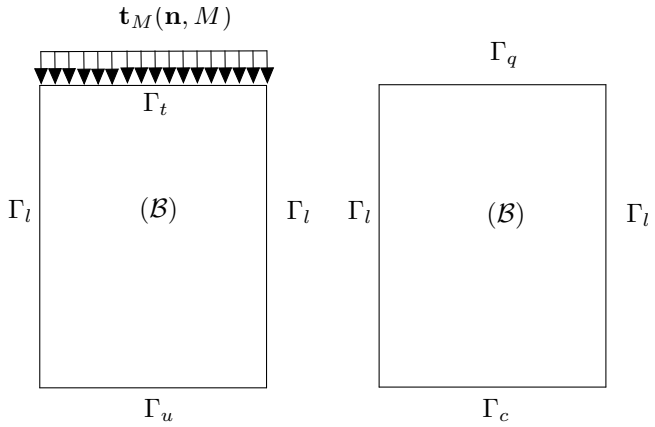


Figure 1. *Boundary conditions : mechanical boundary conditions (left), heat transfer boundary conditions (right).*

media $\mathcal{B} \in \mathcal{R}^d$ with the $V^{\mathcal{B}}$ the volume. The boundary of $\mathcal{B}$ is $\partial\mathcal{B} = \Gamma_q \cup \Gamma_l \cup \Gamma_c$ with the surface $S^{\mathcal{B}}$. For each time $t \in \mathcal{R}_+$ this volume is under heat production density ρr, a heat flux q_0 on one parts of the boundary of $\mathcal{B}$ and with a prescribed temperature θ_0 on other parts of the boundary of $\mathcal{B}$. The heat boundary can written as :

$$\begin{cases} \mathbf{Q}\cdot\mathbf{n} = q_0 & \text{on } (\Gamma_q \times [0,T]) \\ \mathbf{Q}\cdot\mathbf{n} = 0 & \text{on } (\Gamma_l \times [0,T]) \\ \mathbf{Q}\cdot\mathbf{n} = k_c(\theta - \theta_\infty) & \text{on } (\Gamma_c \times [0,T]) \\ \textbf{I.C } \theta(t=0,\cdot) := \theta_{ref} & \text{in } (\mathcal{B}\times\{0\}) \end{cases} \tag{4}$$

in which, q_0 is the prescribed heat flux on $(\Gamma_q \times [0,T])$, k_c denotes the convection coefficient and θ_0 is the prescribed temperature, θ_{ref} is the initial local temperature of the sample and θ_0 is the thermodynamic temperature.

By using the definition of the potential ψ and χ in the equation (1), the elastic and viscous parts of the second Piola-Kirchhoff hold :

$$\mathbf{S^e} = \lambda \mathrm{tr}(\mathbf{E})\mathbf{I} + 2\mu\mathbf{E} - (3\lambda + 2\mu)\alpha(\theta - \theta_0)\mathbf{I}; \quad \mathbf{S^v} = \lambda' \mathrm{tr}(\dot{\mathbf{E}})\mathbf{I} + 2\mu'\dot{\mathbf{E}} \tag{5}$$

In order to identify the numerical parameters of the self-heating model with the experimental measurements and for the correlation study, we compute the Cauchy stress tensor in the current configuration. For this purpose, we use the classical formulation with the deformation gradient. Then, the elastic part and the viscous part of the Cauchy stress tensor are given successively by :

$$\begin{aligned} \sigma^{\mathbf{e}} = \frac{\lambda}{J}\mathrm{tr}(\mathbf{E})\mathbf{F}\mathbf{I}\mathbf{F}^{\mathbf{T}} + 2\frac{\mu}{J}\mathbf{F}\mathbf{E}\mathbf{F}^{\mathbf{T}} - (3\lambda + 2\mu)\frac{\alpha}{J}(\theta - \theta_0)\mathbf{F}\mathbf{I}\mathbf{F}^{\mathbf{T}} \\ \sigma^{\mathbf{v}} = \frac{\lambda'}{J}\mathrm{tr}(\dot{\mathbf{E}})\mathbf{F}\mathbf{I}\mathbf{F}^{\mathbf{T}} + 2\frac{\mu'}{J}\mathbf{F}\dot{\mathbf{E}}\mathbf{F}^{\mathbf{T}} \end{aligned} \tag{6}$$

Starting from the expression of the heat flux $\mathbf{Q} = -\kappa\theta\nabla\theta$ in $(\mathcal{B} \times [0, T])$, by using the divergence theorem and rearranging the terms in the heat equation, the governing equation (2) can be written as :

$$\begin{cases} \mathbf{Div}\left[\rho(\lambda \mathrm{tr}(\mathbf{E})\mathbf{F}\mathbf{I} + 2\mu\mathbf{F}\mathbf{E}) - (3\lambda + 2\mu)\alpha(\theta - \theta_0)\mathbf{F}\mathbf{I}\right] + \mathbf{Div}\left[\lambda' \mathrm{tr}(\dot{\mathbf{E}})\mathbf{F}\mathbf{I} + 2\mu'\mathbf{F}\dot{\mathbf{E}}\right] \\ +\rho\mathbf{b} = \rho\dfrac{\partial^2\mathbf{u}}{\partial t^2} \text{ in } (\mathcal{B} \times [0, T]) \\ \rho\dfrac{c_v}{\theta_0}\dfrac{\partial\theta}{\partial t}\theta = (3\lambda + 2\mu)\alpha\theta \mathrm{tr}\dot{\mathbf{E}} + \lambda' \mathrm{tr}^2\dot{\mathbf{E}} + 2\mu' \mathrm{tr}\dot{\mathbf{E}}^2 - \kappa\theta\Delta\theta + \kappa||\nabla\theta||^2 + \rho r \\ \textbf{B.C} \text{ and } \textbf{I.C} \text{ (Cf. eq.(3) and (4))} \end{cases} \tag{7}$$

We assume two cases :

– Case 1 : **Local self-heating model** $\kappa \equiv 0$, $\mathbf{Q} \equiv 0$ For the hydrogel HEMA-EGDMA, the heat conductivity coefficient is very small ($\kappa \equiv 0$), then the heat flux by conduction in the sample is neglected ($\mathbf{Q} \equiv 0$). Analogously, the change in internal energy caused by the sources of heat is local vanishes and there is no heat diffusion in the media.

Hypothesis 3.2 (**Local self-heating model**). *We assume for this case that we have a local heat production. The internal heat production is not function of the space but just function of time $\theta := \theta(t)$. In this case, the quantity $\mathbf{Div}\left[(3\lambda + 2\mu)\alpha(\theta - \theta_0)\mathbf{F}\mathbf{I}\right] \equiv 0$ (effect of the temperature change on stress) in the governing equation* (7)*. In fact, we have the effect of the velocity on the internal heat production.*

For the second approximation we assume that, for the hydrogel HEMA-EGDMA, the heat conductivity coefficient of the sample is significant ($\kappa \neq 0$), then the heat flux by conduction in the sample is also significant ($\mathbf{Q} \neq 0$). Indeed, the change in internal energy is caused by the sources of heat and the deformation.

– Case 2 : $\kappa \neq 0$,

Hypothesis 3.3 (**Total self-heating model**). *In this case, we assume that the total heat is function of the space, the gradient of temperature and displacement. In fact, the heat conductivity is not neglected, then, the internal heat production is function of the space and time $\theta := \theta(x, t)$. In this case, the quantity $\mathbf{Div}\left[(3\lambda + 2\mu)\alpha(\theta - \theta_0)\mathbf{F}\mathbf{I}\right] \neq 0$ (effect of the temperature on stress) in the governing equation* (7)*. In fact, we have the two coupling terms : the effect of the velocity on the internal heat production and the effect of the temperature change on the stress.*

The character of the initial boundary value problem of structural mechanics depends on the types of structure and loading that have to be described, which, on the other hand, decisively affect the modeling of the load-carrying behavior. In the previous sections, the essential modeling aspects were already discussed on geometrical and material levels. In summary, the modeling can be categorized, in essence, according to the aspects of geometrical linearity or non-linearity, material linearity or non-linearity, and time-dependence or time-independence. The various approximation levels differ significantly in the complexity of the numerical solution of the underlying physical problem. The correlation between the simplification of the physical problem and the complexity of the numerical solution is illustrated in this work. Furthermore, the dynamic or static formulation of the problem is decisive for the effort expanded on the numerical solution.

We assume linearity of the temperature and the displacement. For physical consideration, the sample dimension is small for the hydrogel HEMA-EGDMA, we therefore assume that the heat production in the sample is local.

Hypothesis 3.4 (**Linearity in temperature**). *We assume small variation of the temperature distribution in the sample the prescribed cyclic displacement. The temperature $\theta \in \mathcal{R}_+$ is expessed as a reference temperature $\theta_0 \in \mathcal{R}_+$ plus the perturbation $\delta\theta \in \mathcal{R}_+$. We have : $\theta = \theta_0 + \delta\theta$ and $\dot{\theta} = \delta\dot{\theta}$.*

– Case 1 : **Local self-heating model** $\kappa \equiv 0$, $\mathbf{Q} \equiv 0$, Cf. hypothesis 3.2
The governing equation can be written as follows :

$$\begin{cases} \mathbf{Div}\left[\rho(\lambda \mathrm{tr}(\mathbf{E})\mathbf{FI} + 2\mu\mathbf{FE}) - (3\lambda+2\mu)\alpha\delta\theta\mathbf{FI}\right] + \mathbf{Div}\left[\lambda' \mathrm{tr}\dot{\mathbf{E}}\mathbf{FI} + 2\mu'\mathbf{F}\dot{\mathbf{E}}\right] \\ +\rho\mathbf{B} = \rho\dfrac{\partial^2\mathbf{u}}{\partial t^2} \text{ in } (\mathcal{B}\times[0,T]) \\ \rho c_v \dfrac{\partial\delta\theta}{\partial t} = (3\lambda+2\mu)\alpha(\theta_0+\delta\theta)\mathrm{tr}\dot{\mathbf{E}} + \lambda'\mathrm{tr}^2\dot{\mathbf{E}} + 2\mu'\mathrm{tr}\dot{\mathbf{E}}^2 + \rho r \text{ in } (\mathcal{B}\times[0,T]) \\ \textbf{B.C} \text{ and } \textbf{I.C} \text{ (Cf. eq.(3) and (4))} \end{cases} \tag{8}$$

– Case 2 : **Total self-heating model** $\kappa \neq 0$, $\mathbf{Q} \neq 0$, Cf. hypothesis 3.3.
The governing equation can be written as follows :

$$\begin{cases} \mathbf{Div}\left[\rho(\lambda \mathrm{tr}(\mathbf{E})\mathbf{FI} + 2\mu\mathbf{FE}) - (3\lambda+2\mu)\alpha\delta\theta\mathbf{FI}\right] + \mathbf{Div}\left[\lambda' \mathrm{tr}(\dot{\mathbf{E}})\mathbf{FI} + 2\mu'\mathbf{F}\dot{\mathbf{E}}\right] \\ +\rho\mathbf{B} = \rho\dfrac{\partial^2\mathbf{u}}{\partial t^2} \text{ in } (\mathcal{B}\times[0,T]) \\ \rho c_v \dfrac{\partial\delta\theta}{\partial t} = (3\lambda+2\mu)\alpha(\theta_0+\delta\theta)\mathrm{tr}\dot{\mathbf{E}} + \lambda'\mathrm{tr}^2\dot{\mathbf{E}} + 2\mu'\mathrm{tr}\dot{\mathbf{E}}^2 - \kappa\theta_0\Delta\delta\theta + \rho\mathbf{R} \\ \text{in } (\mathcal{B}\times[0,T]) \\ \textbf{B.C} \text{ and } \textbf{I.C} \text{ (Cf. eq.(3) and (4))} \end{cases} \tag{9}$$

In order to show the solution of the problem with the applicability of the thermoviscoelastic model as defined in the equation (9), we firstly assume one and two dimensional problem.

4. 2D and 1D approaches

As preliminary steps, it is important to recall the two and monodimensional formulation. The thermomechanical formulation will help us to understand each term appearing

in the equation (9). We assume one-dimensional compression. For the deformation analysis of two-dimensional continua, the plane stress and the plane strain states are of interest. The plane strain state is mostly used in cases where the dimension in one direction is very large with the loading in this direction remaining unchanged. The derivation of these equations can be found in the following sections.

Hypothesis 4.1 (Small strain assumption)**.** *As a first approximation the essential components of the description are small, linear elastic deformations*

We used the dimensionless form of the govering equation. For this purpose, we introduce new variables as defined in the equation (21) :

$$\hat{x} = \frac{x}{\ell}; \quad \hat{u} = \frac{u}{u_0}; \quad \hat{\dot{u}} = \frac{\dot{u}}{\dot{u}_0}; \quad \hat{t} = \frac{t}{t_0}; \quad \delta\hat{\theta} = \frac{\delta\theta}{\theta_0}. \tag{10}$$

The governing equation with the initial and the boundary conditions, and keeping the notation u but not $\hat{u}$ can be written in the following form :

$$\begin{cases} \dfrac{A}{C}\left(\dfrac{\partial^2 u}{\partial x^2}\right) + \dfrac{G}{C}\left(\dfrac{\partial \delta\theta}{\partial x}\right) + \dfrac{B}{C}\left(\dfrac{\partial^2 \dot{u}}{\partial x^2}\right) + \rho\bar{b} = \dfrac{\partial^2 u}{\partial t^2} \quad \text{in } (\mathcal{B}\times[0,T]) \\ \dfrac{\partial \delta\theta}{\partial t} = \dfrac{D}{F}(\theta_0+\delta\theta)\left(\dfrac{\partial \dot{u}}{\partial x}\right) + \dfrac{E}{F}\left(\dfrac{\partial^2 \dot{u}}{\partial x^2}\right) + \dfrac{H}{F}\dfrac{\partial^2 \theta\theta}{\partial x^2} + \rho\bar{R} \quad \text{in } (\mathcal{B}\times[0,T]) \\ \delta\theta(x,0) = \theta_{ref}; \quad \left(-\kappa\dfrac{\partial \delta\theta}{\partial x}\right)_{x=0} = 0; \quad \left(-\kappa\dfrac{\partial \delta\theta}{\partial x}\right)_{x=\ell} = 0; \quad u(0,t) = 0 \\ u(\ell,t) = u_\ell \sin(\omega t); \quad u(x,0) = 0; \quad \dot{u}(x,0) = 0 \end{cases} \tag{11}$$

In which,

$$A = \rho\frac{(\lambda+2\mu)}{\ell^2}u_0^2; \quad B = \frac{(\lambda'+2\mu')}{\ell^2}\dot{u}_0^2; \quad C = \rho\frac{u_0}{t_0^2}; \quad F = \rho c\frac{\theta_0}{t_0}; \tag{12}$$

$$D = \frac{(3\lambda+2\mu)\alpha\theta_0}{\ell}\dot{u}_0; \quad E = \frac{(\lambda'+2\mu')}{\ell^2}\dot{u}_0^2; \quad G = \frac{(3\lambda+2\mu)\rho\alpha}{\ell}\theta_0; \quad H = \frac{\kappa\theta_0^2}{\ell^2}. \tag{13}$$

– Case 1 : **Local self-heating model**, $\kappa \equiv 0$, $\frac{G}{C}\frac{\partial\theta}{\partial x} \equiv 0$, $\frac{H}{F}\frac{\partial^2\theta}{\partial x^2} \equiv 0$, Cf. hypothesis 3.3.

For the first approximation, we assume that the heat source $\rho r = 0$ and the body force $\rho\bar{b} = 0$, then, we introduce $K_1 := \frac{A}{C}$, $K_2 := \frac{B}{C}$, $K_3 := \frac{D}{F}$, $K_4 := \frac{E}{F}$. For the first equation, we use the variable (space-time) separation $u(x,t) = \phi(x)T(t)$ in the first equation, for a physic solution we have $\ddot{T}(t) + K_2k^2\dot{T}(t) + K_1k^2T(t) = 0$. The characteristic equation is given by $r^2 + K_2k^2r + K_1k^2 = 0$, the discriminant is $\Delta = K_2^2k^4 - 4K_1k^2$. We define a critical damping for $\Delta = 0$, $K_2^c = 2\frac{\sqrt{K_1}}{k}$, the damping coefficient is defined as $\zeta := \frac{K_1}{K_2^c} = \frac{K_2k}{2\sqrt{K_1}}$. We denote by $\Omega_0 = K_1k^2$, we have $\ddot{T}(t) + 2\zeta k\Omega_0\dot{T}(t) + \Omega_0^2T(t) = 0$ The characteristics equation is given by $s^2 + 2\zeta\Omega_0 s + \Omega_0^2 = 0$, the discriminant is $\Delta_s = 4\Omega_0^2\left(\zeta^2-1\right)$. For the solution, we assume that $T(0) = T_0$, $\dot{T}(0) = 0$ and consider following three cases :

1) **Critical damping** $\zeta = 1$, $\Delta_s = 0$, $s = -\Omega_0$ The solution is $T(t) = ae^{st} = ae^{-\Omega_0 t}$, the expression $bte^{-\Omega_0 t}$ also satisfies the differential equation. We have $T(t) = (a+bt)e^{st}$, in which $a = T_0$ and $b = T_0\omega_0$, In this case we have

$$T(t) = T_0(1 + \Omega_0 t)e^{-\Omega_0 t}$$

$$u(x,t) = T_0 \sum_{n=1}^{+\infty} u_\ell \frac{\sin(\omega_n t)}{\sin(k\ell)} \sin(k_n x)(1 + \Omega_0 t) e^{-\Omega_0 t}$$

$$\delta\theta(t;x) = \sum_{n=1}^{+\infty} \frac{K_4}{K_3} k\tan(k_n x)$$

$$+ \sum_{n=1}^{+\infty} \exp\left(- \frac{e^{-\Omega_0 t} k\cos(k_n x)\sin(\omega_n t) K_3 T_0 u_\ell \, (1 + \Omega \, t_0)}{\sin(k\ell)} \right) \left(\theta_{ref} - \frac{K_4}{K_3} k\tan(k_n x) \right)$$

$$S_{33} = \sum_{n=1}^{+\infty} \frac{T_0 u_\ell k\ell}{\sin(k\ell)} e^{-\Omega_0 t} \cos(k_n x) \left[\omega\cos(\omega_n t) K_2 \, (1 + \Omega_0 t) \right]$$

$$\sum_{n=1}^{+\infty} \frac{T_0 u_\ell k\ell}{\sin(k\ell)} e^{-\Omega_0 t} \cos(k_n x) \left[+\sin(\omega_n t) \left(-K_2 \Omega_0^2 t + K_1 \, (1 + \Omega_0 t) \right) \right]$$

2) **Sub-critical damping** $\zeta < 1$, $\Delta_s < 0$

$$s_1 = -\Omega_0 \left(\zeta + j\sqrt{1-\zeta^2} \right), \qquad s_2 = -\Omega_0 \left(\zeta - j\sqrt{1-\zeta^2} \right), \qquad j^2 = -1 \tag{14}$$

We denote by $\Omega = \Omega_0 \sqrt{1-\zeta^2}$ the solution can be written as :

$$T(t) = \frac{T_0}{2} \left[\left(1 + \frac{j\zeta\Omega_0}{\Omega} \right) e^{-(\Omega_0\zeta + j\Omega)t} + \left(1 - \frac{j\zeta\Omega_0}{\Omega} \right) e^{-(\Omega_0\zeta - j\Omega)t} \right]$$

$$T(t) = \frac{T_0}{2} e^{-\Omega_0\zeta t} \left[\left(1 + \frac{j\zeta\Omega_0}{\Omega} \right) e^{-j\Omega t} + \left(1 - \frac{j\zeta\Omega_0}{\Omega} \right) e^{j\Omega t} \right] \tag{15}$$

Using the transformation of $e^{-j\Omega t}$ and $e^{j\Omega t}$, we have

$$T(t) = T_0 e^{-\Omega_0\zeta t} \left[\cos(\Omega t) + \frac{\Omega_0\zeta}{\Omega} \sin(\Omega t) \right]$$

$$u(x,t) = T_0 \sum_{n=1}^{+\infty} u_\ell \frac{\sin(\omega_n t)}{\sin(k\ell)} \sin(k_n x) e^{-\Omega_0\zeta t} \left[\cos(\Omega t) + \frac{\Omega_0\zeta}{\Omega} \sin(\Omega t) \right]$$

$$\delta\theta(t;x) = \sum_{n=1}^{+\infty} \frac{K_4}{K_3} k\tan(k_n x)$$

$$+ \sum_{n=1}^{+\infty} \exp\left(- \frac{e^{-\zeta\Omega_0 t} k\cos(k_n x)\sin(\omega_n t) K_3 T_0 u_\ell \, (\Omega\cos(\Omega t) + \zeta\sin(\Omega t)\Omega_0)}{\sin(k\ell)\Omega} \right)$$

$$\left(\theta_{ref} - \frac{K_4}{K_3} k\tan(kx) \right)$$

$$S_{33} = \sum_{n=1}^{+\infty} \frac{T_0 u_\ell k\ell}{\sin(k_n x)\Omega} e^{-\zeta\Omega_0 t} \cos(k_n x) \left[\omega\cos(\omega_n t) K_2 \, (\Omega\cos(\Omega t) + \zeta\sin(\Omega t)\Omega_0) \right]$$

$$+ \sum_{n=1}^{+\infty} \frac{T_0 u_\ell k\ell}{\sin(k_n x)\Omega} e^{-\zeta\Omega_0 t} \cos(k_n x) \left[\sin(\omega_n t) \left(K_1 \, (\Omega\cos(\Omega t) + \zeta\sin(\Omega t)\Omega_0) \right) \right]$$

$$+ \sum_{n=1}^{+\infty} \frac{T_0 u_\ell k\ell}{\sin(k_n x)\Omega} e^{-\zeta\Omega_0 t} \cos(kx) \left[\sin(\omega_n t) \left(-\sin(\Omega t) K_2 \left(\Omega^2 + \zeta^2\Omega_0^2 \right) \right) \right]$$

3) **Super-critical damping** $\zeta > 1$, $\Delta_s > 0$

$$s_1 = -\Omega_0 \left(\zeta + \sqrt{\zeta^2 - 1} \right), \qquad s_2 = -\Omega_0 \left(\zeta - \sqrt{\zeta^2 - 1} \right) \tag{16}$$

The solution is

$$T(t) = \frac{T_0}{2} e^{-\zeta\Omega_0 t}\left[(1-Y)\, e^{-\Omega_0\sqrt{\zeta^2-1}t} + (1+Y)\, e^{\Omega_0\sqrt{\zeta^2-1}t}\right]$$

$$u(x,t) = \frac{T_0}{2}\sum_{n=1}^{+\infty} u_\ell \frac{\sin(\omega_n t)}{\sin(k\ell)}\sin(k_n x) e^{-\zeta\Omega_0 t}\left[(1-Y)\, e^{-\Omega_0\sqrt{\zeta^2-1}t}\right]$$

$$\frac{T_0}{2}\sum_{n=1}^{+\infty} u_\ell \frac{\sin(\omega_n t)}{\sin(k\ell)}\sin(k_n x) e^{-\zeta\Omega_0 t}\left[+(1+Y)\, e^{\Omega_0\sqrt{\zeta^2-1}t}\right]$$

$$\delta\theta(t;x) = \sum_{n=1}^{+\infty}\frac{K_4}{K_3}k\tan(k_n x)$$

$$+\sum_{n=1}^{+\infty}\exp\left[-\frac{2e^{-\zeta\Omega_0 t}k\cos(k_n x)\sin(\omega_n t)\zeta\sinh\left(Y_s\Omega_0 t\right)K_3 T_0 u_\ell}{\sin(k\ell)\sqrt{\zeta^2-1}}\right]\left(\theta_{ref} - \frac{K_4}{K_3}k\tan(k_n x)\right)$$

$$+\sum_{n=1}^{+\infty}\exp\left[-\frac{2e^{-\zeta\Omega_0 t}k\cos(k_n x)\sin(\omega t)Y_s\cosh\left(Y_s\Omega_0 t\right)K_3 T_0 u_\ell}{\sin(k\ell)Y_s}\right]\left(\theta_{ref} - \frac{K_4}{K_3}k\tan(k_n x)\right)$$

$$S_{33} = \sum_{n=1}^{+\infty}\frac{T_0 u_\ell k\ell\cos(k_n x)}{\sin(kl)Y_s}e^{-(\zeta+Y_s)\Omega_0 t}\left\{\left[\left(-1+e^{2Y_s\Omega_0 t}\right)\zeta + \left(1+e^{2Y_s\Omega_0 t}\right)Y_s\right]\right.$$

$$\sin(\omega_n t)K_1 + K_2\left[\left(-1+e^{2Y_s\Omega_0 t}\right)\zeta + \left(1+e^{2Y_s\Omega_0 t}\right)Y_s\right]\omega\cos(\omega_n t)$$

$$\left.-K_2\left(-1+e^{2Y_s\Omega_0 t}\right)\sin(\omega_n t)\Omega_0\right\}$$

In which $Y = \frac{\zeta}{\sqrt{\zeta^2-1}}$ and $Y_s = \sqrt{\zeta^2-1}$.

– Case 2 : **Total self-heating model**, $\kappa \neq 0$, $\frac{G}{C}\left(\frac{\partial\theta}{\partial x}\right) \neq 0$, $\frac{H}{F}\theta\frac{\partial^2\theta}{\partial x^2} \neq 0$, Cf. hypothesis 3.3.

Remark 4.1. *The local behavior of a thermoviscoelastic body for one dimensional problem was totally described in the previous section by means of the initial boundary value problem. Generally, the solution of this differential equation is not analytically explicit. Therefore, approximation methods, in particular the Finite Element Method, are used in order to find an approximate solution. This method does not solve the strong form of the differential equation. It merely solves its integral over the domain, the so-called weak form of the differential equation. This weak formulation forms the basic prerequisite for the application of approximation methods.*

5. Identification of the model parameters

For a given thermodynamic potential, the main problem after the formulation is to calculate or measure the physical constants in the model. If the physical constants can be identified with the experimental measurement, it is appropriate to determine these constants by using classical identification procedures. In the opposite case, we need to identify these constants by using analytical/numerical approaches. For that, we use the one dimension analytical description in order to identify the physical constant in the model.

5.1. Cost functions

According to the classical method of optimization, the identification method of physical constant in the model of self-heating (thermoviscoelasticity) can be expressed using complex parameters. The parameters to be identified are α, λ', μ' and κ

Definition 5.1 (Cost functions)**.** *The cost function for the self-heating model is given by the following equation and we have to minimize the following coupled cost function :*

$$\mathbf{x} = \inf_{\alpha\in\mathcal{R}_+}\inf_{\lambda'\in\mathcal{R}_+}\inf_{\mu'\in\mathcal{R}_+}\inf_{\kappa\in\mathcal{R}_+}\left\{\begin{array}{l} f\left(\left(\frac{2}{J}\mathbf{FSF^T}\right)^{comp}_{/33}(\alpha,\lambda',\mu') - \sigma^{obse}\right) \\ g\left(\delta\theta^{comp}(\alpha,\lambda',\mu',\kappa) - \delta\theta^{obse}\right)\end{array}\right\} \quad (17)$$

Where f and g are the functions used to measure the difference between the computed and observed quantity, in general we use the square function $f, g := \frac{1}{2}\|\cdot\|^2$.

$$\mathbf{S} = \lambda\mathrm{tr}(\mathbf{E})\mathbf{I} + 2\mu\mathbf{E} - (3\lambda+2\mu)\alpha(\theta-\theta_0)\mathbf{I} + \lambda'\mathrm{tr}(\dot{\mathbf{E}})\mathbf{I} + 2\mu'\dot{\mathbf{E}} \quad (18)$$

Definition 5.2 (Least square cost functions)**.** *For the first approximation, we define least square cost functions to identify the physical parameters of the model :*

$$\mathbf{x} = \inf_{\alpha\in\mathcal{R}_+}\inf_{\lambda'\in\mathcal{R}_+}\inf_{\mu'\in\mathcal{R}_+}\inf_{\kappa\in\mathcal{R}_+}\frac{1}{2}\left\{\begin{array}{l} \left\|\left(\frac{2}{J}\mathbf{FSF^T}\right)^{comp}_{/33}(\alpha,\lambda',\mu') - \sigma^{obse}\right\|^2 \\ \left\|\delta\theta^{comp}(\alpha,\lambda',\mu',\kappa) - \left(\delta\theta^{obse}+273.15\right)\right\|^2\end{array}\right\} \quad (19)$$

5.2. Computation, splitting

We present in this section the computation setting using splitting methods. The main step is summarized by the following scheme.

1) **Define** : Initialization $[\alpha_0, \lambda'_0, \mu'_0, \kappa_0]$; $\quad \delta\theta_0 = \theta_0 + 273.15$, ν

2) **Minimize Self-heating model** :

•LOOP ($k = 0\cdots n$)

---

a) **Minimize wave equation** : (input $[\alpha_k, \lambda'_k, \mu'_k, \kappa_k]$)

$$\mathbf{x} = \inf_{\alpha_k\in\mathcal{R}_+}\inf_{\lambda'_k\in\mathcal{R}_+}\inf_{\mu'_k\in\mathcal{R}_+}\inf_{\kappa_k\in\mathcal{R}_+}\frac{1}{2}\left\|\left(\frac{2}{J}\mathbf{FSF^T}\right)^{comp}_{/33}(\alpha_k,\lambda'_k,\mu'_k) - \frac{F^{obs}(t)}{S^{\mathcal{B}}}\right\|^2$$

$$\textbf{if} \quad \frac{\lambda'_k}{2(\lambda'_k+\mu'_k)} \geq \nu \text{ (physical condition)}$$

LOOP wave equation ($k \longleftarrow k+1$)

else

End (output $[\alpha_k, \lambda'_k, \mu'_k, \kappa_k]$)

---

b) **Minimize heat equation** : (input $[\alpha_k, \lambda'_k, \mu'_k, \kappa_k]$)

$$\mathbf{x} = \inf_{\alpha_k \in \mathcal{R}_+} \inf_{\lambda'_k \in \mathcal{R}_+} \inf_{\mu'_k \in \mathcal{R}_+} \inf_{\kappa_k \in \mathcal{R}_+} \frac{1}{2} \left\| \delta\theta^{comp}\left(\alpha_k, \lambda'_k, \mu'_k, \kappa_k\right) - \left(\delta\theta^{obse} + 273.15\right) \right\|^2$$

$$\textbf{if} \quad |\alpha_k - \alpha_{k+1}, \lambda'_k - \lambda'_{k+1}, \mu'_k - \mu'_{k+1}, \kappa_k - \kappa_{k+1}| \geq \epsilon$$

•LOOP ($k \longleftarrow k+1$)
else
End (output $[\alpha_k, \lambda'_k, \mu'_k, \kappa_k]$)

Hypothesis 5.1 (Cost functions for one dimensional model)**.** *For the one dimensional model, the constant K_1 is known via λ, μ. The unknowns are K_2, K_3, K_4. We have to minimize the following cost function.*

$$\mathbf{x} = \inf_{K_2 \in \mathcal{R}_+} \inf_{K_3 \in \mathcal{R}_+} \inf_{K_4 \in \mathcal{R}_+} \frac{1}{2} \left\{ \begin{array}{l} \left\| (2\mathbf{S})^{comp}_{/33}\left(K_1, K_2\right) - \dfrac{F^{obs}(t)}{S^{\mathcal{B}}} \right\|^2, \\ \| \ \delta\theta^{comp}\left(K_1, K_2, K_3, K_4\right) - \left(\delta\theta^{obse} + 273.15\right) \ \|^2 \end{array} \right\} \tag{20}$$

6. Numerical approximations

In this section, we propose a finite element method for a 2D stess elasticity problem. The equations established in the previous section are solved using a finite elements discretization in space. In time, an implicit Euler scheme is applied for the time integration. In fact, we consider finite element approximations of the pure dynamic displacement traction/compression boundary value in three-dimensional nonlinear thermomechanical viscoelasticity associated with a homogenous viscoelastic material. We use the following weak form of the governing equation. The corresponding weak formulation in space-time is obtained by multiplying by the test functions : firstly, for the balance of momentum, by the scalar product with a vector-valued test function $\delta\mathbf{u}$ which has to be compatible with the geometric boundary conditions. Then, this equation is integrated over the volume of the sample.

$$\left\{ \begin{array}{l} \displaystyle\int_{\mathcal{B}} \mathbf{Div}\left[\rho(\lambda \mathrm{tr}(\mathbf{E})\mathbf{FI} + 2\mu\mathbf{FE}) - (3\lambda + 2\mu)\alpha\delta\theta\mathbf{FI}\right]\delta\mathbf{u}\, dV^{\mathcal{B}} \\ \displaystyle + \int_{\mathcal{B}} \mathbf{Div}\left[\lambda' \mathrm{tr}(\dot{\mathbf{E}})\mathbf{FI} + 2\mu'\mathbf{F}\dot{\mathbf{E}}\right]\delta\mathbf{u}\, dV^{\mathcal{B}} + \int_{\mathcal{B}} \rho\mathbf{B}\delta\mathbf{u}\, dV^{\mathcal{B}} = \int_{\mathcal{B}} \rho\frac{\partial\mathbf{v}}{\partial t}\delta\mathbf{u}\, dV^{\mathcal{B}} \\ \displaystyle \int_{\mathcal{B}} \rho c_v \frac{\partial\delta\theta}{\partial t}\delta\theta^*\, dV^{\mathcal{B}} = \int_{\mathcal{B}} (3\lambda + 2\mu)\alpha(\theta_0 + \delta\theta)\mathrm{tr}\dot{\mathbf{E}}\delta\theta^*\, dV^{\mathcal{B}} + \int_{\mathcal{B}} \lambda' \mathrm{tr}^2\dot{\mathbf{E}}\delta\theta^*\, dV^{\mathcal{B}} \\ \displaystyle + \int_{\mathcal{B}} 2\mu' \mathrm{tr}\dot{\mathbf{E}}^2\delta\theta^*\, dV^{\mathcal{B}} - \int_{\mathcal{B}} \kappa\theta_0\Delta\delta\theta\delta\theta^*\, dV^{\mathcal{B}} + \int_{\mathcal{B}} r\delta\theta^*\, dV^{\mathcal{B}} \quad \forall\delta\theta^* \in [H^1(\mathcal{B})]^d \\ \mathbf{v} = \dfrac{\partial\mathbf{u}}{\partial t} \quad \text{in } (\mathcal{B} \times [0, T]) \end{array} \right. \tag{21}$$

For all $[\delta\Phi] = (\delta\mathbf{u}, \delta\theta^*)$. In which, $dV^{\mathcal{B}}$ and $dS^{\mathcal{B}}$ are respectively the volume and surface element. Using the divergence theorem and taking into account the boundary conditions,

the final representation of the weak form of the coupled self-heating model reads as follows :

$$\begin{cases} -\displaystyle\int_{\mathcal{B}} \rho(\lambda \mathrm{tr}(\mathbf{E})\mathbf{I} + 2\mu\mathbf{E}) : \nabla\delta\mathbf{u}^{\mathbf{T}}(\nabla\mathbf{u} + \mathbf{I})\, dV^{\mathcal{B}} - \int_{\mathcal{B}} (3\lambda + 2\mu)\rho\alpha\delta\theta\mathbf{I} : \nabla\delta\mathbf{u}^{\mathbf{T}} \\ (\nabla\mathbf{u} + \mathbf{I})\, dV^{\mathcal{B}} + \displaystyle\int_{\mathcal{B}} \left(\lambda' \mathrm{tr}(\dot{\mathbf{E}})\mathbf{I} + 2\mu'\dot{\mathbf{E}}\right) : \nabla\delta\mathbf{u}^{\mathbf{T}}(\nabla\mathbf{u} + \mathbf{I})\, dV^{\mathcal{B}} + \int_{\mathcal{B}} \rho\mathbf{B}\delta\mathbf{u}\, dV^{\mathcal{B}} \\ = \displaystyle\int_{\mathcal{B}} \rho\frac{\partial\mathbf{v}}{\partial t}\delta\mathbf{u}\, dV^{\mathcal{B}} \quad \forall\delta\mathbf{u} \in [H^1(\mathcal{B})]^d \\ \displaystyle\int_{\mathcal{B}} \rho c_v \frac{\partial\delta\theta}{\partial t}\delta\theta^*\, dV^{\mathcal{B}} = \int_{\mathcal{B}} (3\lambda + 2\mu)\alpha(\theta_0 + \delta\theta)\mathrm{tr}\dot{\mathbf{E}}\delta\theta^*\, dV^{\mathcal{B}} + \int_{\mathcal{B}} \lambda' \mathrm{tr}^2\dot{\mathbf{E}}\delta\theta^*\, dV^{\mathcal{B}} \\ + \displaystyle\int_{\mathcal{B}} 2\mu' \mathrm{tr}\dot{\mathbf{E}}^2\delta\theta^*\, dV^{\mathcal{B}} + \int_{\mathcal{B}} \kappa\theta_0\nabla\delta\theta.\nabla\delta\theta^*\, dV^{\mathcal{B}} - \int_{\partial\mathcal{B}} \kappa\theta\nabla\delta\theta.\mathbf{n}.\delta\theta^*\, dS^{\mathcal{B}} \\ + \displaystyle\int_{\mathcal{B}} \rho r\delta\theta^*\, dV^{\mathcal{B}} \quad \forall\delta\theta^* \in [H^1(\mathcal{B})]^d \\ \mathbf{v} = \dfrac{\partial\mathbf{u}}{\partial t} \quad \text{in } (\mathcal{B} \times [0, T]) \end{cases} \tag{22}$$

6.1. Computations

For the computation we use Comsol Multiphysics to compute the model by using general form of PDE. This tool allows us to solve systems of time-dependent or stationary partial differential equations in one, two, and three dimensions with complex geometry. There are two forms of the partial differential equations available, the general form and the coefficient form. They read

$$\begin{aligned} e_a\frac{\partial^2\mathbf{u}}{\partial t^2} + d_a\frac{\partial\mathbf{u}}{\partial t} + \nabla\cdot\Gamma = \mathbf{F} \text{ in } (\mathcal{B}\times[0,T]) \\ -\mathbf{n}\cdot\Gamma = G + \left(\frac{\partial R}{\partial\mathbf{u}}\right)^T \mu; 0 = R \text{ on } (\partial\mathcal{B}\times[0,T]) \end{aligned}$$

$$\begin{aligned} e_a\frac{\partial^2\mathbf{u}}{\partial t^2} + d_a\frac{\partial\mathbf{u}}{\partial t} + \nabla\cdot(-c\nabla\mathbf{u} - a\mathbf{u} + \gamma) + a\mathbf{u} + \beta\cdot\nabla\mathbf{u} = \mathbf{f} \text{ in } (\mathcal{B}\times[0,T]) \\ -\mathbf{n}(-c\nabla\mathbf{u} - a\mathbf{u} + \gamma) + q\mathbf{u} = g - h^T\mu; h\mathbf{u} = R \text{ on } (\partial\mathcal{B}\times[0,T]) \end{aligned} \tag{23}$$

respectively. The second kind of equation (coefficient form) can only be used for mildly nonlinear problems. For most nonlinear problems, the general form needs to be used.

Remark 6.1. *The coefficients of the coefficient form may depend both on x, t, and $\mathbf{u}$. Observe that a dependence on $\mathbf{u}$ is not recommended. The flux vector Γ and the scalar coefficient F, G and R can be function of the spatial coordinates the solution $\mathbf{u}$ and the space and time derivatives of $\mathbf{u}$. The variable μ is the Lagrange multiplier, and T denotes the transpose. q and g are respectively the boundary absorption coefficient and the boundary source term.*

The second method, to solve numerically the non-linear mechanics in this software is to define directly the thermodynamic potential in the software. The thermodynamic conditions as convexity must be verified before introducing the thermodynamic potential.

$$\nabla\cdot(\sigma^e + \sigma^v) + \rho\mathbf{b} = \rho\frac{\partial^2\mathbf{u}}{\partial t^2}; \quad \sigma^e = J^{-1}\mathbf{F}\mathbf{S}^{\mathbf{e}}\mathbf{F}^{\mathbf{T}} \quad \text{in } (\mathcal{B}\times[0,T])$$

$$\sigma^v = J^{-1}\mathbf{F}\mathbf{S}^{\mathbf{v}}\mathbf{F}^{\mathbf{T}} \quad \text{in } (\mathcal{B}\times[0,T]) \quad \mathbf{F}=\nabla\mathbf{u}+\mathbf{I}; \quad J=det\mathbf{F};$$
$$\mathbf{E}=(\mathbf{C}-\mathbf{I})/2; \quad \mathbf{C}=\mathbf{F}^{\mathbf{T}}\mathbf{F}=\mathbf{I}+\nabla\mathbf{u}+\nabla\mathbf{u}+\nabla^{\mathbf{T}}\mathbf{u}\nabla\mathbf{u}/2$$
$$\mathbf{S}^{\mathbf{e}}=2\rho\frac{\partial\psi}{\partial\mathbf{C}}; \quad \mathbf{S}^{\mathbf{v}}=2\frac{\partial\chi}{\partial\dot{\mathbf{C}}} \quad \text{in } (\mathcal{B}\times[0,T]) \tag{24}$$

In which, $\mathbf{F}$ is the deformation gradient, $\mathbf{I}$ is the identity matrix, $\mathbf{E}$ and $\mathbf{C}$ denote respectively the Green-Lagrange and the Cauchy-Green strain tensors. To solve numerically the self-heating model we assume : for the first approximation, we use the general form of PDE given by the equation (23) (first equation) for the wave and the heat equations. In a second approximation, we use the second method (24), it consists to introduce directly the thermodynamic potential for the wave equation and the general form of PDE for the heat equation. In this work, we use these methods to compare the numerical solution of the self-heating model.

$$\begin{bmatrix} e_a^{\mathbf{u}} & 0 \\ 0 & e_a^{\theta} \end{bmatrix}\frac{\partial^2}{\partial t^2}\begin{pmatrix}\mathbf{u}\\ \delta\theta\end{pmatrix}+\begin{bmatrix} d_a^{\mathbf{u}} & 0 \\ 0 & d_a^{\theta} \end{bmatrix}\frac{\partial}{\partial t}\begin{pmatrix}\mathbf{u}\\ \delta\theta\end{pmatrix}+\nabla\cdot\begin{bmatrix}\Gamma^{\mathbf{u}}\\ \Gamma^{\theta}\end{bmatrix}+\nabla\cdot\begin{bmatrix}\Gamma^{\dot{\mathbf{u}}}\\ \Gamma^{\dot{\theta}}\end{bmatrix}$$
$$=\begin{pmatrix}\mathbf{F}^{\mathbf{u}}\\ \mathbf{F}^{\theta}\end{pmatrix}$$
$$-\mathbf{n}\cdot(\Gamma^{\mathbf{u}}+\Gamma^{\dot{\mathbf{u}}})=0, \quad G=0, \quad \text{on } (\Gamma_\ell\times[0,T])$$
$$\mathbf{R}=-\mathbf{u} \quad \text{on } (\Gamma_u\times[0,T])$$
$$\mathbf{R}=-\mathbf{u}-\mathbf{u}_0 \quad \text{on } (\Gamma_t\times[0,T])$$
$$-\mathbf{n}\cdot(\Gamma^{\theta}+\Gamma^{\dot{\theta}})=0, \quad G=0, \quad \text{on } (\partial\mathcal{B}-\Gamma_c\times[0,T])$$
$$-\mathbf{n}\cdot\Gamma^{\theta}=h(\delta\theta-\delta\theta_{ref}), \quad G=0, \quad \text{on } (\Gamma_c\times[0,T]) \tag{25}$$

Implementation in Comsol Multiphysics software is based on the equation 25.

6.2. Numerical approximations for local self-heating

Using the hypothesis for local self-heating in the sample, (Cf. hypothesis 3.2). The equation (**??**) becomes :

$$\begin{bmatrix} \rho & 0 \\ 0 & 0 \end{bmatrix}\frac{\partial^2}{\partial t^2}\begin{pmatrix}\mathbf{u}\\ \delta\theta\end{pmatrix}+\begin{bmatrix} 0 & 0 \\ 0 & \rho c_v \end{bmatrix}\frac{\partial}{\partial t}\begin{pmatrix}\mathbf{u}\\ \delta\theta\end{pmatrix}+\nabla\cdot\begin{bmatrix}\mathbf{F}\mathbf{S}^{\mathbf{e}}\\ 0\end{bmatrix}$$
$$+\nabla\cdot\begin{bmatrix}\mathbf{F}\mathbf{S}^{\mathbf{v}}\\ 0\end{bmatrix}=\rho\begin{pmatrix}\mathbf{B}\\ r\end{pmatrix} \tag{26}$$

In which

$$\mathbf{B}=0; \qquad r=(3\lambda+2\mu)\alpha(\theta_0+\delta\theta)\mathrm{tr}\dot{\mathbf{E}}+\lambda'\mathrm{tr}^2\dot{\mathbf{E}}+2\mu'\mathrm{tr}\dot{\mathbf{E}}^2$$
$$\mathbf{S}^{\mathbf{e}}=\lambda\mathrm{tr}(\mathbf{E})\mathbf{I}+2\mu\mathbf{E}-(3\lambda+2\mu)\alpha(\theta-\theta_0)\mathbf{I}; \qquad \mathbf{S}^{\mathbf{v}}=\lambda'\mathrm{tr}(\dot{\mathbf{E}})\mathbf{I}+2\mu'\dot{\mathbf{E}} \tag{27}$$

Implementation in Comsol Multiphysics software is based on the equation (26).

6.3. Numerical approximations for non-local self-heating

Cf. hypothesis 3.3. The equation (25) becomes :

$$\begin{bmatrix} \rho & 0 \\ 0 & 0 \end{bmatrix}\frac{\partial^2}{\partial t^2}\begin{pmatrix}\mathbf{u}\\ \delta\theta\end{pmatrix}+\begin{bmatrix} 0 & 0 \\ 0 & \rho c_v \end{bmatrix}\frac{\partial}{\partial t}\begin{pmatrix}\mathbf{u}\\ \delta\theta\end{pmatrix}+\nabla\cdot\begin{bmatrix}\mathbf{F}\mathbf{S}^{\mathbf{e}}\\ \kappa\nabla\delta\theta\end{bmatrix}$$
$$+\nabla\cdot\begin{bmatrix}\mathbf{F}\mathbf{S}^{\mathbf{v}}\\ 0\end{bmatrix}=\rho\begin{pmatrix}\mathbf{B}\\ r\end{pmatrix} \tag{28}$$

in which

$$\mathbf{B} = 0; \qquad r = (3\lambda + 2\mu)\alpha(\theta_0 + \delta\theta)\mathrm{tr}\dot{\mathbf{E}} + \lambda'\mathrm{tr}^2\dot{\mathbf{E}} + 2\mu'\mathrm{tr}\dot{\mathbf{E}}^2$$
$$\mathbf{S^e} = \lambda\mathrm{tr}(\mathbf{E})\mathbf{I} + 2\mu\mathbf{E}; \qquad \mathbf{S^v} = \lambda'\mathrm{tr}(\dot{\mathbf{E}})\mathbf{I} + 2\mu'\dot{\mathbf{E}} \tag{29}$$

Implementation in comsol multiphysics software is based on the equation (28).

7. Experimental and numerical results

As a first result, we want to verify that the experimental measurement of the temperature in the sample is not biaised by the friction between the hydrogel and the temperature sensor in the microcalorimeter during the deformation. We can then conclude that there is no temperature increase due to the friction and, then, eventual temperature increase will be due to self-heating phenomenon of the tested sample. The effect of the self-heating and corresponding temperature increase in the hydrogel is presented in figure 2. A clear

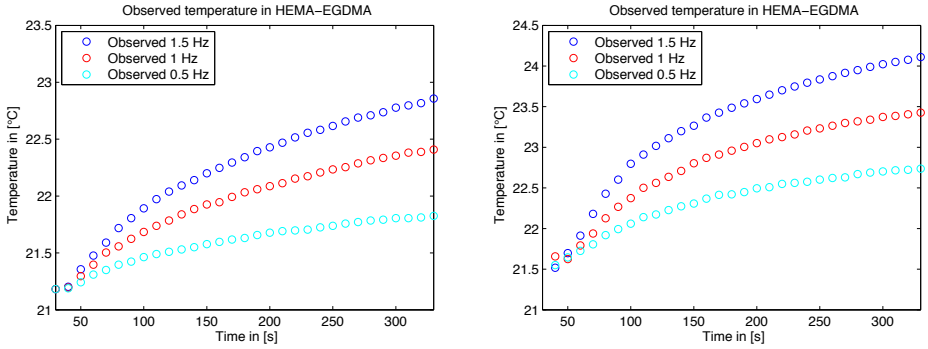


Figure 2. *Observed temperature in the sample of HEMA-EGDMA vs. time for* $\phi = 6\%$ *(left) and* $\phi = 8\%$ *(right),* $f = 0.5$ *[Hz],* $f = 1$ *[Hz] and* $f = 1.5$ *[Hz].*

temperature increase is obtained over time for the three different frequencies and two different cross-linkers concentration. The temperature increases between the initial and last cycles read 2.5^oC. There is clear dependency of the temperature increase to the applied frequency. The higher the frequency is, the higher the temperature increases. These experimental temperature evolution were used to identify the parameters present in the analytical 1D model. A good correlation is obtained between the experimental data and the model as shown in figure 3.

Based on the these correlations, the obtained identified parameters of the model are reported in Table 1.

Samples	$\lambda'[MPa.s]$	$\mu'[MPa.s]$	$\alpha[1/K]$
Sample 1	357.93	39.77	1.9e-4
Sample 2	393.646	51.701	2.1e-4

Tableau 1. *Optimized constants of the samples after equation* (20)

Finally the parameters reported on table 1 were injected in the FEM model (see equation (21)) and the computed temperature evolutions were then plotted in figure 4 It can

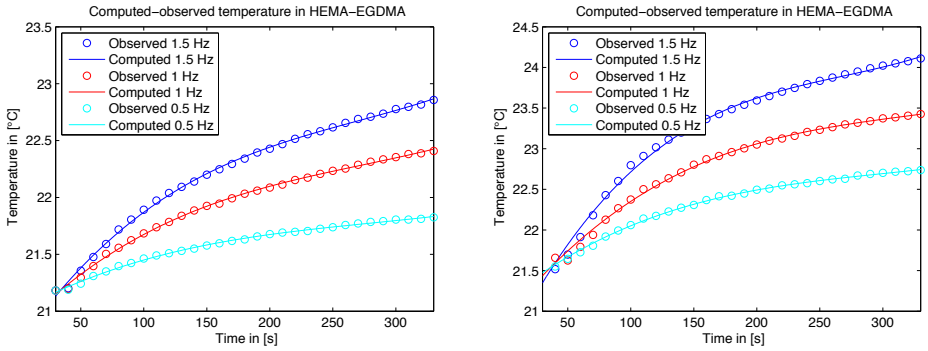


Figure 3. *Correlation between computed (analytical solution) and observed temperature in the sample of HEMA-EGDMA vs. time. for* $\phi = 6\%$ *(left) and* $\phi = 8\%$ *(right),* $f = 0.5$ *[Hz],* $f = 1$ *[Hz] and* $f = 1.5$ *[Hz].*

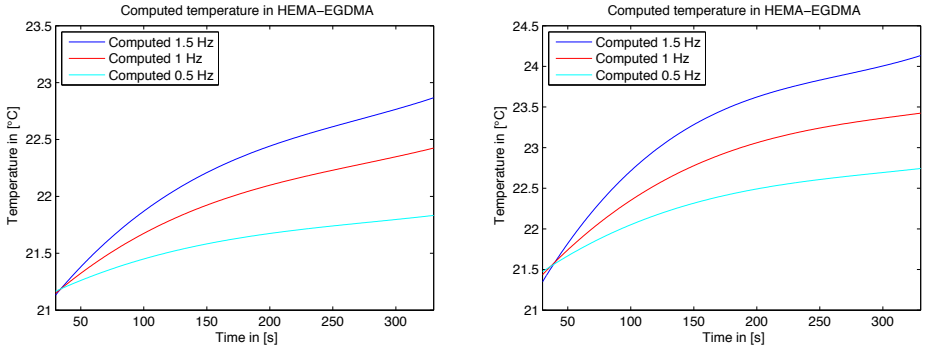


Figure 4. *Computed (numerical model) temperature in the sample of HEMA-EGDMA vs. time for* $\phi = 6\%$ *(left) and* $\phi = 8\%$ *(right),* $f = 0.5$ *[Hz],* $f = 1$ *[Hz] and* $f = 1.5$ *[Hz].*

be obtained that the obtained curves closely match the experimental measurement of the hydrogel self-heating, not only the frequency dependence, but also the cross-linkers dependence could be caught by the developed model.

7.1. Influence of the cross-link density on the self-heating

In order to have a closer look to the influence of cross-link density on the self-heating, we report on the same graph the temperature evolution of the hydrogels for the two different cross-linker density (6% and 8%). It can be observed on figure 5 that the decrease in the cross-linker density caused a significant change in the heat production and consequently a more limited temperature increase during cyclic loading. The effect of the cross-link density is implicitly taken into account in the model through the dependency of the cross-link density in the model parameters.

7.2. Dissipation in function of frequency and cross-link density

In this subsection, we present the experimental results for the dissipation in the hydrogel obtained from the force-displacement hysteresis curves. We evaluate the effect of the temperature increase on the dissipation during the different phase of the test (preloading, cyclic loading and relaxation).

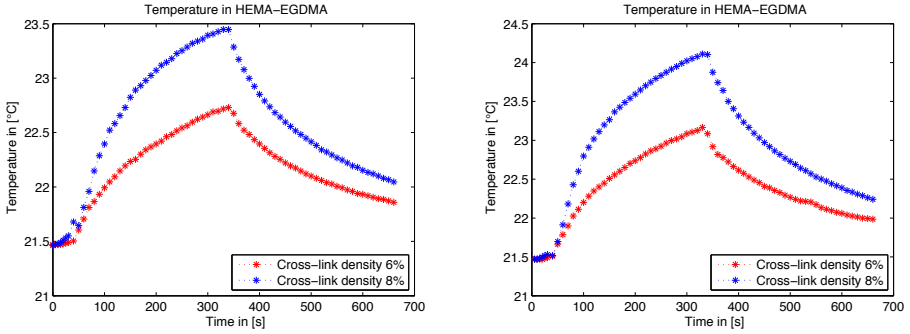


Figure 5. *Temperature (in [^{o}C]) vs. time (in [s]) in the HEMA-EGDMA samples. The curves show the effect of the cross-link density ϕ on the temperature during test (preloading, cyclic loading and relaxation). $f = 1$ [Hz] for the cyclic loading.*

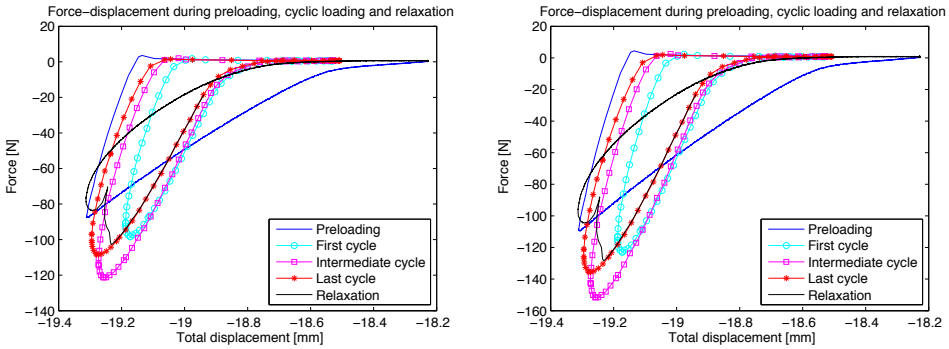


Figure 6. *Hysteresis cycle. The curves represent the response of the sample, force in function of the total displacement (during the test, preloading, cyclic loading 5 [mn] and relaxation). $\phi = 6\%$, $f = 0.5$ [Hz] (left) and $f = 1$ [Hz] (right).*

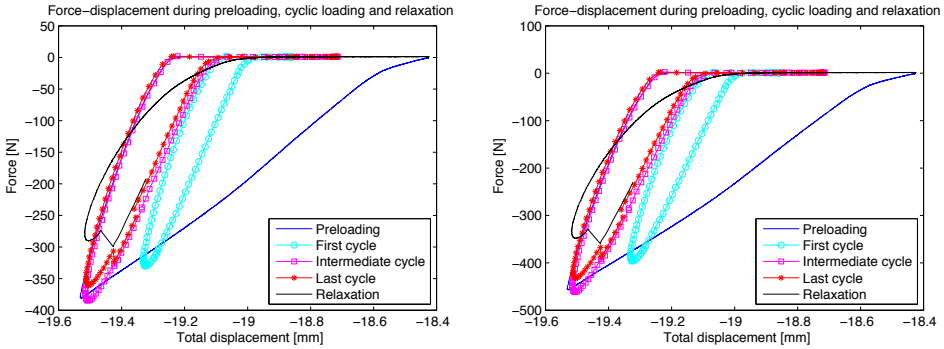


Figure 7. *Hysteresis cycle. The curves represent the response of the sample, force in function of the total displacement (during the test, preloading, cyclic loading 5 [mn] and relaxation). $\phi = 8\%$, for $f = 0.5$ [Hz] (left) and $f = 1$ [Hz] (right).*

We also illustrate the variation of the hydrogel dissipation in function of the cross-link density and the frequency. Without surprise, it can be seen in figures 6 and 7 that

the dissipation is function of the cross-link density and the frequency of loading as for the temperature evolution. More interestingly, we can also observe from this figure that the shape of the hysteresis curves depends on the number of loading cycles. For the same sample under the same loading condition, the shape of the hysteresis curves is completely different if we consider the first, the intermediate or the last cycles. As there is a direct correspondence between the number of cycles and the corresponding temperature in the sample (through the temperature evolution presented in figure 2 (for example), we can deduce that the dissipation is then also function of the temperature.

Indeed, a closer look to the Figures 6 and 7 highlights that the behavior of the hydrogel presents a shift between elastic, viscoelastic and again elastic behaviors at two critical temperatures. This unexpected (and to the best of our knowledge not reported before) behavior was observed for all tested samples.

8. Concluding remarks

In this paper a combined analytical-numerical-experimental approach was developed to evaluate the self-heating phenomenon in a specific hydrogel. The proposed methods are general enough to be used to characterize other types of materials. We demonstrate in this study that the developed model could adequately describe the self-heating behavior of the hydrogel. The influence of two main parameters (cross-link density and loading frequency) on the temperature evolution could also be taken into account in the model. We have to mention that the ranges of the frequency in this work were limited to 0.1-2 Hz for the numerical approaches and to 0.5-1.5 Hz for the experimental measurements. The cross-link density of the hydrogel was limited to 6% and 8% and the percentage in water is prescribed to 40%.

From the experimental data, it has been observed that the hysteresis characterizing the dissipation through the loop force-displacement during the harmonic loading changes its shape in function of the cycle numbers. Two phenomena could be taken into account to explain this observation. First, we can consider that during the loading, the internal structure of the hydrogel changes adapting its structure to the loading. The second phenomenon, which could explain the change of the hysteresis curve over time, is the change in temperature of the self-heating hydrogel. As the number of cycles increases so do the hydrogel temperature. It can then be considered that the increase of temperature changes the mechanical parameters of the hydrogel. For example, in the situation where the elastic parameters would increase with the temperature, as the same displacement was experimentally imposed on the hydrogel, an increase mechanical energy will then be transmitted to the hydrogel.

In general, the developed model could be useful in the phase of design of the hydrogel for a particular application. For example, with the idea of using this kind of dissipative hydrogel for the controlled delivery of a drug through the temperature increase , a link has to be established between the number of cycles and the targeted temperature increase. The developed model would then be useful in this situation to determine the cross-link density needed and/or the mechanical loading regime that the hydrogel should be exposed to. In another application, it has been shown that the toughness of the hydrogel could be increased by increasing its dissipative properties . Again in this situation, the developed model could be used to design the most dissipative hydrogel under known mechanical conditions.

Acknowledgment

Financial supports by the International Doctoral College (CDI) of the Brittany European University (UEB), the Brittany Region Council (France) and the Laboratory of Biomechanical Orthopedics (Lausanne, Switzerland) are greatly appreciated.

9. Bibliographie

[1] A.M. LOWMAN, N.A. PEPPAS, « Hydrogels,Encyclopedia of Drug Delivery », *John Wiley& Sons*, 1999.

[2] N.A. PEPPAS, « Hydrogels in medicine », *Boca Raton, CRC Press*, 1987.

[3] A.S. HOFFMAN , « Hydrogels for biomedical applications, », *Advanced Drug Delivery Reviews*, n° 60, 2002.

[4] M. NASSAJIAN MOGHADAM, V. KASELOV, A. VOGEL, H-A. KLOK « Controlled release from a mechanically-stimulated thermosensitive self-heating composite hydrogel, », *Biomaterials*, n° 35, 2014.

[5] P. ABDEL-SAYED, M. NASSAJIAN MOGHADAM, R. SALOMIR, D. TCHERNIN, D. PIOLETTI « Intrinsic viscoelasticity increases temperatures in knee cartilage under physiological loading, », *Journal of the mechanical behaviour of biomedical materials*, n° 30, 2014.

[6] L. RAKOTOMANANA, D. PIOLETTI, « Non-linear viscoelastic laws for soft biological tissues », *Eur. J. A/Solids*, n° 19, 2000.

[7] TRUESDELL, COLLEMAN, NOLL « The non-linear theories of mechanics », *Springer*, 1992.

[8] M. NASSAJIAN MOGHADAM, D. PIOLETTI, « Improving hydrogels toughness by increasing the dissipative properties of their network », *J Mech Behav. Biomed. Mat.*, n° 41, 2015.

[9] L. RAKOTOMANANA, « Elément de dynamique des solides et structures déformables », *Presses Polytechniques et Universitaires Romandes*, 2009.

[10] R. ZHENG, P. KENNEDY, N. PHAN-THIEN, X-J. FAN, « Thermoviscoelastic simulation of thermally and pressure-induced stresses in injection moulding for the prediction of shrinkage and warpage for fiber-renforced thermoplastics », *Journal of Non-Newtonian Fluid Mechanics*, n° 84, 1999.

[11] C. MORIN, Z. MOUMNI, W. ZAKI « Thermomechanical coupling in shape memory alloys under cyclic loadings : Experimental analysis and consitutive modeling », *Journal maths pures Applied*, n° 27 2011.

[12] N. SANTATRINIAINA, J. DESEURE, T.Q. NGUYEN, T.Q. NGUYEN, H. FONTAINE, C. BEITIA, L. RAKOTOMANANA « Coupled system of PDEs to predict the sensitivity of some materials constituents of FOUP with the AMCs cross–contamination », *International Journal of Applied Mathematical Research*, n° 3 2014.

[13] N. SANTATRINIAINA, J. DESEURE, T.Q. NGUYEN, T.Q. NGUYEN, H. FONTAINE, C. BEITIA, L. RAKOTOMANANA « Mathematical modeling of the AMCs cross–contamination removal in the FOUPs : Finite element formulation and application in the FOUP's decontamination », *International Journal of Mathematical, Computational Science Engineering*, n° 8 2014.

[14] K. KUNISCH , G. LEUGERING , G. LEUGERING , J.SPREKELS , T. FREDI « Optimal Control of the Coupled Systems of Partial Differential Equations », *International Series of Numerical Mathematics*, n° 158, 2009.

Arima

Modélisation en dynamique des populations

Impacts des changements climatiques sur les populations de tiques

Leila Khouaja * — Slimane Ben Miled** — Hassan Hbid***

* ENIT-LAMSIN
Université de Tunis el Manar, BP 37, 1002 Tunis, Tunisie
leilakhouaja85@gmail.com
** ENIT-LAMSIN & Institut Pasteur de Tunis
Université de Tunis el Manar, BP 37, 1002 Tunis, Tunisie
slimane.benmiled@pasteur.rns.tn
*** Faculté des Sciences Semlalia
Université Cadi Ayyad, BP 2390 Marrakech, Maroc
hbid@ucam.ac.ma

ABSTRACT. Epidemiology had an important development these last years allowing the resolution of a large number of problems and had good prediction on disease evolution. However, the transmission of several vector-borne diseases is closely connected to environmental protagonists, specially in the parasite-host interaction. Moreover, understanding the disease transmission is related to studying the ecology of all protagonists. These two levels of complexity(epidemiology and ecology) cannot be separated and have to be studied as a whole in a systematic way. Our goal is to understand the interaction of climate change on the evolution of a disease when the vector has ecological niche that depends on physiological state of development. We are particularly interested in tick vector diseases which are serious health problem affecting humans as well as domestic animals in many parts of the world. These infections are transmitted through a bite of an infected tick, and it appears that most of these infections are widely present in some wildlife species.

RÉSUMÉ. L'épidémiologie a connu un développement important ces dernières années. Cette discipline a permis une meilleure compréhension de l'évolution de maladies. Cependant, plusieurs maladies à transmission vectorielle sont étroitement liées aux protagonistes environnementaux. Ce constat est particulièrement vrai dans le contexte des interactions du parasite avec son hôte. De plus, comprendre la transmission de maladie est lié à l'étude de l'écologie de tous les protagonistes. Notre objectif est de comprendre l'influence du changement climatique sur l'évolution des maladies lorsque la niche écologique du vecteur dépend de l'état de développement physiologique de son hôte. Nous sommes particulièrement intéressés par les maladies vectorielles à tiques qui constituent un grave problème de santé touchant l'être humain et les animaux domestiques dans de nombreuses régions du monde. Ces infections sont généralement transmises par la piqûre d'une tique infectée et il apparaît que la plupart de ces infections sont largement présentées dans certaines espèces fauniques.

KEYWORDS : Epidemiology, McKendrick-Von Foerster equation, Partial differential equation(PDE), Transport equation

MOTS-CLÉS : Epidemologie, Equation de McKendrick-Von Foerster, Equations aux dérivées partielles(EDP), Equation de transport

1. Introduction

Tick-borne diseases (theileriosis, rickettsiosis, Lyme disease, Ehrlichiosis, relapsing fever, TBE(tick-borne encephalitis)) are serious health problem affecting humans as well as domestic animals in many parts of the world. These infections are generally transmitted through a bite of an infected tick, and it appears that most of these infections are widely present in some wildlife species; hence, an understanding of tick population dynamics and its interaction with hosts is essential to understand and control such diseases [6]. For example, the vector of tropical theileriosis in North Africa, the tick Hyalomma detritum, has seasonal activity, while Hyalomma anatolicum is active throughout the year in several parts of Africa and Asia leading to animals being challenged with infection all over the year, this provides a solid immunity during the year contrasting with a very high infection leading to possible endemic stability.

The object of the present work is to develop a tick-borne biology model specific to Hyalomma detritum species in Tunisia. The model will be fitted to field data that have been previously gathered from several Tunisian farms [1].

Our ultimate goal in this paper is to construct models in order to study :

– Epidemiology: The effect of climate change on the evolution of tick-borne diseases particularly Theileriosis.

– Ecological question: What is the most important fact of tick life cycle regulation: Seasonality vs food.

– Control result: The effect of different control actions on tick population.

In order to achieve our goal, we need to solve the two following steps:

1) Modeling the tick life cycle, taking account of temperature fluctuation and seasonality: In this part our objective is to model tick life cycle in order to study the effect of temperature and seasonality on density of the ticks. The model used here will be a partial differentiable equation. The model will be tested using the data from [12] that have been previously gathered from several Tunisian farms [1]. This model will be the foundation of the late epidemiological model.

2) Integration the preceding model of tick life cycle into an epidemiological model:

- Tick: SI model and host: SIR model.

Our work is organized as follows: in the next section, we describe the biology of tick population and present the epidemiological interactions between ticks and their hosts. In section three, we describe both the tick life cycle and its mathematical models; introduce the model which represents the host-parasite epidemiological interaction. Section four is devoted to the conclusion and recommendation.

2. Biological Model

Several field observations on tick biology show a huge polymorphism in their biology (prolificity, mortality, phenology). This polymorphism is enhanced during the parasitic stages of the tick (during feeding stages) because of the interaction between the tick and the host (immunity of the host, surface of exposure, biology of the host). This degree of interaction is again more complicated when the tick-borne infections are considered. Describing this biology of the tick is possible by monitoring infested animals and questing instars and presenting the observations as descriptive results. Nevertheless, understanding

and predicting the mechanisms leading to a determined phenology is quite impossible. Moreover, the prediction of the impact of different control actions is difficult. Modeling represents a powerful tool offering the opportunity to counter account these difficulties. It is possible to model in silico both tick dynamic and the impact of different control options before implementing them, offering then a dramatic decrease of the control costs.

Mathematical modeling represents a powerful tool offering the opportunity to avoid these difficulties. Indeed, it is possible to model in silico both tick-host and epidemiological interactions in order to investigate and understand climate change on disease evolution. Moreover, modeling offer tools to test impact of different control options before implementing them, offering then a dramatic decrease of the control costs.

2.1. Effect of vector life cycle on disease transmission

The tick life cycle includes three post-embryonic developmental stages: larva, nymph and adult. Each stage can be subdivided in turn according to the activity phases: 'questing', in which the unfed tick seeks a host and 'feeding', in which the attached tick feeds, becomes engorged and drops off. After dropping off their hosts, *the cattle*, ticks go through a period of development, after which they emerge as questing ticks at the next stage (or eggs hatch, if the feeding ticks are adult females). The transition from one stage to an other depends closely on the successful questing period that depends on host density. Moreover, the physiological development depends on temperature fluctuations . These two phenomenons are strictly connected to climate change. Indeed, on one hand, it is evident that temperature fluctuation depends on the climate change and on the other hand cattle populations are strictly connected to the agricultural habit which depends on environment.

A variety of approaches have been used to model the tick population with various degrees of complexity. Models often describe in a discrete way the various stages of tick development from egg-larvae-nymph-adult, whether the ticks are attached to hosts, and if disease is part of the model, whether the ticks themselves are infected [15], [16].

Therefore, we propose in this paper two kind of models. The first model is a system of ordinary differential equations with delay where physiological structure is described in a discrete form. This time delay cannot be ignored because the development of the ticks between stages takes time. Moreover, the time delay depends on the weather and climate situation. For this first model our aim is to model tick life cycle in order to study the effect of temperature and seasonality on ticks density. For the second model, the transition from one physiological stage to an other is considered as a continuous process. In this case, we propose to build a PDE model where tick population density satisfies the McKendrick-Von Foerster model with or without blood meal as a limiting factor. All models constructed will be tested using the data from [12], data that have been previously gathered from several Tunisian farms [1] and several data from laboratory colonies. These models will be the foundation of the previous epidemiological model.

2.2. Host - Tick epidemiological interaction

The infection transmission is incorporated into models by adding more states to record the infected status of the ticks and hosts. Typically a mass action law assumption is adopted by the rate of new infections which is directly proportional to the product of susceptible hosts and infectious ticks. However, if larval and nymphs bites are statistically independent, then such clustering would tend to reduce tick and host infection prevalence. In this sense, as positive co-variance of larval and nymphal bites would tend to increase

infection prevalence, as larval bites would be clustered on the host individuals most likely to be infected and infective. An alternative approach to explicitly modeling the host and tick populations was provided by [4] who instead consider the life cycle of the Theileria parva parasite as it progresses through the vertebrate and tick hosts and estimates the time in days (from infection) of disease characteristics in cattle considering challenges from different numbers of infective ticks. It is often assumed that infected ticks behave in the same way as uninfected ones with the mortality of ticks being independent of their infection status. Although as has been seen a pathogen may have a negative impact on the tick in the same manner as a host. Generally models do not consider non-systemic infection (see above) although in a study by [16] this possibility was introduced. However trans-ovarian infection is usually excluded due to the lack of evidence for this in the literature. Reservoir decay or host turnover might enhance positive feedback of infection transmission, for example an increasing prevalence of infected nymphs would increase the frequency at which hosts are re-infected, keeping hosts in a state of high specific infective with a greater probability of infecting the next generation of ticks.

Infection is a one-way through the tick vectors, larvae/nymphae can transmit (trans-stadially) to the hosts of the adults they become, and adults can transmit (trans-ovally) to hosts of the larvae/nymphae that they become.There are two basic frameworks: those who treat the tick density as a parameter, and those who include the processes determining the density of ticks.

3. Implementation of models

3.1. Tick life cycle models

A structured population model is a summary of rules specifying how the number and distribution of individuals within a population changes over time [17]. Most structured population models fall into one of three categories: matrix models, ordinary differential equation (ODE) models, and partial differential equation (PDE) models. In this classification, model type is determined by whether time is discrete (matrix) or continuous (ODE, PDE), and whether the individual-level state is treated as a discrete (matrix, ODE) or a continuous (PDE) variable.

Matrix projection models are popular, because they have relatively simple structure and provide useful information. The eigenvalues and eigenvectors of the projection matrix provide estimates of the population growth rate, the stable age or stage distribution, reproductive value, and the sensitivities of population growth rate to changes in life history parameters [2]. However, whenever a matrix projection or ODE model is applied to population characterized by a continuous state variable (e.g., age, mass, or physiological stage), individuals must be divided into a discrete set of classes.

In partial differential equation models, the individual-level state variables are continuous, and individuals are not lumped into categories. Like the matrix models, PDE models can incorporate a variety of biological situations, including density dependence, and stage- or age-structured populations [7], that's why the basic model structure is the same in all cases. Furthermore, tick population dynamics can be expressed by the McKendrick-Von Foerster equation which is based on partial differential equations (PDE).

As we announced previously, our first objective is to construct a physiological stage dependent PDE model for the tick population dynamics. This model will be in order to fit

to field data from Tunisia that have been previously gathered from several Tunisian farms and several data from laboratory colonies.

As a second step we will investigate the effects of climate on geographic range and seasonality of the tick and compare our results with the ones in [9].

3.2. Parameter definition

We denote by s the tick physiological parameter and t the time parameter and suppose that host populations are fixed at given densities H. To understand the relationship between our PDE model and classical ODE model: we use a physiological parameters s and interstadal development rate, g and let us define s_{egg}^{max}, s_{larvae}^{max}, s_{nymph}^{max},s_{adult}^{max}, the maximum length in the eggs, larvae, nymph and adult class.

To properly model the tick population the rates of tick mortality, reproduction rate (egg-laying) K, and interstadal development rate, g, must be obtained, while to prevent the tick population exponentially increasing issues regarding density dependence should be addressed.

3.3. Mathematical Models

We describe here the mathematical models that we propose to study.

3.3.1. Model 1

The functional equation considered in this model is derived from a physiological-structured model for a population divided into several stages in which individuals change their stage when a certain magnitude reaches a predetermined threshold value. This means that the physiological parameter s of passing from one stage to the next is time-dependent, giving rise to a moving boundary. More details can be found in [10].

To illustrate the ideas underlying the model, consider a population divided into two stages, larvae (l) and adults (L), each one being structured by the age in the stage.

Denote by $l(s,t)$ the density of larvae, $n(s,t)$ the density of nymph and $a(s,t)$ the density of adult at time t and in physiological state s. Capitals, L, N and A, denote the total population of larvae, nymph and adult respectively at time t.

3.3.1.1. Transition from larvae to nymph stage

Let us describe the passage through the larvae stage. We assume that the larvae turn adult when some variable reaches a prescribed value. For example, in [10] the passage to (n) is described in terms of a blood meal which can be measured by *weight function of larvae* $w_l(s,t)$ representing the quantity of blood eaten until time t by an individual until reaches stage s. Larvae turn nymph when the food index reaches a prescribed value $Q_l > 0$. We also assume that there is a finite maximum age $s_{larvae}^{max} > 0$ for individuals in the larval stage: individuals which have not acquired the amount Q of food past $s_{larvae}^{max} > 0$ will die or never reach the nymph stage.

In the model considered in [10], the weight function of larvae depends on the total population of larvae, so that:

$$w_l(s,t) := \int_{t-a}^{t} \frac{K_l}{L(\sigma) + J_l}\, d\sigma; \qquad L(t) := \int_0^{s_{larvae}^{max}} l(s,t)\, ds \tag{1}$$

which means that the quantity of food available is shared in equal parts by all the individuals occupying the same space at time t. $K_l > 0$ is the quantity of food entering the species habitat per unit of volume and per unit of time, which for simplicity is considered

to be constant. The constant $J > 0$ represents the food (converted into a number of individuals) taken per unit of volume by consumers other than larvae.

Then, the age of passage to the (l) stage, denoted by $s_l^*(t)$, is defined by the *threshold condition*:

$$w_l(s,t) = Q \tag{2}$$

so that

$$s_l^*(t) = \begin{cases} s(t) & \text{solution to (2), if it exists and satisfies } \quad 0 \leq s(t) \leq s_{larvae}^{max}, \\ s_{larva}^{max} & \text{otherwise.} \end{cases}$$

Bearing in mind the above considerations, the density of larvae $l(s,t)$ satisfies the following model:

$$\begin{cases} \dfrac{\partial l}{\partial t}(s,t) + \dfrac{\partial}{\partial s}\left(g_l(s,t)l(s,t)\right) = -\mu_l(s)l(s,t), \quad 0 < s < s_l^*(t),\ t > 0, \\ l(s,t) = 0, \quad s_l^*(t) \leq s \leq s_{larvae}^{max}, \quad t > 0, \\ l(s,0) = 0, \quad 0 \leq s \leq s_{larvae}^{max}, \\ l(0,t) = B(t), \quad t > 0, \end{cases}$$

where $\mu_l(s)$ is the age-dependent mortality rate of larvae, the initial condition expresses the fact that at time $t = 0$ no individuals are in the (l) stage and $B(t)$ is the recruitment of larvae at time t. We will assume that μ_l is a nonnegative continuous function on $[0, s_{larva}^{max})$ such that

1) $l(s,t)$ is the density of larvae that have absorbed a quantity of blood s at time t.

2) $n(s,t)$ is the density of nymphs that have absorbed a quantity of blood s at time t.

3) $a(s,t)$ is the density of adults that have absorbed a quantity of blood s at time t.

The tick population dynamic is given by the following system which is composed by three PDE:

Equation of larvae:

$$\begin{cases} \dfrac{\partial l(s,t)}{\partial t} & + & \dfrac{\partial}{\partial s}\left(g_l(s,t)l(s,t)\right) = -\mu_l(s)l(s,t), \\ l(0,t) & = & \lambda \displaystyle\int_0^{s_l^{max}} B(\sigma)a(\sigma,t)d\sigma, \\ l(s,0) & = & \psi_l(s). \end{cases} \tag{3}$$

Equation of nymphs

$$\begin{cases} \dfrac{\partial n(s,t)}{\partial t} & + & \dfrac{\partial}{\partial s}\left(g_n(s,t)n(s,t)\right) = -\mu_n(s)n(s,t), \\ n(0,t) & = & \delta \displaystyle\int_{t-\bar{s_n}}^{t} l(s_n^*(\sigma),\sigma)d\sigma, \\ l(s,0) & = & \psi_n(s), \end{cases} \tag{4}$$

with $\bar{s_n} = \sup_{[0,t]} s_n^*(t)$.

Dynamic equation of adults:

$$\begin{cases} \dfrac{\partial a(s,t)}{\partial t} & + & \dfrac{\partial}{\partial s}\left(g_a(s,t)a(s,t)\right) = -\mu_a(s)a(s,t), \\ n(0,t) & = & \gamma \displaystyle\int_{t-\bar{s_a}}^{t} n(s_a^*(\sigma),\sigma)d\sigma, \\ l(s,0) & = & \psi_a(s), \end{cases} \tag{5}$$

with $\bar{s_a} = \sup_{[0,t]} s_a^*(t)$.

- The functions g_l, g_n and g_a are the growth of blood's quantity which have absorbed by ticks at stages l, n and a respectively.

3.3.2. Model 2

The tick population density varies satisfying the following model for all $t \in [0, \mathbb{T}]$ and $s \in [s_{min}, s_{max}]$ given by

$$\begin{cases} \dfrac{\partial n(s,t)}{\partial t} & + & \dfrac{\partial}{\partial s}\left(g(s,t)n(s,t)\right) = -\mu(n(s,t))n(s,t), \\ n(s_{min},t) & = & \displaystyle\int_{s_{min}}^{s_{max}} K(n(s,t))n(s,t)\mathrm{d}s, \\ n(s,0) & = & n_0(s), \end{cases}$$

with s_{min} and s_{max} the min and the max physiological stage of the tick life cycle and $\mathbb{T}$ the maximum study time.

We are going to work on a set of differentiable C^1 periodic functions on t, and C^1 non-negative functions on s such that $n(s_{max}, t) = 0$ for all times t.

On this basis, since we are dealing with C^1 non-negative functions on s, we may consider that $a(s)$ and $b(s)$, functions appearing in the somatic growth rate $g(s,t)$ as non-negative functions of s so that $g(T) \leq a(s)$ for all $s \in [s_{min}, s_{max}]$ and $t \in [0, \mathbb{T}]$. Moreover, we may suppose that $a(s)$ is a bounded function i.e there exists a constant $\mathcal{C} > 0$ such that

$$a(s) \leq \mathcal{C}, \qquad \forall s \in [s_{min}, s_{max}].$$

Thus,

$$g(T) \leq \mathcal{C}, \qquad \forall s \in [s_{min}, s_{max}], \qquad \forall t \in [0, \mathbb{T}].$$

Also let $\mathcal{N}$ be the maximum tick population density so that

$$n(s,t) \leq \mathcal{N}, \qquad \forall s \in [s_{min}, s_{max}], \qquad \forall t \in [0, \mathbb{T}].$$

3.3.3. Model 3

The following model aims to analyze the impact of climate change on life cycle tick and especially on hibernation period. Let $x_1(t)$ be the density of larvae at time$t$, $x_2(t)$ the density of nymphs after hibernation at time t and $x_3(t)$ the density of adults at time t.

We suppose that $T(t)$ the temperature at time t, $\rho(T)$ the fertility rate of adult females at time t and $\mu(t,T)$ the mortality rate. Thus, the system equations are written in the following way:

$$\begin{cases} \frac{dx_1(t)}{dt} & = & \displaystyle\int_{t-\tau}^{t} \rho(T(\sigma))x_3(\sigma)d\sigma - \mu_1(t,T(t))x_1(t) - f_1(t,T,H,x_1), \\ \frac{dx_2(t)}{dt} & = & \lambda_1\alpha_1(t,T,H,x_1(t-r(t,T(t))))x_1(t-r(t,T(t))) - \mu_2(t,T(t))x_2(t) - f_2(t,T,H,x_2(t)), \\ \frac{dx_3(t)}{dt} & = & \lambda_2\alpha_2(t,T,H,x_2(t))x_2(t) - \mu_3(t,T(t))x_3(t). \end{cases}$$

With

– $f_1(t, T, H, x_1) = \alpha_1(t, T, H, x_1(t))x_1(t)$.
– $f_2(t, T, H, x_1) = \alpha_2(t, T, H, x_2(t))x_2(t)$.

3.4. Physiological SIS for ticks and SIR ODE for host Model

We consider that the number of tick is governed by the equation of section 3.2 and that the tick are subdivided in two class Susceptible and Infected. Let $n_T^s(t)$ and $n_T^i(t)$ be respectively the frequencies of susceptible and infected tick parasites, i.e.,

$$n_T^s(t) + n_T^i(t) = P(t).$$

But

$$P(t) = \int_{a_{min}}^{a_{max}} (l(a,t) + n(a,t) + a(a,t))da, \qquad \forall a \in [a_{min}, a_{max}],$$

where $P(t)$ is total population of tick parasites, a_{min} and a_{max} are the min and the max physiological age of the tick life cycle.

And that the host population number is constant, let $n_H^s(t)$, $n_H^i(t)$ and $n_H^r(t)$ be respectively the frequencies of susceptible, infected and removed host population, i.e. $n_H^s(t) + n_H^i(t) + n_H^r(t) = 1$.

Consider the following assumptions:

- Let $\phi(n)$ be the factor representing the influence of ticks on the host. Thus the model representing the host-parasite epidemiological interaction is given by the following SIR-SIS model:

$$\frac{\mathrm{d}n_H^s(t)}{\mathrm{d}t} = -K_1 n_H^s(t) n_H^i(t) \tag{6}$$

$$\frac{\mathrm{d}n_H^i(t)}{\mathrm{d}t} = K_1 n_H^s(t) n_H^i(t) - K_2 n_T^i(t) n_H^i(t) \tag{7}$$

$$\frac{\mathrm{d}n_H^r(t)}{\mathrm{d}t} = K_2 n_T^i(t) n_H^i(t) \tag{8}$$

$$\frac{\mathrm{d}n_T^i(t)}{\mathrm{d}t} = K_1 n_H^s(t) n_H^i(t) - \phi(n) K_2 n_T^i(t) n_H^i(t) + n_T^s(t) \tag{9}$$

$$n_T^s(t) = 1 - n_T^i(t) \tag{10}$$

4. Conclusion and Recommendation

In this paper, we present the various stages of tick population dynamics which is composed by three partial differential equations.

Our aim in the future is to adapt the numerical method developed in subsection 3.2 in order to solve the correlated EDO-EDP equations. This method will be used to test impact of climate change on transmission of Tick disease, for that the model will be fitted to field data that have been previously gathered from several Tunisian farms [1] and several data from laboratory colonies (Darghouth, unpublished data) and with data from Tunisian National Institute of Meteorology. Furthermore, our goal is to develop new numerical methods in order to approximate solutions of the previous type of equations in subsection 3.3.2.

5. References

[1] A. BOUATTOUR, M. A. DARGHOUTH, L. BEN MILED. "Cattle infestation by Hyalomma ticks and prevalence of Theileria in H. detritum species in Tunisia", *Vet. Parasitol.*, vol. 65:233-245, 1996.

[2] H. CASWELL, "Matrix Population Models: construction, analysis, and interpretation", *Sinauer Associates Inc. Sunderland, Massachusetts*, 2001.

[3] HOPE-CAWDERY, M.J., GETTINBY, G. AND GRAINGER, J.N.R., "Mathematical models for predicting the prevalence of liver fluke disease and its control from biological and meteorological data. In: T.E. Gibson (Editor), Weather and Parasitic Animal Disease", *World Meteorological Organisation Technical Note*, 159:21-38, 1978.

[4] W P GARDINER, G GETTINBY, AND J S GRAY "Modes based on weather for the development phases of the sheep tick, ixodes-ricinus", *Veterinary Parasitology*, 9(1):75-86, 1981.

[4] G GETTINBY, W BYROM "The dynamics of east coast fever: A modelling perspective for the integration of knowledge", *Parasitology Today*, 5(3):68-73, 1989.

[5] W S C GURNEY, R M NISBET, N GURNEY, "Ecological dynamics", *Oxford University Press New York*, 1998.

[6] PETER J HUDSON, ANDY P DOBSON, ISABELLA M CATTADORI, DAVID NEWBORN, DAN T HAYDON, DARREN J SHAW, TIM G BENTON , BRYAN T GRENFELL, "Trophic interactions and population growth rates: describing patterns and identifying mechanisms ", *Philos Trans R Soc Lond B Biol Sci*, 357(1425):1259-1271, September 2002.

[7] J A J METZ, O DIEKMANN, "The dynamics of physiologically structured populations", 1986.

[8] R NORMAN, R G BOWERS, M BEGON, P J HUDSON, "Persistence of tick-borne virus in the presence of multiple host species: tick reservoirs and parasite mediated competition. *J Theor Biol*, 200(1):111-118, September 1999.

[9] N H OGDEN, M BIGRAS-POULIN, C J O'CALLAGHAN, I K BARKER, L R LINDSAY, A MAAROUF, K E SMOYER-TOMIC, D WALTNER-TOEWS, D CHARRON "A dynamic population model to investigate effects of climate on geographic range and seasonality of the tick ixodes scapularis", *Int J Parasitol*, 35(4):375-389, 2005.

[10] O.ARINO, M.L.HBID, R.BRAVO DE LA PARRA, "A Mathematical model of population of fish in the larval stage: density-dependence effects", *Math.Biosci*, 150,1-20, 1998.

[11] S RANDOLPH, "Epidemiological uses of a population model for the tick rhipicephalus appendiculatus", *Trop Med Int Health*, 4(9):A34-A42, September 1999.

[12] S E RANDOLPH, "Tick ecology: processes and patterns behind the epidemiological risk posed by ixodid ticks as vectors ", *Parasitology*, 129 Suppl:S37-S65, 2004.

[13] S E RANDOLPH, "Dynamics of tick-borne disease systems: minor role of recent climate change", *Rev Sci Tech*, 27(2):367-381, 2008.

[14] SARAH E RANDOLPH, "Abiotic and biotic determinants of the seasonal dynamics of the tick rhipicephalus appendiculatus in south africa", *Medical and Veterinary Entomology*, 11(1):25-37, 1997.

[15] R ROSÀ, A PUGLIESE, R NORMAN, P J HUDSON, "Thresholds for disease persistence in models for tick-borne infections including non-viraemic transmission, extended feeding and tick aggregation", *Journal of theoretical biology*, 224(3):359-376, 2003.

[16] R ROSÀ, ROBERTO ROSÀ, ANDREA PUGLIESE, "Effects of tick population dynamics and host den- sities on the persistence of tick-borne infections", *Mathematical Biosciences*, 208(1):216-240, 2007.

[17] S TULJAPURKAR, H CASWELL, " Structured-population models in marine, terrestrial, and freshwater systems ", *Kluwer Academic Pub*, 1997.

Stochastic modeling of the anaeorobic model AM2b

Models at different scales

F. Campillo[a] — M. Chebbi[b] — S. Toumi[c]

[a] Inria-France
Fabien.Campillo@inria.fr
[b] ENIT-Laboratoire LAMSIN-University Tunis el Manar-Tunisie
mohsenchebbi@ymail.com
[c] INSAT-Carthage University-Tunisie
salwa.toumi@gmail.com

ABSTRACT. The AM2B model is conventionally represented, in large population, as a system of ordinary differential equations . Our goal is to build several models at different scales. At the microscopic scale (the scale of the individual), we propose a pure jump stochastic model. This model can be exactly simulated. However, when the size of the population is large that type of exact simulation is not feasible, hence we propose approximated simulation methods in discrete time, of the Poisson type or of the diffusive type. The diffusive type of approximated simulation method can be seen as a discretization of a stochastic differential equation. Finally, we present informally a law of large numbers/central limit theorem of the functional type and how they can be used to provide models at different scales or hybrid models.

RÉSUMÉ. Le modèle AM2b est classiquement représenté, en grande population, par un système d'équations différentielles. Notre objectif est d'établir plusieurs modèles à différentes échelles. À l'échelle microscopique (individuelle), on propose un modèle stochastique de saut pur. Ce modèle peut être simulé de façon exacte. Lorsque la taille de la population est grande ce genre de simulation n'est pas praticable, et nous proposons des méthodes de simulation, à pas de temps discret, de type poissonnien ou de type diffusive. La méthode de simulation de type diffusive peut être vue comme une discrétisation d'une équation différentielle stochastique. Nous présentons enfin de façon informelle un résultat de type loi des grands nombres/théorème central limite fonctionnelle et comment ce résultat peut conduire à des modèles selon l'échelle considérée ou encore à des modèles hybrides.

KEYWORDS : AM2b model, pure jump process, ordinary differential equation, diffusion approximation, stochastic differential equation

MOTS-CLÉS : modèle AM2b, processus de saut pure, équation différentielle ordinaire, approximation diffusion, équation différentielle stochastique

Stochastic models recently gain more credibility and numerical efficiency in chemistry [7], biotechnology [8], system biology [12] where deterministic models have been extensively used. Taking the example of a biotechnological model, we explain how a stochastic modeling approach deepens the insights allowed by the deterministic classical models.

Wastewater treatment plant aims at reducing the volume of pollutants rejection, producing potential energy like CH_4 in anaerobic treatment, providing treated water for agriculture and industry. Among these technologies anaerobic membrane bioReactors are promising technologies provided that membrane fouling phenomenon could be reduced. AM2b is a mathematical model of anaerobic membrane bioreactors developed by [1, 2], it is a variant of model AM2 (2-steps Acidogenesis-Methanogenesis model, see [3]) with soluble microbial products (SMP) dynamics. The production and the degradation of SMP play an important role in the membrane fouling phenomenon.

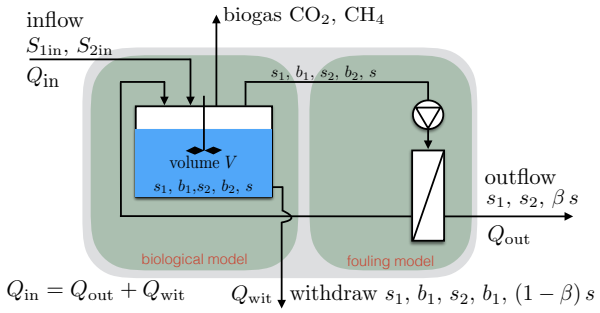
We present the original ODE model of the AM2b. Then we introduce a pure jump Markov model of the same device and a exact Monte Carlo simulation method of this process. Next we propose faster approximated simulation methods. Finally we present a stochastic differential equation (SDE) model of the AM2b. The validity of these models depends on the scale considered for the problem, this could be cleared with a law of large numbers/central limit theorem of the functional type presented in the last section.

1. The ODE model

The state variable of this model is:

$$x \stackrel{\text{def}}{=} \begin{pmatrix} s_1 \\ b_1 \\ s_2 \\ b_2 \\ s \end{pmatrix} \begin{array}{l} \text{concentration of organic matter} \\ \text{concentration of acidogenic biomass} \\ \text{concentration of volatile fatty acids (VFA)} \\ \text{concentration of methanogenic biomass} \\ \text{concentration of the soluble microbial products (SMP)} \end{array}$$

The AM2b model describes the dynamics of biological and anaerobic wastewater treatme
str
in
pr



On the right of these scheme the membrane fouling model is represented; the separation of mater is as follow: the substrates s_1 and s_2 go through the membrane without retention (the size of their molecules is assumed to be smaller than pore diameter), the biomass

b_1 and b_2 is retained by the membrane, and a fraction $\beta\, s$ of the SMP go through the membrane and leaves the reactor ($(1-\beta)\, s$ will me considered as macromolecules).

In Appendix A.1 we describe the AM2b model as a reaction network of $J = 15$ reactions, and in Appendix A.2 we explain how this reaction network can be translated into an ODE system thanks to the laws of mass action and mass conservation. The ODE system reads:

$$\dot{s}_1 = D_{\text{in}}\,(S_{1\text{in}} - s_1) \; - k_1\,\mu_1(s_1)\,b_1\,, \tag{1a}$$
$$\dot{b}_1 = \big(\mu_1(s_1) + \mu(s) - D_{\text{dec}} - D_{\text{wit}}\big)\,b_1\,, \tag{1b}$$
$$\dot{s}_2 = D_{\text{in}}\,(S_{2\text{in}} - s_2) \; - k_2\,\mu_2(s_2)\,b_2 + \big(c_{12}\,\mu_1(s_1) + c_{02}\,\mu(s)\big)\,b_1\,, \tag{1c}$$
$$\dot{b}_2 = \big(\mu_2(s_2) - D_{\text{dec}} - D_{\text{wit}}\big)\,b_2\,, \tag{1d}$$
$$\dot{s} = \big(c_{10}\,\mu_1(s_1) + D_{\text{dec}} - k_0\,\mu(s)\big)\,b_1 + \big(c_{20}\,\mu_2(s_2) + D_{\text{dec}}\big)\,b_2 - M\,s \tag{1e}$$

or

$$\dot{x}(t) = f(x(t))\,,\; x(0) = x_0 \tag{2}$$

where:

- k_i degradation rate of s_i by b_i,
- k_0 degradation rate of s by b_1,
- c_{12} production rate of s_2 by b_1 from s_1,
- c_{02} production rate of s_2 by b_1 from s,
- c_{i0} production rate of s by b_i from s_i,
- $S_{i\text{in}}$ input concentrations of s_i,
- $M = \beta\, D_{\text{out}} + (1-\beta)\, D_{\text{wit}}$
- β SMP fraction passing through the membrane,
- D_{in} dilution rate ($= Q_{\text{in}}/V$),
- D_{dec} decay rate of biomass,
- D_{wit} withdrawal rate of biomass ($= Q_{\text{wit}}/V$),
- D_{out} outflow rate of the bioreactor ($= Q_{\text{out}}/V$)

($i = 1, 2$). To ensure a constant volume we state that $Q_{\text{in}} = Q_{\text{out}} + Q_{\text{wit}}$. The rate M states that a proportion β of the SMP will leave the bioreactor through the membrane, at rate D_{in} and a proportion $1-\beta$ through the withdrawal process, at rate D_{wit}.

The growth functions are:

$$\mu_1(s_1) = m_1\,\tfrac{s_1}{K_1+s_1}\,, \qquad \mu(s) = m\,\tfrac{s}{K+s}\,, \qquad \mu_2(s_2) = m_2\,\tfrac{s_2}{K_2+s_2+s_2^2/K_{\text{i}}}\,. \tag{3}$$

μ_1 and μ are of Monod type; μ_2 is of Haldane type to model the phenomenon of inhibition of the eventual accumulation of the volatile fatty acids in the bioreactor during the methanogenesis (the major problem of the anaerobic digestion).

Model (1) relies on the fact that the stochastic effects can be neglected or at least can be averaged out. Although this level of description is sufficient for a number of applications of interest, it could be a valuable way of accounting for the stochastic nature of the system. Indeed, at small population sizes the AM2b model could present stochastic behaviors. Moreover, whereas the experimental results observed in well-mastered laboratory conditions match closely the ODE theoretical behavior, a noticeable difference may occur in operational conditions. In these cases, stochastic features may not be neglected. We aim to build a model that still relies on a mass balance principle and that encompasses the useful stochastic information.

2. Pure jump Markov model

Following the approach described in [4], we propose a representation of the AM2b model as a pure jump Markov process:

$$X(t) = (S_1(t), B_1(t), S_2(t), B_2(t), S(t))^* \tag{4}$$

taking values in $\mathbb{R}_+^5$. This process will encompass the J reactions: each reaction j is now characterized by it's intensity functions $\lambda_j(x)$ and it's jump functions $\nu_j(x)$, see details in Section A.3. The dynamic of the process $X(t)$ is described as follows: $X(0) = x_0$ and conditionally on $X(t) = x$, we set

$$X(t+\Delta t) = \begin{cases} x + \nu_j(x) & \text{with probability } \lambda_j(x)\,\Delta t + o(\Delta t) \quad 1 \le j \le J \\ x & \text{with probability } 1 - \sum_{j=1}^{J} \lambda_j(x)\,\Delta t + o(\Delta t) \end{cases} \tag{5}$$

where $(\lambda_j(x), \nu_j(x))_{1\le j\le J}$ is given by (12) and (13).

2.1. Simulation and representation of $X(t)$

The process $X(t)$ can be simulated according to the following SSA (stochastic simulation algorithm) [6]:

$$\begin{array}{l} X \leftarrow x_0,\, t \leftarrow 0 \\ \textbf{while } t < T_{\max} \textbf{ do} \\ \quad \tau \leftarrow \sum_{j=1}^{J} \lambda_j(X) \\ \quad S \sim \mathrm{Exp}(\tau) \\ \quad \text{sample } j \text{ according to the distribution } (\frac{\lambda_1(X)}{\tau}, \ldots, \frac{\lambda_J(X)}{\tau}) \\ \quad t \leftarrow t + S \\ \quad X \leftarrow X + \nu_j(X) \\ \textbf{end while} \end{array} \tag{6}$$

This Monte Carlo procedure allows us to simulate *exact* trajectories of the process X_t, the only approximation resides in the algorithms used for simulating the basic probability distributions.

Algorithm (6) is an exact representation of the process $X(t)$, and it leads to the following representation of the process:

$$X_t = X_0 + \sum_{j=1}^{J} \int_{[0,t]\times\mathbb{R}_+} 1_{[0,\lambda_j(X_{s^-})]}(v)\ \nu_j(X_{s^-})\ \mathcal{N}_j(\mathrm{d}s, \mathrm{d}v) \tag{7}$$

where $\mathcal{N}_j(\mathrm{d}s, \mathrm{d}v)$ are independent Poisson random measures of intensity measure $\mathrm{d}s \times \mathrm{d}v$ (the Lebesgue measure on $\mathbb{R}_+^2$).

3. Discrete time approximations

The SSA simulates each reaction of the ecosystem asynchronously in time. In many situations this detailed simulation is too cumbersome, this is why synchronous discrete time approximations have been proposed. Let $t_m = m\,\Delta t$, for $\Delta t > 0$ fixed.

Poisson approximation

We construct an approximation $(\tilde{X}(t_m))_{m\ge1}$. On the interval $[t_m, t_{m+1})$ suppose that the different rate functions are approximated by:

$$\lambda_j(x) \simeq \lambda_j(\tilde{X}(t_m))\,,\ \forall x \in \mathbb{R}_+^5$$

so that each of the J reactions are independent and occur at constant rates $\lambda_j(\tilde{X}(t_m))$, that is the number of reactions of type j is a Poisson process of intensity $\lambda_j(\tilde{X}(t_m))$.

Hence, on the time interval $[t_m, t_{m+1})$ the number of reactions of type j follows a Poisson distribution of parameter $\Delta t\, \lambda_j(\tilde{X}(t_m))$. We obtain the following approximation also called τ-leaping:

$$\tilde{X}(t_{m+1}) = \Big[\tilde{X}(t_m) + \sum_{j=1}^{J} \nu_j(\tilde{X}(t_m))\, \rho_{j,m}\Big]_+ \tag{8}$$

where $\rho_{j,m}$ are independent Poisson distribution variables with parameter $\Delta t\, \lambda_j(\tilde{X}(t_m))$ and $[x]_+$ is the projection of $\mathbb{R}^5$ onto $\mathbb{R}^5_+$ (the positive part of each component).

Diffusion approximation

The Poisson distribution with parameter $\Delta t\, \lambda_j(\tilde{X}(t_m))$, for $\Delta t\, \lambda_j(\tilde{X}(t_m))$ large enough, can be approximated by a Normal distribution of mean $\Delta t\, \lambda_j(\tilde{X}(t_m))$ and variance $\Delta t\, \lambda_j(\tilde{X}(t_m))$. From (8) we get:

$$\tilde{\xi}(t_{m+1}) = \Big[\tilde{\xi}(t_m) + \sum_{j=1}^{J} \nu_j(\tilde{\xi}(t_m)) \left\{\Delta t\, \lambda_j(\tilde{\xi}(t_m)) + \sqrt{\Delta t\, \lambda_j(\tilde{\xi}(t_m))}\, w_{j,m}\right\}\Big]_+$$

where $w_{j,m}$ are independent $N(0,1)$ random variables. This last equation can be rewritten:

$$\tilde{\xi}(t_{m+1}) = \Big[\tilde{\xi}(t_m) + F(\tilde{\xi}(t_m))\, \Delta t + \sum_{j=1}^{J} \frac{1}{\sqrt{N_j}}\, g_j(\tilde{\xi}(t_m))\ [W_j(t_{m+1}) - W_j(t_m)]\Big]_+ \tag{9}$$

where $W_j(t)$ are independent standard Brownian motions so that $W_j(t_{m+1}) - W_j(t_m)$ are independent and $N(0, \Delta t)$; $F(x)$ is defined in (15) and:

$$g_j(x) \stackrel{\text{def}}{=} \sqrt{N_j}\, \sqrt{\lambda_j(x)}\, \nu_j(x) = \sqrt{\tilde{\lambda}_j(x)}\, N_j\, \nu_j(x)\,.$$

Let

$$\tilde{g}_j(x) \stackrel{\text{def}}{=} \sqrt{\tilde{\lambda}_j(x)}\, \tilde{\nu}_j$$

so that:

$$g_j(x) - \tilde{g}_j(x) = 1_{x\in\mathcal{D}}\, \big(g_j(x) - \tilde{g}_j(x)\big)$$

and $|g_j(x) - \tilde{g}_j(x)| \le C\, 1_{x\in\mathcal{D}}\, \sqrt{1+|x|}$.

Stochastic differential equation

Equation (9), is an Euler-Maruyama discrete time approximation of the following stochastic differential equation (SDE):

$$\mathrm{d}\xi(t) = F(\xi(t))\, \mathrm{d}t + \sum_{j=1}^{J} \frac{1}{\sqrt{N_j}} g_j(\xi(t))\, \mathrm{d}W_j(t)\,,\ \xi(0) = x_0\,. \tag{10}$$

4. Scales and asymptotics

According to Section A.3, the scale parameters N_j are connected to the "size" of the jumps in the reactions ⓙ. We can assume that the F_j's range from 10^4 to 10^9. When a reaction involves only substrate molecules the corresponding F_j's range from 10^7 to 10^9; when a reaction involves only bacteria the corresponding F_j's range from 10^4 to 10^6. Hence for reasonable concentrations, the simulation algorithm (6) will not be feasible as it simulates every single reaction.

First suppose that $N = N_j$ for all j and that N is large. The first well know result can be understood as a functional law of large numbers (originally proved in this context by Tom Kurtz [9, 10]), it states that:

$$\sup_{0\leq t\leq T} |X(t) - x(t)| \xrightarrow[N\to\infty]{} 0 \tag{11}$$

in L^2 or in probability. It is clear that in (11) we can replace $X(t)$ by $\xi(t)$. So under mild conditions, when the population sizes are large and so the number of reaction, the ODE model (2) is adapted to this scale.

At an intermediate scale, a functional central limit theorem states that the process $\sqrt{N}\,(X(t) - x(t))$ can be approximated in law by $\sum_{j=1}^{J} \int_0^t g_j(x(s))\ \mathrm{d}\tilde{W}_j(s)$ where the $\tilde{W}_j(s)$ are independent standard Brownian motions, that is formally:

$$X(t) \simeq x(t) + \frac{1}{\sqrt{N}} \sum_{j=1}^{J} \int_0^t g_j(x(s))\, \mathrm{d}\tilde{W}_j(s)\,.$$

This also proves that the SDE model (10) is adapted to this scale.

In many situation ODE and SDE models are not valid. This is the case when one of the bacterial population is of "small" size but still affects the global dynamic of the process. This so-called "molecular randomness" may influence the global dynamic even when the population sizes are not so small [5]. In this case we may adopt hybride approaches. We just present an example where we separate the dynamics of the substrates from the dynamics of the biomasses: The idea is to break down the reactions between substrate type reactions and biomass type reactions, then to describe the first ones as a system of ODE's and to describe the second ones as a pure jump Markov process. For example we can obtain a system of ODE's describing the continuous evolution of the substrates and the SMP concentrations:

$$\begin{aligned}
\dot{s}_1 &= D_{\text{in}}\,(S_{1\text{in}} - s_1)\ - k_1\,\mu_1(s_1)\,B_1\,,\\
\dot{s}_2 &= D_{\text{in}}\,(S_{2\text{in}} - s_2)\ - k_2\,\mu_2(s_2)\,B_2 + \big(c_{12}\,\mu_1(s_1) + c_{02}\,\mu(s)\big)\,B_1\,,\\
\dot{s} &= \big(c_{10}\,\mu_1(s_1) + D_{\text{dec}} - k_0\,\mu(s)\big)\,B_1 + \big(c_{20}\,\mu_2(s_2) + D_{\text{dec}}\big)\,B_2 - M\,s
\end{aligned}$$

coupled to a 2-dimensional pure jump process describing the discrete evolution of the biomasses concentrations:

jump	rate	jump	rate
$B_1 \to B_1 + \delta_1$	$\mu_1(s_1)\,B_1/\delta_1$	$B_2 \to B_2 + \delta_2$	$\mu_2(s_2)\,B_1/\delta_1$
$B_1 \to B_1 + \delta_1$	$\mu(s)\,B_1/\delta_1$	$B_2 \to B_2 - \delta_2$	$D_{\text{dec}}\,B_1/\delta_2$
$B_1 \to B_1 - \delta_1$	$D_{\text{dec}}\,B_1/\delta_1$	$B_2 \to B_2 - \delta_2$	$D_{\text{wit}}\,B_1/\delta_2$
$B_1 \to B_1 - \delta_1$	$D_{\text{wit}}\,B_1/\delta_1$		

This type of model, known as *Piecewise-deterministic Markov process*, is very promising and will be investigated in the near future.

5. Conclusion

We show that an ODE model of microbial dynamics (1) contains all the ingredients that can be used to establish a pure jump Markov model, see Section 2. This pure jump Markov model can be exactly simulated with the Monte Carlo technique (6). This exact Monte Carlo method is not feasible in large population size cases and we proposed a Poissonian discrete time approximation (8) (also called τ-leaping) and a diffusion discrete time approximation (9). This last equation is the Euler-Maruyama time discretization of the SDE (10). This SDE is valid in high population size and different recalling so that alternative formulations of this SDE can be established. In Section 4, we describe the validity of these different models according to the scales at which the process should be simulated. All these models share the same ingredient, however the have very different qualitative properties.

Acknowledgments

This work was funded by the NuWat LIRIMA Inria project which funded this work.

6. References

[1] Boumédiène Benyahia, Tewfik Sari, Brahim Cherki and Jérôme Harmand. Anaerobic membrane bioreactor modeling in the presence of Soluble Microbial Products (SMP) – the Anaerobic Model AM2b. *Chemical Engineering Journal*, 228(0):1011–1022, 2013.

[2] Boumédiène Benyahia. *Modélisation et observation des bioprocédés à membranes: application à la digestion anaérobie*. PhD thesis, Université de Tlemcen et Université de Montpellier 2, 2012.

[3] Olivier Bernard, Zakaria Hadj-Sadock, Denis Dochain, Antoine Genovesi, and Jean-Philippe Steyer. Dynamical model development and parameter identification for an anaerobic wastewater treatment process. *Biotechnology and Bioengineering*, 75:424–438, 2001.

[4] Fabien Campillo, Marc Joannides, and Irène Larramendy-Valverde. Stochastic modeling of the chemostat. *Ecological Modelling*, 222(15):2676–2689, 2011.

[5] Fabien Campillo and Claude Lobry. Effect of population size in a Predator-Prey model. *Ecological Modelling*, 246:1–10, 2012.

[6] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[7] Desmond J. Higham. Modeling and simulating chemical reactions. *SIAM Review*, 50(2):347–368, 2008.

[8] Sebastiaan Kops, Krist Gernaey, Olivier Thas and Peter A. Vanrolleghem. The modelling of noise processes in stochastic differential equations: Application to biotechnological processes. In *Proceedings 7th IFAC Conference on Computer Applications in Biotechnology CAB7. Osaka, Japan, May 31 - June 4*, pages 67–72, 1998.

[9] Thomas G. Kurtz. Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of Applied Probability*, 7(1):49–58, 1970.

[10] Thomas G. Kurtz. Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *Journal of Applied Probability*, 8:344–356, 1971.

[11] Mukhtar Ullah and Olaf Wolkenhauer. *Stochastic approaches in systems biology*. Springer, 2011.

[12] Darren J. Wilkinson. *Stochastic Modelling for Systems Biology*. Chapman & Hall/CRC, 2006.

A. Appendix: the AMB2b models

A.1. AMB2b as a reaction network

The first set of reactions describes the biochemical reactions:

$$\begin{array}{rl}
\textit{Acidogenesis and SMP production} & ① \; k_1 \mathrm{S}_1 + \mathrm{B}_1 \xrightarrow{r_1} 2\,\mathrm{B}_1 + c_{12}\,\mathrm{S}_2 + c_{10}\,\mathrm{S} \\
\textit{Methanogenesis and SMP production} & ② \; k_2 \mathrm{S}_2 + \mathrm{B}_2 \xrightarrow{r_2} 2\,\mathrm{B}_2 + c_{20}\,\mathrm{S} \\
\textit{SMP degradation} & ③ \; k_0 \mathrm{S} + \mathrm{B}_1 \xrightarrow{r_3} 2\,\mathrm{B}_1 + c_{02}\,\mathrm{S}_2 \\
\textit{SMP production from biomass decay} & ④ \; \mathrm{B}_1 \xrightarrow{r_4} \mathrm{S} \quad ⑤ \; \mathrm{B}_2 \xrightarrow{r_5} \mathrm{S}
\end{array}$$

The second set of reactions describes the *substrate inflow* and the *substrate outflow through the membrane*:

$$⑥ \; \emptyset \xrightarrow{r_6} \mathrm{S}_1 \quad ⑦ \; \emptyset \xrightarrow{r_7} \mathrm{S}_2 \quad ⑧ \; \mathrm{S}_1 \xrightarrow{r_8} \emptyset \quad ⑨ \; \mathrm{S}_2 \xrightarrow{r_9} \emptyset \quad ⑩ \; \mathrm{S} \xrightarrow{r_{10}} \emptyset$$

The third and last set of reactions describes the *biomass and substrate withdrawal* with :

$$⑪ \; \mathrm{S}_1 \xrightarrow{r_{11}} \emptyset \quad ⑫ \; \mathrm{S}_2 \xrightarrow{r_{12}} \emptyset \quad ⑬ \; \mathrm{S} \xrightarrow{r_{13}} \emptyset \quad ⑭ \; \mathrm{B}_1 \xrightarrow{r_{14}} \emptyset \quad ⑮ \; \mathrm{B}_2 \xrightarrow{r_{15}} \emptyset$$

The second and third set of reactions are not biochemical reactions they just describe the inflows and outflows in the AM2B process. In reaction ⑩ only a proportion β of the SMP goes through the membrane, and in reaction ⑬ a proportion $1-\beta$ of the SMP is withdrawn, this mechanism will be explicated in the definition of the rates. Let j the index of the reaction and $J = 15$ the number of reactions. $\{(ⓙ, r_j); j = 1, \ldots, J\}$

A.2. AMB2b as an ODE system

The AMB2b reaction network described in Section A.1 is "translated" to an ODE system thanks to the laws of mass action and conservation of mass. The state variables of the ODE system are the concentration $s_i = [\mathrm{S}_i]$, $b_i = [\mathrm{B}_i]$ $(i = 1, 2)$ and $s = [\mathrm{S}]$. For example, for the ①, the *rate of reaction* also called *speed of reaction*, is defined by:

$$r_1 = \tfrac{b_1(t+\Delta)-b_1(t)}{\Delta} = -k_1 \tfrac{s_1(t+\Delta)-s_1(t)}{\Delta} = c_{12} \tfrac{s_2(t+\Delta)-s_2(t)}{\Delta} = c_{10} \tfrac{s(t+\Delta)-s(t)}{\Delta}$$

(the equalities are due to the mass conservation). This reaction, like reactions ② and ③, is of order two and the mass action law states that $r_1 = s_1\, b_1$, but saturation/inhibition phenomena suggest to replace in this last expression s_1 by $\mu_1(s_1)$ indeed:

Biochemical reactions: The mass action law applied to the second order reactions ① ② ③ states that:

$$r_1 = \mu_1(s_1)\, b_1\,, \qquad r_2 = \mu_2(s_2)\, b_2\,, \qquad r_3 = \mu(s)\, b_1$$

where the growth functions μ_i and μ are chosen as (3), indeed for low substrat concentration, these growth functions are linear accordingly to the mass action law, but for higher

substrate concentrations, saturation and inhibition phenomena have to be taken into account. The mass action law applied to first order reactions ④ ⑤ gives $r_4 = D_{\text{dec}}\, b_1$ and $r_5 = D_{\text{dec}}\, b_2$ where D_{dec} is decay rate of biomass.

Inflow and outflows: Inflow is done at rate D_{in}, outflow though the membrane at rate D_{out} and only a proportion β of the SMP is affected by the outflow, so the rates of these reactions are:

$$r_6 = D_{\text{in}}\, S_{1\text{in}}\,, \quad r_7 = D_{\text{in}}\, S_{2\text{in}}\,, \quad r_8 = D_{\text{out}}\, s_1\,, \quad r_9 = D_{\text{out}}\, s_2\,, \quad r_{10} = \beta\, D_{\text{out}}\, s\,.$$

Withdrawal: The withdrawal is done at rate D_{wit} and only a proportion $1-\beta$ of the SMP affected by the withdrawal, so the rates of these reactions are:

$$r_{11} = D_{\text{wit}}\, s_1\,, \quad r_{12} = D_{\text{wit}}\, s_2\,, \quad r_{13} = (1-\beta)\, D_{\text{wit}}\; s\,, \quad r_{14} = D_{\text{wit}}\, b_1\,, \quad r_{15} = D_{\text{wit}}\, b_2\,.$$

Summing up these expressions and applying the mass conservation law lead to the system of differential equations (1).

A.3. AMB2b as a pure jump Markov process

The AMB2b reaction network described in Section A.1 is "translated" into a pure jump Markov process thanks to the stochastic law of mass action [12]. Now $X(t) = [S_1(t),\, B_1(t), S_2(t), B_2(t), S(t)]^*$ is a pure jump Markov process defined by (5) where each reaction ② is described as an instantaneous jump $X(t) \to X(t)+\nu_j(X(t))$ occurring with intensity $\lambda_j(X(t))$ defined respectively by:

$$\lambda_j(x) \stackrel{\text{def}}{=} N_j\, \tilde{\lambda}_j(x)\,, \qquad\qquad \nu_j(x) \stackrel{\text{def}}{=} [x + \tfrac{1}{N_j}\, \tilde{\nu}_j]_+ - x \tag{12}$$

($[x]_+$ the orthogonal projection of x onto $\mathbb{R}^5_+$) with

j	$\tilde{\lambda}_j(x)$	$\tilde{\nu}_j^*$				
①	$\mu_1(s_1)\, b_1$	$[-k_1$	$+1$	$+c_{12}$	0	$+c_{10}\,]$
②	$\mu_2(s_2)\, b_2$	$[\;\;0$	0	$-k_2$	$+1$	$+c_{20}\,]$
③	$\mu(s)\, b_1$	$[\;\;0$	$+1$	$+c_{02}$	0	$-k_0\,]$
④	$D_{\text{dec}}\, b_1$	$[\;\;0$	-1	0	0	$+1\,]$
⑤	$D_{\text{dec}}\, b_2$	$[\;\;0$	0	0	-1	$+1\,]$
⑥	$D_{\text{in}}\, S_{1\text{in}}$	$[\;+1$	0	0	0	$0\,]$
⑦	$D_{\text{in}}\, S_{2\text{in}}$	$[\;\;0$	0	$+1$	0	$0\,]$
⑧	$D_{\text{out}}\, s_1$	$[-1$	0	0	0	$0\,]$
⑨	$D_{\text{out}}\, s_2$	$[\;\;0$	0	-1	0	$0\,]$
⑩	$\beta\, D_{\text{out}}\, s$	$[\;\;0$	0	0	0	$-1\,]$
⑪	$D_{\text{wit}}\, s_1$	$[-1$	0	0	0	$0\,]$
⑫	$D_{\text{wit}}\, b_1$	$[\;\;0$	-1	0	0	$0\,]$
⑬	$D_{\text{wit}}\, s_2$	$[\;\;0$	0	-1	0	$0\,]$
⑭	$D_{\text{wit}}\, b_2$	$[\;\;0$	0	0	-1	$0\,]$
⑮	$(1-\beta)\, D_{\text{wit}}\, s$	$[\;\;0$	0	0	0	$-1\,]$

(13)

About the second equation of (12)**:** Basically the jumps are $\frac{1}{N_j}\, \tilde{\nu}_j$, but near the border of $\mathbb{R}^5_+$ to avoid jumps that can lead to negative concentration values, we adopt truncated jumps so that $x + \nu_j(x) \in \mathbb{R}^5_+$ for all $x \in \mathbb{R}^5_+$. Indeed, note that $-\tilde{\nu}_{\min} \le \tilde{\nu}_{ij} \le \tilde{\nu}_{\max}$, so that if $x_i \ge \tilde{\nu}_{\min} / \min_j N_j$ then $\nu_{ij}(x) = \tilde{\nu}_{ij}/N_j$, where $\tilde{\nu}_{ij}$ and $\nu_{ij}(x)$ denote the ith component of $\tilde{\nu}_j$ and $\nu_j(x)$ respectively. Define:

$$\mathcal{D} \stackrel{\text{def}}{=} \{x \in \mathbb{R}^5_+ ; x_i \ge \tilde{\nu}_{\min} / \min_j N_j\} \tag{14}$$

so that $x \in \mathcal{D}$ implies that $\nu_j(x) = \frac{1}{N_j}\tilde{\nu}_j$.

The drift coefficient: Given $X(t) = x$, the expectation of $X(t + \Delta t)$ is

$$\begin{aligned}
&\mathbb{E}[X(t+\Delta t)|X_t = x] \\
&= \textstyle\sum_{j=1}^{J}\big(x+\nu_j\big)\,\mathbb{P}[\text{reaction } j|X_t = x] + x\,\mathbb{P}[\text{no reaction }|X_t = x] \\
&\simeq \textstyle\sum_{j=1}^{J}\big(x+\nu_j\big)\,(\lambda_j(x)\,\Delta t) + x\,(1 - \sum_{j=1}^{J}\lambda_j(x)\,\Delta t) \\
&\simeq x + \textstyle\sum_{j=1}^{J}\lambda_j(x)\,\nu_j(x)\,\Delta t = x + F(x)\,\Delta t
\end{aligned}$$

where

$$F(x) \stackrel{\text{def}}{=} \textstyle\sum_{j=1}^{J}\lambda_j(x)\,\nu_j(x) \tag{15}$$

So locally in time, $\mathbb{E}(X(t))$ evolves according to the drift coefficient $F(x)$ (note that $\mathbb{E}(X(t))$ is not solution of an ODE as the function F is non linear). We can easily check that:

$$F(x) = 1_{x\in\mathcal{D}}\,\tilde{F}(x) + 1_{x\notin\mathcal{D}}(x)\,F(x)$$

where

$$F(x) \stackrel{\text{def}}{=} \textstyle\sum_{j=1}^{J}\tilde{\lambda}_j(x)\,\tilde{\nu}_j \tag{16}$$

does not depend on the N_j's. Finally from $|F(x)| + |\tilde{F}(x)| \leq C\,(1 + |x|)$ and $F(x) - \tilde{F}(x) = 1_{x\notin\mathcal{D}}\big(F(x) - \tilde{F}(x)\big)$, we get:

$$|F(x) - \tilde{F}(x)| \leq C\,1_{x\notin\mathcal{D}}(1 + |x|)\,.$$

Rubrique

Identification of source for the bidomain equation using topological gradient

Jamila Lassoued[1], Moncef Mahjoub[1], and Nejib Zemzemi[2]

[1] University of Tunis El Manar
National Engineering School of Tunis $LAMSIN - ENIT$
BP 37, 1002 Tunis Belvedere, Tunisia
jamila.lassoued@enit.rnu.tn
moncef.mahjoub@lamsin.rnu.tn

[2]University of Bordeaux I, INRIA
200 Avenue de la vielle Tour 33405 Talence Cedex France.
nejib.zemzemi@inria.fr

RÉSUMÉ. Nous présentons une approche pour estimer les sources électriques dans le coeur à partir de mesures non invasives enregistrées sur la surface externe du thorax. L'approche est basée sur la méthode du gradient topologique. Cette méthode consiste à étudier le comportement d'une fonction coût via une perturbation locale du domaine. Nous montrons que l'approche proposée est capable d'identifier un terme source quand le support de la source est réduit dans l'espace.

ABSTRACT. We present an approach for estimating electrical sources within the heart domain from noninvasive measurements recorded on the outer surface of the torso. The approach is based on the topological gradient method. This method studies the behavior of a cost function during a local perturbation of the domain. We show that the proposed approach based on the topological gradient method has actually been able to identify the source terms when they are clustred in space.

MOTS-CLÉS : Le modèle bidomaine, électrophysiologie cardiaque, gradient topologique, analyse de sensibilité.

KEYWORDS : Bidomain model, cardiac electrophsiology , topological gradient, analysis sensibility.

1 Introduction

In order to localize the electrical sources in the heart, we make use of a recent method based on the topological gradient introduced by Sokolowski [7] and Masmoudi [6]. The topological gradient was originally used as part of the optimization shapes in solid mechanics [5]. This approach has subsequently been applied to a large number of areas : in imaging, it was first used for the detection of contours [4], in image classification [1], inpainting [2] and segmentation [3]. The calculation of topological sensitivity associated with the cost function of the inverse problem provides good qualitative information on the location of obstacles.

In this work, we are interested in the identification of the source term f from the boundary data obtained from the solution of the following system of equations :

$$\begin{cases} -div((\boldsymbol{\sigma_{\mathbf{i}}} + \boldsymbol{\sigma_{\mathbf{e}}})\nabla u_e) & = f & \text{in } \Omega_H \\ -div(\boldsymbol{\sigma_{\mathbf{T}}}\nabla u_T) & = 0 & \text{in } \Omega_T \\ \boldsymbol{\sigma_{\mathbf{T}}}\nabla u_T.n_T & = 0 & \text{on } \Gamma_{\text{ext}}. \\ u_e & = u_T & \text{on } \Sigma, \\ \boldsymbol{\sigma_{\mathbf{e}}}\nabla u_e.n + \boldsymbol{\sigma_{\mathbf{T}}}\nabla u_T.n_T & = 0 & \text{on } \Sigma, \end{cases} \tag{1}$$

where Ω_H (respectively Ω_T) is the heart (respectively, torso) domain (see figure 1), $\Sigma = \partial\Omega_H$ is the epicardial boundary and Γ_{ext} is the body surface. The tensors $\boldsymbol{\sigma_{\mathbf{i}}}, \boldsymbol{\sigma_{\mathbf{e}}}$ and $\boldsymbol{\sigma_{\mathbf{T}}}$ are respectively the intracellular, extracellular and thoracic conductivity tensors. The torso potential is denoted by u_T. The source term f is defined by

$$f = div(\boldsymbol{\sigma_{\mathbf{i}}}\nabla V_m)$$

where $V_m = u_i - u_e$ with u_e and u_i are respectively the extra-cellular potential and the intra-cellular potential. If we consider the dynamic of the electrical wave, the transmembrane potential V_m is governed by a reaction diffusion equation and is coupled to the extra-cellular potential, following these equations

$$\begin{cases} \chi_m \partial_t V_m + I_{ion}(V_m, w) - div(\boldsymbol{\sigma}_{\text{i}}\nabla V_m) - div(\boldsymbol{\sigma}_{\text{i}}\nabla u_e) & = I_{app} \quad \text{in } \Omega \times (0,T), \\ \partial_t w + G(V_m, w) & = 0 \quad \text{in } \Omega \times (0,T) \\ \boldsymbol{\sigma}_{\text{i}}\nabla V_m.n & = 0 \quad \text{on } \partial\Omega \times (0,T), \end{cases} \tag{2}$$

where I_{app} is a given external current stimulus. w represents the concentrations of different chemical species and variables representing the openings or closures of some gates of the ionic channels. The ionic current $I_{ion}(V_m, w)$ and the function $G(V_m, w)$ are described by the Mitchell and Schaeffer model [8]. Note that the equation (1) represents the diffusion of the electrical potential at a given time. The combination of equations (1) and (2) provides the model of the electrical wave propagation in the heart and the torso. This is known in the litterature as the the bidomain-torso coupled problem. In this study, the dynamic of the electrical wave is not considered in the identification of the source, we only consider (1). The bidomain-torso coupled problem is only used to generate synthetical observations.

By defining $\Omega = \Omega_H \cup \Omega_T$, $u = \begin{cases} u_e & \text{in } \Omega_H \\ u_T & \text{in } \Omega_T \end{cases}$ and $\boldsymbol{\sigma} = \begin{cases} (\boldsymbol{\sigma_{\mathbf{i}}} + \boldsymbol{\sigma_{\mathbf{e}}}) & \text{in } \Omega_H \\ \boldsymbol{\sigma_{\mathbf{T}}} & \text{in } \Omega_T \end{cases}$, the problem (1) could be rewritten as follows

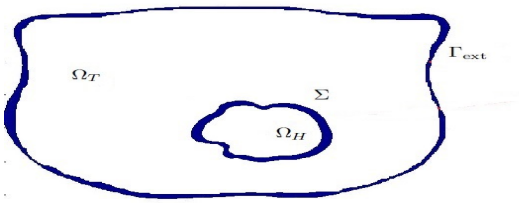


Figure 1 – The heart and torso domains

$$\begin{cases} -div(\boldsymbol{\sigma}\nabla u) = f\mathbf{1}_{\Omega_H} & \text{in } \Omega \\ \boldsymbol{\sigma}\nabla u.n_T = 0, & \text{on } \Gamma_{\text{ext}}. \end{cases} \tag{3}$$

2 Topological gradient method

We use two notations of the cost function to be minimized : $j(\Omega_H)$ and $\mathfrak{J}(u_{e,\Omega_H})$, where u_{e,Ω_H} is the solution to the system (1). The idea of topological asymptotic analysis is to measure the effect of a perturbation of the domain Ω_H on the cost function. For a small $\epsilon \geq 0$, let $\Omega_\epsilon := \Omega_H \setminus \overline{\theta_\epsilon}$ be the perturbed domain by the insertion of an inclusion $\theta_\epsilon = x_0 + \epsilon\theta$, where $x_0 \in \Omega_H$ and θ is a given, fixed and bounded domain of $\mathbb{R}^d$, containing the origine, whose boundary $\partial\omega$ is C^1. The topological sensitivity theory provides then an asymptotic expansion of the considered cost function when the size of ω_ϵ tends to zero. It takes the general form :

$$j(\Omega_\epsilon) - j(\Omega_H) = \rho(\epsilon)g(x_0) + o(\rho(\epsilon)),$$

where $\rho(\epsilon)$ is an explicit positive function going to zero with ϵ, and $g(x_0)$ is the topological gradient at point x_0. Then in order to minimize the criterion, one has to insert small inclusion at points where the topological gradient is the most negative. In our case, the source would be identified in the zones where the topological gradient is the most negative. $j(\Omega_\epsilon)$ would be a function minimizing the gap between the solution u_ϵ solution of the following problem and a given observed data.

$$\begin{cases} -div(\boldsymbol{\sigma}\nabla u_\epsilon) = f_\epsilon\mathbf{1}_{\Omega_H} & \text{in } \Omega, \\ \boldsymbol{\sigma}\nabla u_\epsilon.n_T = 0, & \text{on } \Gamma_{\text{ext}}, \end{cases} \tag{4}$$

where

$$f_\epsilon = \begin{cases} f_1 & \text{on } \theta_\epsilon \\ f_0 & \text{on } \Omega_\epsilon. \end{cases}$$

is the unknown source to be identified.

2.1 Variational formulation

The solution of the problem (4) is defined up to a constant, thus we define the suitable functional space by

$$V = \{v \in H^1(\Omega) \quad , \quad \int_{\Omega_H} v = 0\}$$

and the bilinear form A_ϵ and the linear form l_ϵ as

$$A_\epsilon(u_\epsilon, v) = \int_\Omega \boldsymbol{\sigma} \nabla u_\epsilon \nabla v \quad and \quad l_\epsilon(v) = \int_\Omega f_\epsilon v \quad \forall v \in V$$

Then the variational formulation of this problem reads such that

$$\int_\Omega \boldsymbol{\sigma} \nabla u_\epsilon \nabla v = \int_\Omega f_\epsilon v, \forall v \in V.$$

The solution u_ϵ is solution of $A_\epsilon(u_\epsilon, v) = l_\epsilon(v), \forall v \in V$. To determine the topological gradient we need to compute the adjoint solution of this problem.

2.2 Adjoint problem

We consider the direct solution u_ϵ satisfying $A_\epsilon(u_\epsilon, v) = l_\epsilon(v)$ and we define the lagrangian $L_\epsilon(u, p) = \mathfrak{J}(u) + A_\epsilon(u, p) - l_\epsilon(p)$, for every $u, p \in V$. One could check that if u_ϵ is solution of (4) we have

$$L_\epsilon(u_\epsilon, v) = \mathfrak{J}(u_\epsilon)$$

We denote $D_u L_\epsilon$ and $D_u \mathfrak{J}$ the derivative of L_ϵ and $\mathfrak{J}$ respectively, so

$$D_u L_\epsilon(u_\epsilon, v) = D_u \mathfrak{J}(u_\epsilon)$$

Then we define the abstract adjoint equation by

$$(D_u L_\epsilon, \psi) = 0, \forall \psi \in V$$

we have

$$(D_u \mathfrak{J}(u), \psi) + \int_\Omega \boldsymbol{\sigma} \nabla p \nabla \psi = 0$$

So

$$\int_\Omega \boldsymbol{\sigma} \nabla p \nabla \psi = -(D_u \mathfrak{J}(u), \psi)$$

Finally the adjoint solution p associated of the cost function $\mathfrak{J}$ is given by

$$\begin{cases} -div(\boldsymbol{\sigma} \nabla p) & = -D_u \mathfrak{J}(u) \text{ in } \Omega \\ \nabla p \cdot n_T & = 0 \qquad\qquad \text{on } \Sigma, \end{cases} \tag{5}$$

We remarque that the computation time and memory space required by the state adjoint method are largely reasonable. In the next section we will derive the variation of the cost function j with respect to the insertion of a small subdomain ω_ϵ in the cardiac domain Ω_H. We begin our analysis by giving the main hypothesis 1, then the main result of this section is presented by Theorem 1. It concerns the topological asymptotic expansion of a cost function $\mathfrak{J}$.

2.3 Main result

Let us consider the following hypothesis :

hypothesis 1 *We assume That*
(i) $\mathfrak{J}$ *is differentiable with respect to u, we denote* $D\mathfrak{J}(u)$*its derivative.*
(ii) There exists a real number $\partial J(x_0)$ *such that*

$$\mathfrak{J}(u_\epsilon) - \mathfrak{J}(u_0) = D\mathfrak{J}(u_0)(u_\epsilon - u_0) + \epsilon^d|\omega_\epsilon|\partial J(x_0) + o(\epsilon^d)$$

(iii) $\|u_\epsilon - u\|^2_{L^2(\partial\Gamma_{ext})} = o(\epsilon^d)$
(iv) $\|\nabla(u_\epsilon - u)\|^2_{L^2(\partial\Gamma_{ext})} = o(\epsilon^d)$

The expression of the topological gradient for this problem is given by the following result :

Theorem 1 *Under the hypothesis above the cost function j has the following asymptotic expansion :*

$$j(\Omega_\epsilon) - j(\Omega_H) = \epsilon^d|\omega_\epsilon|\partial J(x_0) - \epsilon^d|\omega_\epsilon|(f_1 - f_0)p(x_0)$$

In other words, the topological gradient at x_0 *is :*

$$g(x_0) = \partial J(x_0) - (f_1 - f_0)p(x_0)$$

where p is the adjoint solution.

Proof 1 *We always seek to minimize the function* $\mathfrak{J}$ *defined above. We consider the lagrangian*

$$L_\epsilon(u, v) = \mathfrak{J}(u) + A_\epsilon(u, v) - l_\epsilon(v)$$

u_ϵ *is solution of problem 4, then we have*

$$j(\Omega_\epsilon) = L_\epsilon(u_\epsilon, v)$$

So the first variation of the cost function with respect to ϵ *is given by*

$$\begin{aligned} j(\Omega_\epsilon) - j(\Omega_H) &= L_\epsilon(u_\epsilon, v) - L_0(u_0, v) \\ &= \mathfrak{J}(u_\epsilon) - \mathfrak{J}(u_0) + A_\epsilon(u_\epsilon, v) - A_0(u_0, v) - l_\epsilon(v) + l_0(v) \end{aligned}$$

Then from the definition of A_ϵ *and* l_ϵ *we have :*

$$A_\epsilon(u_\epsilon, v) - A_0(u_0, v) = \int_\Omega \boldsymbol{\sigma}\nabla(u_\epsilon - u_0)\nabla v$$

$$l_\epsilon(v) - l_0(v) = \int_{\omega_\epsilon} (f_1 - f_0)v$$

Choosing $v = p$ *the adjoint solution is solution of* (5)

$$\int_\Omega \boldsymbol{\sigma}\nabla(u_\epsilon - u_0)\nabla p = -D\mathfrak{J}(u_0)(u_\epsilon - u_0)$$

Then we have

$$j(\Omega_\epsilon) - j(\Omega) = \mathfrak{J}(u_\epsilon) - \mathfrak{J}(u_0) - DJ(u_0)(u_\epsilon - u_0) - \int_{\omega_\epsilon} (f_1 - f_0)p$$

From the hypothesis we have

$$j(\Omega_\epsilon) - j(\Omega_H) = \epsilon^d |\omega_\epsilon| \partial J(x_0) - \epsilon^d |\omega_\epsilon| (f_1 - f_0) p(x_0)$$

So we have

$$j(\Omega_\epsilon) - j(\Omega_H) = \rho(\epsilon) g(x_0) + o(\rho(\epsilon))$$

where

$$g(x_0) = \partial J(x_0) - (f_1 - f_0) p(x_0)$$

where $\partial J(x_0)$ depends on the cost function. We will present in the next section some examples of the cost function and the associated $\partial J(x_0)$ term.

3 Numerical results

In this paragraph we aim to recover the source term with the help of the non-invasive observations on the external boundary of the torso. We use the bidomain model in order to create a source term based on reaction diffusion equation. We solve the electrostatic source identification problem at a given time step. The topological gradient method is implemented using the following algorithm :

- Solve the forward solution of the problem 4.
- Compute the adjoint solution of the problem 5.
- Compute the topological gradient g.
- Search for the minimum of the topological gradient.

In order to numerically test the topological gradient method, we consider a two cost functions $\mathfrak{J}_1(u) = \int_{\partial\Gamma_{ext}} |u - u_{obs}|^2 dx$ and $\mathfrak{J}_2(u) = \int_{\partial\Gamma_{ext}} |\nabla u - \nabla u_{obs}|^2 dx$, where u_{obs} is the observed data at the body surface Γ_{ext}. We tested this method for both cost functions in two different scenarios. The first case is for clustered source. The electrical source in this case is obtained by solving the bidomain equation with a single site stimuli until 4ms. The second case is for a distributed source, The electrical source in this case is the gradient of the transmembrane potential at 20 ms after a single site stimuli.

clustered source :

In figure 2 (a), we show the distribution of the extracellular potential in the heart domain after 4ms of a single site stimulation. The topological gradient distribution is shown in figure 2 (b) for the cost function $\mathfrak{J}_1$ and figure2 (b) for the cost function $\mathfrak{J}_2$. The green circle in figures 2 (b,c,e,f) denotes the position of the source at 4 ms and the red point is the source obtained using the topological gradient method. The source at time 4ms could be deduced from figure 2 (e), where we represent the distribution of $f = div(\boldsymbol{\sigma_i} \nabla V_m)$. We distinguish two clustered sources. We remark that the electrical source is globally well localized. The two cost functions seems to capture one of the two sources at time 4ms.

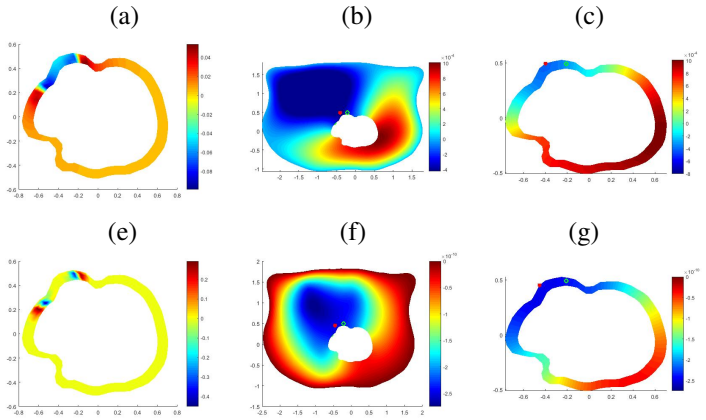


Figure 2 – (a) the solution u_e at 4 ms, (e) : the source. (b) (respectively,(c)) : the topological gradient for the cost function $\mathfrak{J}_1$ (respectively,$\mathfrak{J}_2$) in the heart thorax doamin. (f) (respectively,(g)) : The topological gradient for the cost function $\mathfrak{J}_1$ (respectively,$\mathfrak{J}_2$) in the heart doamin.

Distributed source

Here we test the capability of the method in localizing distributed sources. We run a simulation of a single site stimuli and we extract the data after 20 ms. In figure 3 (a), we show the distribution of the extracellular potential in the heart domain. The topological gradient distribution is shown in figure 3 (b) for the cost function $\mathfrak{J}_1$ and figure3 (b) for the cost function $\mathfrak{J}_2$. The green circle in figures 3 (b,c,e,f) denotes the position of the source at 20 ms and the red point is the source obtained using the topological gradient method. The source at time 20ms could be deduced from figure 3 (e). We distinguish two sources far from each other. We remark that the first cost function still provides an averaged position which is here very far from both real sources figure 3 (e). By the contrary, the second cost function still captures with a good accuracy one of the two sources at time 20 ms.

4 Conclusion

We presented a new approach for localizing electrical sources in the heart. This approach is based on the topological gradient method. We have tested this method on in silico data obtained by solving the bidomain problem. The numerical results show that the method is accurate when dealing with clustered sources. Our investigation shows that the considering the cost function $\mathfrak{J}_2(u) = \int_{\partial\Gamma_{ext}} |\nabla u - \nabla u_{obs}|^2$ is better than considering $\mathfrak{J}_1(u) = \int_{\partial\Gamma_{ext}} |u - u_{obs}|^2$. The first capture one of the two sources. The latter tries to find an averaged position. This works well when the source is clustred but when the sources are far from each other, the function $\mathfrak{J}_2(u)$ seems to localise the source that is the closest to the body surface. These preliminary results have been conducted in 2D simulations and have to be confirmed with much more testing with multiple stimuli and multiple sources for the 2D and the 3D cases. This would be the topic of our future investigations.

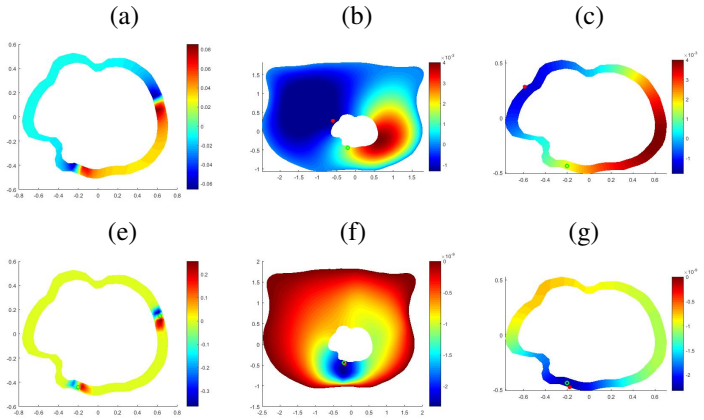


Figure 3 – (a) the solution u_e at 20 ms, (e) : the source. (b) (respectively,(c)) : the topological gradient for the cost function $\mathfrak{J}_1$ (respectively,$\mathfrak{J}_2$) in the heart thorax doamin. (f) (respectively,(g)) : The topological gradient for the cost function $\mathfrak{J}_1$ (respectively,$\mathfrak{J}_2$) in the heart doamin.

Références

[1] AUROUX, DIDIER AND BELAID, L JAAFAR AND MASMOUDI, MOHAMED « Image restoration and classification by topological asymptotic expansion » *Variational formulations in mechanics : theory and applications*, p. 23–42, 2006.

[2] AUROUX, DIDIER AND MASMOUDI, MOHAMED « A one-shot inpainting algorithm based on the topological asymptotic analysis » *Computational & Applied Mathematics*, vol. 25, n° 23, p. 251–267, 2006.

[3] AUROUX, DIDIER « From restoration by topological gradient to medical image segmentation via an asymptotic expansion » *Mathematical and Computer Modelling*, vol. 49, n° 11, p. 2191–2205, 2009.

[4] BELAID, L JAAFAR AND JAOUA, M AND MASMOUDI, M AND SIALA, L « Application of the topological gradient to image restoration and edge detection » *Engineering Analysis with Boundary Elements*, vol. 32, n° 11, p. 891–899, 2008.

[5] ESCHENAUER, HANS A AND KOBELEV, VLADIMIR V AND SCHUMACHER, A « Bubble method for topology and shape optimization of structures » *Structural optimization*, vol. 8, n° 01, p. 42–51, 1994.

[6] MASMOUDI, MOHAMED « The topological asymptotic » *PICOF'02 : problèmes inverses, contrôle et optimisation de formes. Colloque*, p. 285–289, 2002.

[7] SOKOLOWSKI, J AND ZOCHOWSKI, A « On the Topological Derivative in Shape Optimization » *SIAM Journal on Control and Optimization*, vol. 37, n° 04, p. 1251–1272, 1999.

[8] COLLEEN CMITCHELL ANDDAVID G SCHAEFFER., « A two-current model for the dynamics of cardiacmembrane », *Bulletin of mathematical biology*, vol. 5, n° 65, p. 767–793, 2003.