Proceedings of CARI 2018 (African Conference on Research in Computer Science and Applied Mathematics)

Eric Badouel, Nabil Gmati, Bruce Watson

Proceedings of CARI 2018

Actes du CARI 2018

Stellenbosch, South Africa, October 2018

EDITORS : NABIL GMATI - ERIC BADOUEL - BRUCE WATSON

EDITORS : NABIL GMATI - ERIC BADOUEL - BRUCE WATSON

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 quatorzième édition cette année en Afrique du sud. 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 , à Saint-Louis du Sénégal en 2014 et à Tunis in 2016.

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), et l'Agence Universitaire de la Francophonie (AUF). Cette treizième édition, confiée à l'Université de Stellenbosch, sous la coordination du professeur Bruce Watson , a bénéficié également du soutien de l'IFIP.

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 28 communications scientifiques, sélectionnées parmi 100 articles soumis, et des conférences invitées présentées par des spécialistes de renommée internationale. La conférence a été précédée d'une école de recherche sur les méthodes formelles, co-organisée par ICTAC, une conférence internationale sur les aspects théoriques de l'informatique qui s'est tenue conjointement avec le CARI à Stellenbosch.

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 Bruce Watson.

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

Nabil Gmati, Président du CARI
Eric Badouel, Secrétaire du Comité permanent du CARI
Bruce Watson, Organisateur du CARI 2018

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, Saint-Louis du Senegal in 2014, and Tunis in 2016.

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), and the Agence Universitaire de la Francophonie (AUF). This fourteenth edition, entrusted to the University of Stellenbosch, under the coordination of Professor bruce Watson, has also benefited from the support of IFIT.

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 28 scientific contributions, selected from 100 submissions, together with invited talks delivered by acknowledged specialists. It was preceded by a research school on the formal aspects of computing, co-organized by ICTAC, an international conference on the theoretical aspects of computing that is jointly organized with CARI in Stellenbosch.

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 Bruce Watson.

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

Nabil Gmati, Chairman of CARI
Eric Badouel, Secretary of CARI Permanent Committee
Bruce Watson, Chair of CARI 2018

LISTE DES RELECTEURS – *LIST OF REFEREES*

Nahla ABDELLATIF
Yamine AIT AMEUR
Soraya AIT CHELLOUCHE
Hugo ALATRISTA-SALAS
Mejdi AZAIEZ
Jérôme AZE
Eric BADOUEL
Monique BARON
Nicolas BECHET
Hacene BELHADEF
Slimane BEN MILED
Ahmed BOUAJJANI
Asma BOUHAFS
Tarik BOUJIHA
Ibrahim BOUNHAS
Sandra BRINGAY
Patrice BUCHE
Gaoussou CAMARA
Hadda CHERROUN
Yacine CHITOUR
Nadia CHOUAIEB
Loek CLEOPHAS
Laurent DEBREU
Belhassen DEHMAN
Rachid ELLAIA
Abdel ELLATIF SAMHAT
Melhem EL HELOU
Daoudi EL MOSTAFA
Nadjia EL SAADI
Mohamed Faouzi ATIG
Radhouene FEKIH SALEM
Bernd FISCHER
Davide FREY
Abdoulaye GAMATIE
Jean-Frédéric GERBEAU
Nabil GMATI
Stefan GRUNER
Bamba GUEYE
Abdou GUERMOUCHE
Yassine HADJADJ-AOUL
Abderrahmane HABBAL
Jean-Claude HOCHON
Michel HURFIN
Marc IBRAHIM
Roberto INTERDONATO
Frédéric JEAN
Sofiene JELASSI
Dibie JULIETTE
Amira KEBIR
Bouabdellah KECHAR
Eric KERGOSIEN
Hélène KIRCHNER
Kolyang KOLYANG
Derrick KOURIE
Maryline LAURENT
Mohamed Tayeb LASKRI
Julia LAWALL
Philippe LEMOISSON
Christophe LETT
Cédric LOPEZ
Juan Antonio LOSSIO-VENTURA
Stéphane MAAG
Azmi MAKHLOUF
Pierre MANNEBACK
Thomas MAUGEY
Pierre MAUREL
Ludovic ME
Ali MILI
Nathalie MITTON
Valery MONTHE
Ahmed MOUSSA
Ali MOUSSAOUI
Gérard NDENOKA
Thomas Djotio NDIE
Tri NGUYEN-HUU
Boniface NKONGA
Rachid OUIFKI
Patrice QUINTON
Josvah Paul RAZAFIMANDIMBY
Mathieu ROCHE
Ounsa ROUDIES
Houari SAHRAOUI
Abed Ellatif SAMHAT
HASSAN SANEIFAR
Tewfik SARI
Fathia SAIS
Ina SCHAEFER
Mokhtar SELLAMI
Rachid SENOUSSI
Joël SOR
Maurice TCHUENTE
Maguelonne TEISSEIRE
César VIHO
Bruce WATSON
Jean-Daniel ZUCKER

TABLE DES MATIERES / TABLE OF CONTENTS

A multi-seasonal model of the dynamics of banana plant-parasitic nematodes

Israël Tankam[a, b, d, *] - Suzanne Touzeau[e, b] - Frédéric Grognard[b] - Ludovic Mailleret[e] - JJ. Tewa[a, 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 Université Côte d'Azur, Inria, INRA, CNRS, UPMC Univ Paris 06, BIOCORE, France
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, University of Yaoundé I, Faculty of Science P.O. Box 812, Yaoundé, Cameroon
e Université Côte d'Azur, INRA, CNRS, ISA, France
This work is supported by EPITAG, an Inria Associate team part of the LIRIMA (https://team.inria.fr/epitag/).

ABSTRACT. In this paper, a hybrid multi-seasonal model is proposed to describe the action of *Radopholus similis* with banana plants' roots. On one side a general Holling type II predator-prey model with stage structuration of the predators is coupled including a host-parasite dynamic with a parasite free living stage. On the other side, at a certain period called inter-seasonal time, a decay equation is given, consisting essentially in the exponential decay of free living pest when hosts are lacking. The switching between these two continuous systems is given by discrete laws and the switchings are repeated season after season. The proposed model is reduced and analysed and relevant constants like the basic reproduction number and the minimal inter-season duration for pest eradication are computed. Numerical simulation are provided.

RÉSUMÉ. Dans cet article, un modèle hybride multi-saisonnier est proposé pour décrire l'action de *Radopholus similis* sur les racines des bananiers. D'un côté, un modèle proie-prédateur de Holling type II avec une structuration par stade des prédateurs est couplé avec une dynamique hôte-parasite incluant un stade libre des parasites. De l'autre côté, sur un temps dit inter-saisonnier, une équation de désintégration est donnée, consistant essentiellement en la décroissance exponentielle de la population de parasite libre en absence d'hôte. La commutation entre ces deux systèmes continus est donnée par des lois discrètes et les commutations sont répétées saison après saison. Le modèle proposé est réduit puis analysé et des constantes révélatrices comme le taux de reproduction de base et la durée minimale d'inter-saison pour l'éradication des ravageurs sont calculées. Des simulations numériques sont fournies.

KEYWORDS : Mathematical modelling; Radopholus similis; Multi-seasonal model; Semi-discrete model; Parasit-host dynamics; Slow-fast dynamics; Model reduction;

MOTS-CLÉS : Modélisation mathématique; Radopholus similis; Modèle multi-saisonnier; Modèle semi-discret; Dynamique hôte-parasite; Dynamique lent-rapide; Réduction de modèle;

1. Introduction

Banana cultures are hampered by several parasitic factors like plant-parasitic nematodes, insect pests or soil-borne fungi that seriously threaten the sustainability of these systems by decreasing yield, causing plant toppling or requiring intensive pesticide use. The nematode *Radopholus similis* is the most significant parasitic nematode of the banana plant and the banana plantain plant in the world [8]. The infestation by the *Radopholus similis* causes damages going from simple roots lesion reducing the production to the fall of the seedlings. These damages are due to the fact that the nematodes destroy the roots tissue by feeding on. Hence, *Radopholus similis* is one of the most regulated pests of banana plant [5] but its control still implies toxic nematicides which are not always efficient.

After describing briefly the biology of *R. Similis*, we propose a general model that will be reduced and summarily analysed then provide some numerical result.

2. Biological background and model formulation

The burrowing nematode *Radopholus similis* is a phytophagous nematode that attacks the roots of host plants. Like most nematodes in its family, the Pratylenchidae family, it is an obligate parasite. The following observations show that, it can only feed on living roots, which explains why: (i) it occurs mainly in roots and rhizomes, and little in the soil; the ratio of population density in soil and roots (expressed as the number of nematodes per gram) is generally less than 1/100; (ii) in roots, maximum densities are observed at the edges of necrotic areas or between necrotic zones, and not in necrotic root sections [1]. In absence of host, the population of nematodes therefore decreases. Exponential model fairy well describes the decay of the nematode population in the absence of a host [3].

Concerning banana plant biology and cultivation, its roots are continuously produced until the flowering [2]; subsequently, the newly emitted roots are mainly related to successive suckers. After a complete season of culture of banana -10 to 12 months- it is advisable to remove all the old plant roots before planting healthy suckers, in order to avoid the nematodes to directly infest suckers roots from the infested roots. The crop rotation or the fallow strategies consisting in leaving the soil free of any nematodes host for a while in order to insure sufficient decay of nematodes population in the soil and reduce the infestation.

According to the previous biological background, the following assumptions are made:

– There are two compartments for nematodes: free nematodes in the soil (P) and nematodes infesting the roots (X).

– There is one compartment for healthy roots (S).

– There are several cropping seasons with an inter-season that match the duration of the fallow or the alternative culture duration.

– During one cropping season, banana roots grow logistically [4] until the flowering. The duration d until the flowering is usually 7 months and a cropping season usually lasts t_f = 11 months. If τ is the duration of the inter-season, we let $T = t_f + \tau$ be the combined duration of the banana cropping and the alternative cropping (or the fallow).

– The logistic growth of the roots during a cropping season is then given by: $\frac{dS}{dt} = g(t,S)$ where the function $g(t,S)$ is defined by

$$g(t,S) = \rho(t)S\Big(1 - \frac{S}{K}\Big).$$

The form of the function $\rho(t)$ within the $(n+1)$th season follows:

$$\rho(t) = \begin{cases} \rho & , \quad t \in \{0\} \cup]nT, nT+d], \\ 0 & , \quad t \in]nT+d, nT+tf]. \end{cases} \tag{1}$$

Where ρ is the growth rate of the roots during the growth phase.

– At the end of a season, the roots of the plants are torn off. We assume that a small fraction q (that can be null) of infesting nematodes remains in the soil. This fraction corresponds to the nematodes that leave the roots toward the soil at the removal or the non-fresh roots that are left in the soil during the removal.

– When free pests contact plant roots (S) with a rate β, they infest the roots and start feeding on it with a saturated response as well as in Holling type II functional response, which is well-suited for invertebrates [6]. In the absence of experimental data, it seems coherent to rely on this functional response, since nematodes are invertebrates.

– The infesting parasites have a natural mortality μ.

– Infesting nematodes use a part of their food to reproduce inside (proportion γ) or outside (proportion $1-\gamma$) the roots.

– In the absence of hosts, free nematodes undergo an exponential decay with a rate Ω.

This assumptions result in a three-dimensional semi-discrete model; a switched system, coupled to two sets of recurrence equations and one set of ordinary differential equations, define the model.

During the cropping, i.e $t \in \{0\} \cup]nT, nT+tf]$, free nematodes ($P$) and infesting nematodes (X) interact with the healthy roots according to the following switching system:

$$\begin{cases} \dfrac{dP(t)}{dt} &= -\beta P(t)S(t) + \alpha a(1-\gamma)\dfrac{S(t)X(t)}{S(t)+\Delta} - \Omega P(t), \\ \dfrac{dS(t)}{dt} &= \rho(t)S(t)\Big(1 - \dfrac{S(t)}{K}\Big) - a\dfrac{S(t)X(t)}{S(t)+\Delta}, \\ \dfrac{dX(t)}{dt} &= \beta P(t)S(t) + \alpha a\gamma\dfrac{S(t)X(t)}{S(t)+\Delta} - \mu X(t). \end{cases} \tag{2}$$

With the initial conditions $P(0) = P_0$, $S(0) = S_0$, $X(0) = 0$.

Where $\Delta > 0$ is the half-saturation constant.

In the following, we will term the dynamics of (2) during the $\{0\} \cup]nT, nT+d]$ interval "the first subsystem of (2)" while "the second subsystem of (2)" will concern $]nT+d, nT+t_f]$, with $\rho = 0$.

When roots are removed, only free parasites and a fraction q of infesting nematodes survive. So, for $t = nT + t_f$:

$$\begin{cases} P(nT+t_f^+) &= P(nT+t_f) + q.X(nT+t_f), \\ S(nT+t_f^+) &= 0, \\ X(nT+t_f^+) &= 0. \end{cases} \tag{3}$$

where the + superscript will always indicate the instant that directly follows.

When there is no host plant, i.e $t \in]nT+t_f, (n+1)T]$, the remaining free nematodes undergo a decay:

$$\begin{cases} \dfrac{dP(t)}{dt} &= -\Omega P(t), \\ \dfrac{dS(t)}{dt} &= 0, \\ \dfrac{dX(t)}{dt} &= 0. \end{cases} \qquad (4)$$

At the beginning of a new season of banana plants, i.e. $t = (n+1)T$, fresh healthy roots are added through new suckers. The equation translating the process follows:

$$\begin{cases} P((n+1)T^+) &= P((n+1)T), \\ S((n+1)T^+) &= S_0, \\ X((n+1)T^+) &= 0. \end{cases} \qquad (5)$$

The system formed by the equations (2 - 5) represents our multi-seasonal nematodes-banana interaction model. This type of multi-seasonal model exists in the literature for another nematode, *Meloidogyne incognita* [7].

3. Model reduction and analysis

Proposition 1 *– The problem (2- 5) admits a solution that is unique for any initial condition and continuous on any interval* $]nT, nT+t_f]$ *and* $]nT+t_f, (n+1)T]$*, with* $n \in \mathbb{N}$.

– The state variables remain non-negative over the time.

– The tranasition law of free nematodes from one banana cropping season to the next is given by

$$P((n+1)T) = \left[P(nT+t_f) + qX(nT+t_f)\right]e^{-\Omega\tau}. \qquad (6)$$

Proof.

– The equation (4) with initial condition (3) easily satisfies the Cauchy-Lipschitz conditions. Hence the conclusion directly follows on the intervals $]nT+t_f, (n+1)T]$. Furthermore, each subsystem of equation (2) is a well-posed Cauchy problem. The first subsystem has $P_0, S_0, 0$ as initial condition when $n = 0$ and the initial conditions are given by (5) when $n \geq 1$. The second subsystem has the value of the solution of the first subsystem as initial condition. Therefore, since S is bounded and the (P,X) dynamics are linearly bounded, the switched system (2) admits a unique continuous solution on $\{0\} \cup]nT, nT+t_f]$.

– We first consider $n = 0$ and denote by $W = (P, S, X)$ the state vector and $W(0)$ the initial condition. As these state variables represent biological quantities, we set $W(0) \geq 0$. The structure of the model then ensures that the state variables remain non-negative in the course of time. Besides, the discrete rule (5) ensures that if the non-negative orthant is positively invariant for season n then the initial condition for season $n+1$ will be positive, hence the same conclusion will follow for $n \geq 1$.

– Let $t \in]nT+t_f, (n+1)T]$, $n \in \mathbb{N}$.

We solve $\frac{dP}{dt} = -\Omega P$ with initial condition $P(nT+t_f^+) = P(nT+t_f) + q.X(nT+t_f)$ to obtain $P((n+1)T) = \left[P(nT+t_f) + qX(nT+t_f)\right]e^{-\Omega\tau}$. ∎

The proposition 1 shows that the problem (2-5) is well posed. In the next proposition, we reduce the first subsystem of equation (2) to a Rosenzweig-MacArthur model, by introducing a new state variable $N = P + S$ that represents the total number of nematodes and using the singular perturbation theory.

Proofs of propositions (2 - 4) are left in appendix.

Proposition 2 *The first subsystem of equation (2) can be reduced to the system:*

$$\begin{cases} \dot{S} &= \rho S\Big(1-\dfrac{S}{K}\Big) - a\dfrac{SN}{S+\Delta}, \\ \dot{N} &= \alpha a\dfrac{SN}{S+\Delta} - \mu N, \end{cases} \quad (7)$$

With initial conditions $S(nT^+) = S(0) = S_0,\ N(nT^+) = N(0) = P(nT)$.

Where $N = P + X$*; assuming that the primary infestation is very fast (*β *is high) and the free pest* P *tends very fast to 0 and using the singular perturbation theory for slow-fast dynamics [11]. The states of the second subsystem will then be initialized with the final values of the first subsystems taking* $P(nT + d^+) = 0$ *and* $X(nT + d^+) = N(nT + d)$*;*

REMARK. — According to Proposition 2, the number of free pest is null in the reduced first subsystem. That is a good approximation when β has a high value.
Thus, we consider that there is no free pest at the input of the second subsystem of equation (2) and that, at the same input, the number of infesting parasites X is therefore equal to the sum N of the two. In the following proposition, we therefore compute the values of the pest level and the roots biomass in the neighbourhood of the pest free solution.

Proposition 3 *In the neighbourhood of the Pest Free Solution (PFS),*

– For all $n \in \mathbb{N}$ *and* $t \in [0, d]$*, the solution of equation (7) is given on* $]nT, nT + d]$ *by*

$$N(nT+t) = P(nT)e^{-\mu t + \displaystyle\int_0^t \frac{\alpha a S^*(\tau)}{S^*(\tau)+\Delta}d\tau} \quad (8)$$

$$S(nT+t) = S_0 - \int_0^t F(\xi)exp\Big(-\int_0^\xi \rho\Big(1-\frac{2S^*(\tau)}{K}\Big)d\tau\Big)d\xi \times exp\Big(\int_0^t \rho\Big(1-\frac{2S^*(\tau)}{K}\Big)d\tau\Big), \quad (9)$$

– For all $n \in \mathbb{N}$ *and* $t \in]nT + d, nT + tf]$.

There exists a matrix $\Pi(t)$ *detailed in appendix such as*

$$S(t) = S^*(t) - a\frac{S(nT+d)}{S(nT+d)+\Delta}X(t)$$

and

$$\begin{pmatrix} P(t) \\ X(t) \end{pmatrix} = \Pi\big(t-(nT+d)\big).\begin{pmatrix} 0 \\ N(nT+d) \end{pmatrix}$$

Where $S^*(t) = \dfrac{S_0 K}{S_0 + (K - S_0)e^{-\rho t}}$

From this result, one can now compute the basic reproduction number of the pest and the minimal inter-season duration that leads to the disappearance of the pest. That is the aim of the following proposition.

Proposition 4 *(Pest eradication)*

We have the following results:

1) For all $n \in \mathbb{N}$*, the persistence of free nematodes is given by*

$$P(nT) = P_0 e^{-n\Omega\tau}\delta^n\big[\Pi_{1,2}(t_f - d) + q\Pi_{2,2}(t_f - d)\big]^n, \quad (10)$$

Where

$$\delta = exp\Big(-\mu d + \int_0^d \frac{\alpha a S^*(\tau)}{S^*(\tau)+\Delta}d\tau\Big) = N(d)/P_0.$$

2) The basic reproduction number is given by

$$\mathcal{R}_0 = e^{-\Omega\tau}\delta\big[\Pi_{1,2}(t_f - d) + q\Pi_{2,2}(t_f - d)\big]. \tag{11}$$

3) The minimal inter-season duration τ_0 such as the pest will disappear over time as soon as $\tau > \tau_0$ is given by

$$\tau_0 = \frac{ln\Big(\big[\Pi_{1,2}(t_f - d) + q\Pi_{2,2}(t_f - d)\big]\delta\Big)}{\Omega}. \tag{12}$$

4. Numerical simulations

We consider the parameters values in Table 1. In Figure 1, we compare the difference in behaviour between the system and its reduced form for different values of β within a season. In Figure 2, we illustrate the behaviour of the reduced system in the neighbourhood of the pest free equilibrium. Since N = P+X and P = 0 in the first substem, such as N = X in this same subsystem, the value X (number of infesting nematodes) will always be the one represented instead of N. With the considered parameters, we obtain $\tau_0 = 8.04$. The figure 2 illustrates the behaviour of the pest with τ taking different values: $\tau \ll \tau_0$, $\tau < \tau_0$, $\tau > \tau_0$, $\tau \gg \tau_0$.

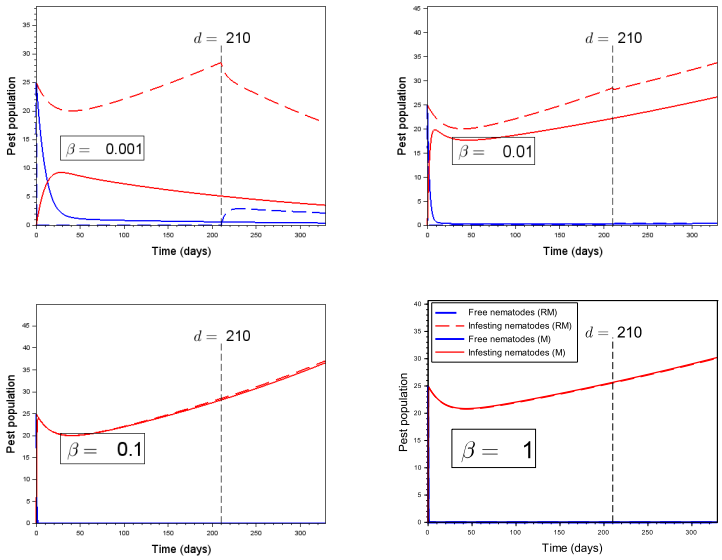


Figure 1. *Comparison between the model (M) and its reduced form (RM) for different values of β.*

Parameter	Value	Ref.
Ω	$0.0495^{(1)}$ day^{-1}	[3]
μ	$0.05 - 0.04^{(2)}$ day^{-1}	[9]
K	$\geq 143^{(3)}$ grammes (g) in Costa Rica	[10]
S_0	$60^{(4)}$ g	[10]
t_f	$300 - 360^{(5)}$ *days*	[12]
d	Berangan: 210-240 *days* Cavendish: 180-210$^{(6)}$*days*	[12]

Parameter	Value
β	10^{-1} g^{-1}.day^{-1}
γ	0.5
a	10^{-4} g.day^{-1}
α	650 g^{-1}
q	5%
P_0	25
Δ	60 g
ρ	0.06 days^{-1}

Table 1. *Values of the parameters.*$^{(1)}$ *We consider the soil as an andosol with a null matric potential.* $^{(2)}$ *We consider the average.* $^{(3)}$ *We consider the value* 150. $^{(4)}$ *We consider an approximation of the sucker survey critical level* $^{(5)}$ *We take the average value.* 330 $^{(6)}$ *We take the value* 210.

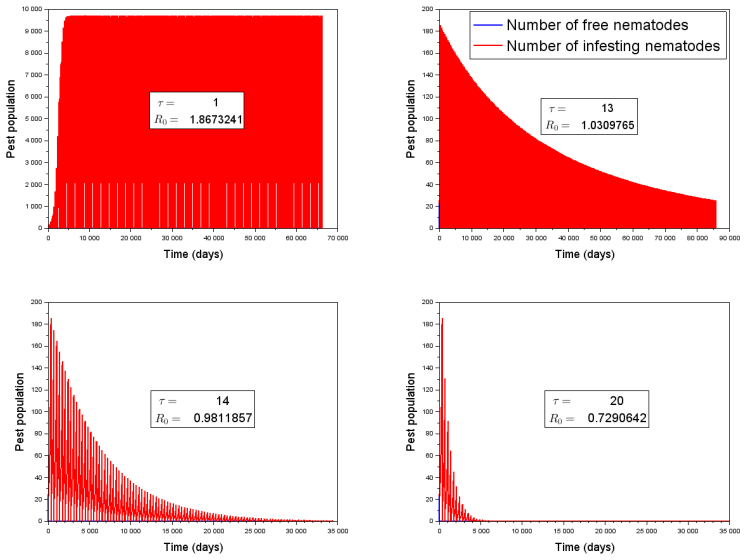


Figure 2. *Pest dynamics for different values of* τ *when* $\tau_0 = 11.78$.

REMARK. — In figure 2, free pest levels are very low and really appear when there is no host (during the inter-season) through the fraction q of infesting remaining in the soil upon extraction of the roots.

Conclusion

Nematode-host models have not undergone enough development in theory and pratical applications in the field of biomathematics in the case of *Radopholus similis*. So, in

this paper, we have studied a simple model for this kind of interaction with a saturated response and taking in account both free and infesting stages of nematodes. We have obtained a threshold on the duration of the inter-season such as the pest level tends to zero over time when that threshold is crossed. We also computed the basic reproduction number of the nematodes. All this work was done after we reduced the model by using the singular perturbation theory for slow-fast dynamics. Our numerical simulation results confirm that when the duration of the inter-season τ passes through the critical value τ_0, the pest tends to disappear. The ability to compute the basic reproduction number and the critical duration of the inter-season developed in this paper might help lead to more sophisticated strategies of control of *Radopholus similis* in agricultural fields.

5. References

[1] ARAYA, M., CENTENO, M., " Recuperacion de Radopholus similis, Helicotylenchus spp., Meloidogyne spp. y Pratylenchus spp. de raiz funcional, no funcional y combinada de banano (Musa AAA)", *Corbana*, vol. 20, pp.11-16, 1995.

[2] BEUGNON, M., CHAMPION, J., " Études sur les racines du bananier", *Fruits*, vol. 21, pp.309-327, 1966.

[3] CHABRIER, C., " Survie et dissémination du nématode Radopholus similis (Cobb) Thorne dans les sols bruns-rouilles à halloysites (nitisols) : effets de l'état hydrique et des flux hydriques.", *Phytopathology and phytopharmacy. Université des Antilles-Guyane*, pp.68-74, 2008.

[4] GREGORY, P., " Plant Roots: Growth Activity and Interactions with Soils", *Bio-Green Elsevier (Exc)*, 2006.

[5] HOCKLAND, S., INSERRA, R.N., MILLAR, L., LEHMAN, P.S., "International plant health-Putting legislation into practice", *Plant Nematology (Perry RN, Moens M. eds.) CAB international, Wallingford, UK*, pp.327-345, 2006.

[6] HOLLING, C. S., " The Functional Response of Invertebrate Predators to Prey Density ", *Memoirs of the Entomological Society of Canada*, pp.5-86, 1966

[7] NILUSMAS, S., MERCAT, M., PERROT, T., TOUZEAU, S., CALCAGNO, V., DJIAN-CAPORALINO, C., CASTAGNONE, P., MAILLERET, L., "A multi-seasonal model of plant-nematode interactions and its use for durable plant resistance deployment strategies", *Acta Horticulturae*, vol. 1182, pp. 211-218, 2017. DOI: 10.17660/ActaHortic.2017.1182.25

[8] SARAH, J.- L., "Variabilité du pouvoir pathogène de Radopholus similis entre populations provenant de différentes zones de production du monde.", *Info Musa*, vol. 2(2), pp. 6, 1993.

[9] SARAH, J.-L., PINOCHET, J., STANTON, J., " Radopholus similis Cobb, nématode parasite des bananiers.", *Fiche technique $n^{\circ}1$, INIBAP, Montpellier, France*, 2 pp, 1996.

[10] SERRANO, E., "Relationship between functional root content and banana yield in costa rica. In Banana Root System: towards a better understanding for its productive management: Proceedings of an international symposium/Sistema Radical del Banano: hacia un mejor conocimiento para su manejo productivo: Memorias de un simposio internacional", pp. 25-34 2003.

[11] VERHULST, F., "Singular perturbation methods for slow-fast dynamics", *Nonlinear Dyn*, vol. 50, pp. 747-753, 2007.

[12] BANANA CULTIVATION GUIDE, "http://mahaprisons.gov.in/Uploads/Dockets_Files/635259935664912504Banana_Cultivation_Guide_%C2%AB_Banana_Planters.pdf", Visited on March 23, 2018.

Appendix 1: Proof of Proposition 2

Let's first consider the first subsystem of equation (2), i.e. $t \in]nT, nT + d[$. Let $N = P + X$ and consider the system in (P, S, N).

Assuming that β is large, let $\beta = \frac{\beta'}{\varepsilon}$, $0 < \varepsilon \ll 1$ and $\tau = \frac{t}{\varepsilon}$. The new time τ is called *fast time*. The system with derivatives according to τ is written:

$$\left\{\begin{array}{rcl} \dfrac{dP}{d\tau} & = & -\beta' PS + \varepsilon\alpha a(1-\gamma)\dfrac{S(N-P)}{S+\Delta} - \varepsilon\Omega P, \\ \dfrac{dS}{d\tau} & = & \varepsilon\rho S\Big(1-\dfrac{S}{K}\Big) - \varepsilon a\dfrac{S(N-P)}{S+\Delta}, \\ \dfrac{dN}{d\tau} & = & \varepsilon\alpha a\gamma\dfrac{S(N-P)}{S+\Delta} + \varepsilon(\mu-\Omega)P - \varepsilon\mu N, \end{array}\right. \tag{13}$$

When $\varepsilon = 0$, we then define the *fast equation* by

$$\frac{dP}{d\tau} = -\beta' PS$$

Which admits an equilibrium $\bar{P} = 0$ that is asymptotically stable because $S > 0$ (we will have proven that the trajectories are positive).

The *slow equation* is written as

$$\left\{\begin{array}{rcl} \dot{S} & = & \rho S\Big(1-\dfrac{S}{K}\Big) - a\dfrac{SN}{S+\Delta}, \\ \dot{N} & = & \alpha a\gamma\dfrac{SN}{S+\Delta} - \mu N, \end{array}\right. \tag{14}$$

Which corresponds to a Rosenzweig-MacArthur model. The Tychonov theorem ensures that

$$\lim_{\varepsilon\to 0} P(t,\varepsilon) = 0, \quad t \in]nT, nT + d[$$

$$\lim_{\varepsilon\to 0}(S(t,\varepsilon), N(t,\varepsilon)) = (\bar{S}(t), \bar{N}(t)), \quad t \in]nT, nT + d[$$

Where $(\bar{S}, \bar{N})$ is the solution of equation (14) and $(P(t,\varepsilon), S(t,\varepsilon), N(t,\varepsilon))$ is the solution of the perturbed system:

$$\left\{\begin{array}{rcl} \varepsilon\dot{P} & = & -\beta' PS + \varepsilon\alpha a(1-\gamma)\dfrac{S(N-P)}{S+\Delta} - \varepsilon\Omega P, \\ \dot{S} & = & \rho S\Big(1-\dfrac{S}{K}\Big) - a\dfrac{S(N-P)}{S+\Delta}, \\ \dot{N} & = & \alpha a\dfrac{S(N-P)}{S+\Delta} + (\mu-\Omega)P - \mu N. \end{array}\right.$$

Appendix 2: Proof of Proposition 3

– The Pest Free Solution is written for all $t \in [0, d]$, $\begin{pmatrix} S^*(t) \\ N^*(t) \end{pmatrix} = \begin{pmatrix} \dfrac{S_0 K}{S_0 + (K - S_0)e^{-\rho t}} \\ 0 \end{pmatrix}$.

Considering the deviation variables $\tilde{S} = S(t) - S^*(t)$ and $\tilde{N} = N(t) - N^*(t) = N(t)$, one can write the deviation system as

$$\begin{cases} \dot{\tilde{S}} &= \rho(\tilde{S} + S^*(t))\Big(1 - \dfrac{\tilde{S} + S^*(t)}{K}\Big) - \dfrac{a(\tilde{S} + S^*(t))\tilde{N}}{\tilde{S} + S^* + \Delta} - \rho S^*(t)\Big(1 - \dfrac{S^*(t)}{K}\Big), \\ \dot{\tilde{N}} &= \dfrac{\alpha a(\tilde{S} + S^*(t))\tilde{N}}{\tilde{S} + S^* + \Delta} - \mu\tilde{N}, \\ & \tilde{S}(0) = 0, \tilde{N}(0) = N(0) = P_0. \end{cases} \tag{15}$$

In a neighbourhood of the PFS, the system is equivalent to

$$\begin{pmatrix} \dot{\tilde{S}} \\ \dot{\tilde{N}} \end{pmatrix} = \begin{pmatrix} \rho\Big(1 - \dfrac{2S^*(t)}{K}\Big) & -\dfrac{aS^*(t)}{S^*(t) + \Delta} \\ 0 & -\mu + \dfrac{\alpha a S^*(t)}{S^*(t) + \Delta} \end{pmatrix} \cdot \begin{pmatrix} \tilde{S} \\ \tilde{N} \end{pmatrix} \tag{16}$$

That leads to the equation in $\tilde{N}$

$$\dot{\tilde{N}} = \Big(-\mu + \frac{\alpha a S^*(t)}{S^*(t) + \Delta}\Big)\tilde{N},$$

whose solution is given by

$$\tilde{N}(t) = P_0 e^{-\mu t + \int_0^t \frac{\alpha a S^*(\tau)}{S^*(\tau) + \Delta} d\tau}.$$

One can now replace this expression in (16) and let $F(t) := \dfrac{aS^*(t)}{S^*(t) + \Delta}\tilde{N}(t)$ to obtain the equation in $\tilde{S}$:

$$\dot{\tilde{S}} = \rho\Big(1 - \frac{2S^*(t)}{K}\Big)\tilde{S}(t) - F(t), \qquad \tilde{S}(0) = 0.$$

That leads to the solution

$$\tilde{S}(t) = -\int_0^t F(\xi) exp\Big(-\int_0^\xi \rho\Big(1 - \frac{2S^*(\tau)}{K}\Big)d\tau\Big)d\xi \times exp\Big(\int_0^t \rho\Big(1 - \frac{2S^*(\tau)}{K}\Big)d\tau\Big).$$

Assuming that the solutions remain close enough to the PFS over the seasons, we obtain the result by changing $P(0)$ in $P(nT^+) = P(nT)$.

– On $]nT + d, nT + t_f]$, the second subsystem of equation 2 is written

$$\begin{cases} \dot{P}(t) &= -\beta P(t)S(t) + \alpha a(1 - \gamma)\dfrac{S(t)X(t)}{S(t) + \Delta} - \Omega P(t), \\ \dot{S}(t) &= -a\frac{S(t)X(t)}{S(t)+\Delta}, \\ \dot{X}(t) &= \beta P(t)S(t) + \alpha a\gamma\dfrac{S(t)X(t)}{S(t) + \Delta} - \mu X(t). \end{cases} \tag{17}$$

With initial conditions $P(nT + d^+) = 0$, $X(nT + d^+) = N(nT + d)$ and $S(nT + d^+) = S(nT + d)$ from the system (14).

The pest free equilibrium (PFE) can be written $Y_P(t) = \begin{pmatrix} P_p(t) \\ S_p(t) \\ X_p(t) \end{pmatrix} = \begin{pmatrix} 0 \\ S^*(d) \\ 0 \end{pmatrix}$.
Considering the deviation variables $\tilde{P}(t) = P(t) - P_p(t) = P(t), \tilde{S}(t) = S(t) - S_p(t), \tilde{X}(t) = X(t) - X_p(t) = X(t)$, one can write the equation in the new variables as:

$$\begin{cases} \dot{\tilde{P}} &= -\beta\tilde{P}(\tilde{S}+S^*(d)) + \alpha a(1-\gamma)\dfrac{(\tilde{S}+S^*(d))\tilde{X}}{\tilde{S}+S^*(d)+\Delta} - \Omega\tilde{P}, \\ \dot{\tilde{S}} &= -a\dfrac{(\tilde{S}+S^*(d))\tilde{X}}{\tilde{S}+S^*(d)+\Delta}, \\ \dot{\tilde{X}} &= \beta\tilde{P}(\tilde{S}+S^*(d)) + \alpha a\gamma\dfrac{(\tilde{S}+S^*(d))\tilde{X}}{\tilde{S}+S^*(d)+\Delta} - \mu\tilde{X} \end{cases} \tag{18}$$

And the Jacobian matrix $J = \begin{bmatrix} -\beta S^*(d) - \Omega & 0 & \alpha a(1-\gamma)\dfrac{S^*(d)}{S^*(d)+\Delta} \\ 0 & 0 & -a\dfrac{S^*(d)}{S^*(d)+\Delta} \\ \beta S^*(d) & 0 & -\mu + \alpha a\gamma\dfrac{S^*(d)}{S^*(d)+\Delta} \end{bmatrix}$.

In the neighbourhood of the PFE, system 18 is then equivalent to the linearised system

$$\dot{\tilde{Y}} = J.\tilde{Y}, \quad \tilde{Y} = (\tilde{P}, \tilde{S}, \tilde{X}). \tag{19}$$

Since the second column of J is null, one just has to compute the exponential of At that will generate a local solution for $\tilde{P}$ and $\tilde{X}$, where $A := \begin{bmatrix} -\beta S^*(d) - \Omega & \alpha a(1-\gamma)\dfrac{S^*(d)}{S^*(d)+\Delta} \\ \beta S^*(d) & -\mu + \alpha a\gamma\dfrac{S^*(d)}{S^*(d)+\Delta} \end{bmatrix}$.

We deduce $\tilde{S}$ from $\dot{\tilde{S}} = -a\dfrac{S^*(d)}{S^*(d)+\Delta}\tilde{X}$, i.e.

$$\tilde{S}(t) = -a\frac{S_0}{S_0+\Delta}\tilde{X}(nT+t).$$

Since A is a Metzler matrix, it admits two distinct real eigenvalues $\lambda_{1,2} = \dfrac{tr(A)}{2} \pm \dfrac{1}{2}\sqrt{tr^2(A) - 4det(A)}$ and we have $\Pi(t) = \begin{pmatrix} \Pi_{1,1}(t) & \Pi_{1,2}(t) \\ \Pi_{2,1}(t) & \Pi_{2,2}(t) \end{pmatrix}$, where

$$\Pi_{1,1}(t) = \tfrac{1}{\lambda_2-\lambda_1}\Big(e^{\lambda_1 t}(\lambda_2 + \beta S^*(d) + \Omega) - e^{\lambda_2 t}(\lambda_1 + \beta S^*(d) + \Omega)\Big)$$

$$\Pi_{1,2}(t) = -\tfrac{1}{\lambda_2-\lambda_1}\Big(\tfrac{\alpha a(1-\gamma)}{S^*(d)+\Delta}(e^{\lambda_1 t} - e^{\lambda_2 t})\Big)$$

$$\Pi_{2,1}(t) = -\tfrac{1}{\lambda_2-\lambda_1}\Big(\beta S^*(d)(e^{\lambda_1 t} - e^{\lambda_2 t})\Big)$$

$$\Pi_{2,2}(t) = \tfrac{1}{\lambda_2-\lambda_1}\Big(e^{\lambda_1 t}\big(\lambda_2 + \mu - \tfrac{\alpha a S^*(d)}{S^*(d)+\Delta}\big) - e^{\lambda_2 t}\big(\lambda_1 + \mu - \tfrac{\alpha a S^*(d)}{S^*(d)+\Delta}\big)\Big)$$

Appendix 3: Proof of Proposition 4

1) From equation (8), we have

$$N(nT+d) = P(nT)e^{-\mu d + \int_0^d \frac{\alpha a S^*(\tau)}{S^*(\tau)+\Delta} d\tau}.$$

Hence, $N(nT+d) = P(nT)\delta$.

Proposition 3 therefore involves

$$\begin{pmatrix} P(nT+tf) \\ X(nT+t_f) \end{pmatrix} = \begin{pmatrix} \Pi_{1,1}(t_f-d) & \Pi_{1,2}(t_f-d) \\ \Pi_{2,1}(t_f-d) & \Pi_{2,2}(t_f-d) \end{pmatrix} \begin{pmatrix} 0 \\ \delta P(nT) \end{pmatrix}$$

Hence,

$$\begin{cases} P(nT+t_f) &= \Pi_{1,2}(t_f-d).P(nT)\delta \\ X(nT+t_f) &= \Pi_{2,2}(t_f-d).P(nT)\delta \end{cases}$$

So, according to the transition rule (6),

$$P((n+1)T) = \Big[\Pi_{1,2}(t_f-d) + q\Pi_{2,2}(t_f-d)\Big]P(nT)\delta e^{-\Omega\tau}$$

From where we deduce

$$P(nT) = P_0 e^{-n\Omega\tau}\delta^n\big[\Pi_{1,2}(t_f-d) + q\Pi_{2,2}(t_f-d)\big]^n.$$

2) Since $P(nT) \to 0$ iff

$$\big(\Pi_{1,2}(t_f-d) + q\Pi_{2,2}(t_f-2)\big)\delta e^{-\Omega\tau} < 1, \tag{20}$$

We deduce $\mathcal{R}_0 = \big(\Pi_{1,2}(t_f-d) + q\Pi_{2,2}(t_f-2)\big)\delta e^{-\Omega\tau}$.

3) We deduce τ_0 from the condition (20) above, by rearranging as $\tau > \frac{ln\big(\big[\Pi_{1,2}(t_f-d) + q\Pi_{2,2}(t_f-d)\big]\delta\big)}{\Omega} \equiv \tau_0$.

HIV infection

Viral dynamics of a delayed HIV-1 infection model with both virus-to-cell and cell-to-cell transmissions, and CTL immune response delay.

M.L. Mann Manyombe[a,*] –D.F. Nkoa Onana[a] –J. Mbang[a] –S. Bowong[b]

[a] Department of Mathematics, University of Yaounde I, PO Box 812 Yaounde, Cameroon, luthermann.3ml@gmail.com, mbangjoh@yahoo.fr, dnkoa@yahoo.com

[b] Department of Mathematics and Computer Science, Faculty of Science, University of Douala, PO Box 24157 Douala, Cameroon, sbowong@gmail.com

[*] Corresponding author : M.L. Mann Manyombe, luthermann.3ml@gmail.com

RÉSUMÉ. Nous considerons un modèle qui décrit la dynamique de l'infection du VIH et, qui tient compte des transmissons virus-cellule et cellule-cellule, de la réponse immunitaire. Ce modèle inclut quatre retards continus qui décrivent respectivement: la latence pour l'infection virus-cellules, l'infection cellule-cellule, la production de nouveaux virions et l'activation de la réponse immunitaire. Quelques innovations de ce modèle sont l'inclusion d'un taux de production des cellules CTL issue du thymus et du retard d'activation de la reponse immunitaire. Nous déterminons le taux de reproduction de base $\mathcal{R}_0$ et montrons que la dynamique global est completement déterminé par la valeur de $\mathcal{R}_0$. Nous montrons que si $\mathcal{R}_0 \leq 1$ alors l'infection peut être éliminé ; alors que si $\mathcal{R}_0 > 1$, il existe un équilibre endémique, et, le système est persistent. Des simulations numériques indiquent que les retards intracellulaires et le retard de la réponse immunitaire peuvent stabilisé et/ou destabilisé l'équilibre endémique.

ABSTRACT. We consider a mathematical model that describes a viral infection of HIV-1 with both virus-to-cell and cell-to-cell transmission, CTL response immune and four distributed delays, in which the first, second and fourth distributed delay respectively describe the intracellular latency for virus-to-cell infection, the intracellular for the cell-to-cell infection and the time period that viruses penetrated into cells and infected cells release new virions, and the third delay describes the activation delay of CTLs cells. One of the main features of the model is that it includes a constant production rate of CTLs export from thymus, and an immune response delay. We derive the basic reproduction number $\mathcal{R}_0$ and establish that the global dynamics is completely determined by the values of $\mathcal{R}_0$. We show that if $\mathcal{R}_0 \leq 1$, then the infection free equilibrium is globally asymptotically stable, meaning that HIV virus can be cleared ; whereas, if $\mathcal{R}_0 > 1$, then there exist a chronic infection equilibrium, and the HIV-1 infection will persist in the host. Numerical simulations indicate that the intracellular delays and immune response delay can stabilize and/or destabilize the chronic infection equilibrium.

MOTS-CLÉS : Dynamique viral, Retards continus, Réponse immunitaire, Persistence

KEYWORDS : Viral dynamics, Distributed delays, CTL immune response, Persistence

1. Introduction

Over the recent years, great efforts have been paid in mathematical modeling of within-host virus dynamics. Mathematical models and their analysis are helpful in understanding the dynamical behavior of many human viruses such as HIV, HTLV-I and HBV (e.g., [2, 3, 4, 5, 6, 8]). Recently, it has been reported that the uninfected cells can also become infected because of direct contact with infected cells. The viral infection model with cell-to-cell transmission and distributed time delay have been proposed in [2, 3, 6, 7]. They observed that the basic reproduction number of their model might be underevaluated if either cell-to-cell spread or virus-to-cell infection is neglected.

Note that the immune response after viral infection is common and is necessary for eliminating or controlling the disease. In most virus infections, cytotoxic T lymphocytes (CTLs) play a critical role in antiviral defense by attacking virus-infected cell. Many existing mathematical models for HIV infection with CTLs response are given by systems of ordinary differential equation (ODE) (see, e.g. [2, 4, 5, 6, 8]). However, time delays can not be ignored when modeling immune response, since antigenic stimulation generating CTLs may need a period of time, that is, the activation rate of CTL response at time t may depend on the population of antigen at a previous time [8]. Moreover, all the aforementioned works not take into account of the constant production rate of CTLs exported from thymus. This consideration of export rate of new CTLs from thymus is considered in [4, 5] and is ignored by many authors.

Motivated by the works in [4, 7], in the present paper, we are concerned by the effect of both virus-to-cell and cell-to-cell transmissions with intracellular delays, and immune response activation delay on the global dynamics of HIV-1 infection model. We consider a within-host viral infection model with both virus-to-cell and cell-to-cell transmissions, immune response and four distributed delays, in which the first, second and fourth delay respectively describes the intracellular latency for virus-to-cell infection, the intracellular latency for the cell-to-cell infection and the time period that viruses penetrated into cells and infected cells release new virions [7], and the third delay describes the activation delay of CTLs cells ([8]). The rest of the paper is organized as follows. In Section 2, the mathematical model is constructed, the preliminaries including the positivity and boundedness of solutions are introduced, the existence of an infection-free equilibrium and its global stability are obtained, the existence of a chronic infection equilibrium and the persistence of infection are also obtained. In section 3, numerical simulations for several cases of the main model are presented. Section 4 concludes the paper.

2. The model formulation

The compartmental model includes the concentrations of healthy target cells $T(t)$ which susceptible to infection, infected cells $T_i(t)$ that produces viruses, cytotoxic T lym-

phocytes (CTLs) cells $T_c(t)$ which are responsible of the destruction of infected cells and viruses $V(t)$. Let β_1 be the virus-to-cell infection rate, β_2 be the cell-to-cell infection rate, δ, μ_1, α and c be death rates of healthy target cells, activated infected cells, cytotoxic CTLs cells and viruses, respectively. Let b be the production rate of healthy target cells, λ be the production rate of CTLs cells export from thymus, a be the proliferation rate of CTLs cells. Infected cells are eliminated by CTLs cells at a rate q, which represent the lytic activity of CTLs cells. $e^{-\mu_1 s_1}$ is the survival rate of cells that are infected by viruses at time t and become activated s_1 time later with a probability distribution $f_1(s_1)$. Then $\int_0^\infty \beta_1 T(t-s_1)V(t-s_1)f_1(s_1)e^{-\mu_1 s_1}ds_1$ describes the newly activated infected target cells which are infected by free viruses s_1 time ago [7]. Similarly, $\int_0^\infty \beta_2 T(t-s_2)T_i(t-s_2)f_2(s_2)e^{-\mu_1 s_2}ds_2$ represents the newly activated infected target cells which are infected by infected cells s_2 time ago [7]. $e^{-\mu_2 s_3}$ is the survival rate of CTLs cells that are activated at time t, and become cytotoxic s_3 time later with a probability distribution $f_3(s_3)$. Then, $\int_0^\infty aT_i(t-s_3)T_c(t-s_3)f_3(s_3)e^{-\mu_2 s_3}ds_3$ represents the newly CTLs cells proliferated at time t [8]. Let s_4 be the random variable that is the time between viral RNA transcript and viral release and maturation with a probability distribution $f_4(s_4)$. Then, $\int_0^\infty kT_i(t-s_4)f_4(s_4)e^{-\mu_3 s_4}ds_4$ describes the mature viral particles produced at time t [7]. k is the average number of viruses that bud out from an infected cell and $e^{-\mu_3 s_4}$ is the survival rates of cells that start budding from activated infected cells at time t and become free mature viruses s_4 time later. Note that s_1,s_2, s_3 and s_4 are all integration variables, without loss of generality, they all will be represented by s. The model is given as follows :

$$\begin{cases} \frac{dT(t)}{dt} &= b-\delta T-\beta_1 TV-\beta_2 TT_i \\ \frac{dT_i(t)}{dt} &= \int_0^\infty \beta_1 T(t-s)V(t-s)f_1(s)e^{-\mu_1 s}ds \\ & \quad +\int_0^\infty \beta_2 T(t-s)T_i(t-s)f_2(s)e^{-\mu_1 s}ds-\mu_1 T_i-qT_iT_c \\ \frac{dT_c(t)}{dt} &= \lambda+a\int_0^\infty T_i(t-s)T_c(t-s)f_3(s)e^{-\mu_2 s}ds-\alpha T_c \\ \frac{dV(t)}{dt} &= k\int_0^\infty T_i(t-s)f_4(s)e^{-\mu_3 s}ds-cV, \end{cases} \tag{1}$$

$f_i(\nu):[0,\infty)\longrightarrow[0,\infty)$ are probability distributions with compact support, $f_i(\nu)\geqslant 0$, and $\int_0^\infty f_i(\nu)d\nu=1,\ i=1,\ldots,4$.

From the modeling perspective, the model (1) extends the basic model developed in [4] by : **(i)** incorporating the cell-to-cell transmission, **(ii)** intracellular delays and **(iii)** immune activation delay. Together with this latter improvement **(iii)**, the incorporation

of a constant production rate of CTLs export from thymus in our model also extend the works in [2, 6, 8]. It is also noticeable that, our model extends the models developed in [3, 7] by including CTL response immune delay.

2.1. Preliminaries

Define the Banach space of fading memory type (see [3, 7])
$\mathcal{C} = \left\{\phi \in C((-\infty, 0]|\phi(\theta)\, e^{\mu\theta} \text{ is continuous for } \theta \in (-\infty, 0] \text{ and } \|\phi\| < \infty\right\}$ where μ is positive constant and the norm $\|\phi\| = \sup_{\theta\leqslant 0} |\phi(\theta)|\, e^{\mu\theta}$. The nonnegative cone of $\mathcal{C}$ is defined by $\mathcal{C}_+ = \mathcal{C}((-\infty, 0], \mathbb{R}_+)$. For $\phi \in \mathcal{C}$, Let $\phi_t(\theta) = \phi(t + \theta)$, $\theta \in (-\infty, 0]$. We consider solutions $(T,\, T_i,\, T_c,\, V)$ of system (1) with initial conditions

$$(T(0),\, T_i(0),\, T_c(0),\, V(0)) \in X := \mathcal{C}_+ \times \mathcal{C}_+ \times \mathcal{C}_+ \times \mathcal{C}_+. \tag{2}$$

By the standard theory of functional differential equations, we can obtain the existence of solutions for $t > 0$. Let $\eta_i = \int_0^\infty e^{-\mu_1 s} f_i(s)ds$, $i = 1, 2$, $\eta_3 = \int_0^\infty f_3(s)e^{-\mu_3 s}ds$, $\eta_4 = \int_0^\infty f_4(s)e^{-\mu_4 s}ds$.

Theorem 2.1 *Solutions of system (1) with initial conditions (2) are positive and ultimately uniformly bounded for $t > 0$.*

Proof 2.1 *The proof of Theorem 2.1 is given in Appendix A.* □

Theorem 2.1 implies that omega limit sets of system(1) are contained in the following bounded feasible region :

$$\Omega = \left\{(T, T_i, T_c, V) \in \mathcal{C}_+^4 : \|T_s\| \leq \frac{b}{\delta},\, \|T_i\| \leq M_1,\, \frac{\lambda}{\alpha} \leq T_c \leq \frac{a}{q}M_3,\, \|V\| \leq M_2\right\}.$$

It can be verified that the region Ω is positively invariant with respect (1) and the system is well posed.

2.2. The infection-free equilibrium and its stability

System (1) has an infection-free equilibrium $E_0 = \left(\frac{b}{\delta}, 0, \frac{\lambda}{\alpha}, 0\right)$. We defined the basic reproduction number as follows :

$$\mathcal{R}_0 = \mathcal{R}_{01} + \mathcal{R}_{02} = \frac{k\,\beta_1\, b\,\eta_1\,\eta_4}{c\delta\left(\mu_1 + \frac{q\lambda}{\alpha}\right)} + \frac{\beta_2 b\eta_2}{\delta\left(\mu_1 + \frac{q\lambda}{\alpha}\right)},$$

which represents the average number of secondary infections. In fact, $\frac{k\,\beta_1\, b\,\eta_1\,\eta_4}{c\delta\left(\mu_1+\frac{q\lambda}{\alpha}\right)}$ is the average number of secondary viruses caused by a virus, that is the basic reproduction number corresponding to virus-to-cell infection mode, while $\frac{\beta_2 b\eta_2}{\delta\left(\mu_1+\frac{q\lambda}{\alpha}\right)}$ is the average number of secondary infected cells that caused by an infected cell, that is the basic reproduction number corresponding to cell-to-cell infection mode. The factors have the biological interpretations as follows :

– $\frac{b\beta_1\eta_1}{\delta}$ is the number of new infections caused by a virus in target susceptible cells ;
– $\frac{q\lambda}{\alpha}$ is the rate at which infected cells are eliminated by the CTLs response ;
– $\frac{1}{\mu_1+\frac{q\lambda}{\alpha}}$ is the average time that an infectious cell survives ;
– $k\eta_4$ is the rate at which infected cells bud into viruses ;
– $\frac{1}{c}$ is gives the average life-span of a virus ;
– $\frac{b\beta_2\eta_2}{\mu_1+\frac{q\lambda}{\alpha}}$ represents the number of new infections caused by an infected cell in target susceptible cells.

The result below follows is straightforward.

Theorem 2.2 *The infection-free equilibrium E_0 of system (1) is locally asymptotically stable in the feasible region Ω whenever $\mathcal{R}_0 < 1$ and unstable otherwise.*

Proof 2.2 *The characteristic equation of system (1) at the equilibrium E_0 is*

$$(\nu+\delta)(\nu+\alpha)\left[(\nu+c)\left(\nu+\mu_1+\frac{q\lambda}{\alpha}-\frac{b\beta_2\overline{\eta}_2}{\delta}\right)-\frac{kb\beta_1}{\delta}\overline{\eta}_1\overline{\eta}_4\right]=0, \qquad (3)$$

where $\overline{\eta}_i = \int_0^\infty e^{-(\mu_1+\nu)s}f_i(s)ds$, $i = 1, 2$, $\overline{\eta}_3 = \int_0^\infty e^{-(\mu_2+\nu)s}f_3(s)ds$ and $\overline{\eta}_4 = \int_0^\infty e^{-(\mu_3+\nu)s}f_4(s)ds$. We see that (3) has eigenvalues $\nu_1 = -\delta$, $\nu_2 = -\alpha$ and other eigenvalues are determined by $(\nu+c)\left(\nu+\mu_1+\frac{q\lambda}{\alpha}-\frac{b\beta_2\overline{\eta}_2}{\delta}\right)-\frac{kb\beta_1}{\delta}\overline{\eta}_1\overline{\eta}_4=0$, which equivalent to

$$\Psi(\nu) := \left(\frac{\nu}{\mu_1+\frac{q\lambda}{\alpha}}+1\right)(\nu+c)-\mathcal{R}_0\left(\frac{\overline{\eta}_2\mathcal{R}_{02}}{\eta_2\mathcal{R}_0}\nu+c\frac{\overline{\eta}_2\mathcal{R}_{02}}{\eta_2\mathcal{R}_0}+c\frac{\overline{\eta}_1\overline{\eta}_4\mathcal{R}_{01}}{\eta_1\eta_4\mathcal{R}_0}\right)=0. \quad (4)$$

Thus, $\Psi(0) = c(1-\mathcal{R}_0) < 0$ when $\mathcal{R}_0 > 1$. Note that $\overline{\eta}_1 \le \int_0^\infty f_1(s)ds = 1$, $i = 1, 2, 3, 4$. Then, we have $\Psi(\nu) \ge \left(\frac{\nu}{\mu_1+\frac{q\lambda}{\alpha}}+1\right)(\nu+c)-\mathcal{R}_0\left(\frac{\mathcal{R}_{02}}{\eta_2\mathcal{R}_0}\nu+c\frac{\mathcal{R}_{02}}{\eta_2\mathcal{R}_0}+c\frac{\mathcal{R}_{01}}{\eta_1\eta_4\mathcal{R}_0}\right) \to +\infty$ as $\nu \to +\infty$. This yields that equation (4) has at least one positive root. Therefore, the infection-free equilibrium E_0 is unstable if $\mathcal{R}_0 > 1$. □

Biologically speaking, Theorem 2.2 implies that infection can be eliminated if the initial sizes of cells are in the basin of attraction of the infection-free equilibrium. Thus, the infection can be effectively controlled if $\mathcal{R}_0 < 1$. One can remark that $\mathcal{R}_0$ depends on λ and is a decreasing function of this rate. Hence, the constant rate λ could be an important control parameter in order to reduce $\mathcal{R}_0$ to a value less than unity. To ensure that the effective control of the infection is independent of the initial size of the cells, a global stability result must be established for the infection-free equilibrium.

Theorem 2.3 *If $\mathcal{R}_0 \le 1$, then the infection-free equilibrium E_0 of system (1) is globally asymptotically stable in Ω.*

Proof 2.3 *The proof of Theorem 2.3 is given in Appendix B.* □

2.3. The chronic infection equilibrium and persistence of infection

In this section, we will show that there exists a chronic infection equilibrium and the model (1) is persistent when $\mathcal{R}_0 > 1$. The infection is endemic if the infected cells persist above a certain positive level.

Denote by $E^* = (T^*, T_i^*, T_c^*, V^*)$ the chronic infection equilibrium of system (1). Then

$$T^* = \frac{b}{\delta + \left(\beta_1 + \frac{\beta_2 c}{k\eta_4}\right) V^*}, T_i^* = \frac{cV^*}{k\eta_4}, T_c^* = \frac{kb\beta_1\eta_1\eta_4 + b\beta_2\eta_2 c - c\mu_1\delta - c\mu_1\left(\beta_1 + \frac{\beta_2 c}{k\eta_4}\right)V^*}{qc\left(\delta + \beta_1 V^* + \frac{\beta_2 c}{k\eta_4}V^*\right)}, \tag{5}$$

where V^* is a positive root of $\lambda + a\eta_3 T_i^* T_c^* - \alpha T_c^* = 0$ $(\star)$. After expansion and substitution of T^*, T_i^*, T_c^* by their expressions, Eq. $(\star)$ is equivalent to polynomial $P(V) = a_2V^2 + a_1V + a_0 = 0$, with the coefficients a_2, a_1 and a_0 given by

$$\begin{aligned} a_2 &= \frac{a\eta_3 c^2\mu_1}{k\eta_4}\left(\beta_1 + \frac{\beta_2 c}{k\eta_4}\right), \\ a_1 &= \frac{a\eta_3 c^2\mu_1\delta}{k\eta_4} - a\eta_3 bc\left(\beta_1\eta_1 + \frac{\beta_2\eta_2 c}{k\eta_4}\right) - \left(\beta_1 + \frac{\beta_2 c}{k\eta_4}\right)(\mu_1 c\alpha + q\lambda c), \\ a_0 &= c\delta\alpha\left(\mu_1 + \frac{q\lambda}{\alpha}\right)(\mathcal{R}_0 - 1). \end{aligned} \tag{6}$$

Using $T_c \geq 0$ one shows that $V \leq V_{max}$, where $V_{max} = \frac{kb\beta_1\eta_1\eta_4 + b\beta_2\eta_2 c - c\mu_1\delta}{c\mu_1\left(\beta_1 + \frac{\beta_2 c}{k\eta_4}\right)} = \frac{c\mu_1\delta(\mathcal{R}_0 - 1) + \mathcal{R}_0\frac{c\delta q\lambda}{\alpha}}{c\mu_1\left(\beta_1 + \frac{\beta_2 c}{k\eta_4}\right)} > 0$, since $\mathcal{R}_0 > 1$. Using $P(0)$ and $P(V_{max})$ one shows that $P(0)P(V_{max}) = -\frac{a_0 b\lambda q(k\beta_1\eta_1\eta_4 + \beta_2\eta_2 c)}{\mu_1} < 0$. Since P is continuous and strictly decreasing on interval $]0; V_{max}[$, the intermediate value theorem implies that P vanishes on $]0; V_{max}[$, which proves the existence and uniqueness of a positive chronic infection equilibrium when $\mathcal{R}_0 > 1$.

In the following, we will shows that the model (1) is persistent when $\mathcal{R}_0 > 1$. To achieve our goal, we will apply Theorem 4.2 in [1]. To this end, let $S(t)$, $t > 0$, be the solution semiflow of model (1), we can prove the following persistence result for (1).

Theorem 2.4 *For system (1), if $\mathcal{R}_0 > 1$, then the solution semiflow $S(t)$ is uniformly persistent ; that is, there exists a $\sigma > 0$ such that any solution of (1) satisfies*

$$\liminf_{t\to\infty} T(t) \geq \sigma, \quad \liminf_{t\to\infty} T_i(t) \geq \sigma, \quad \liminf_{t\to\infty} T_c(t) \geq \sigma, \quad \liminf_{t\to\infty} V(t) \geq \sigma.$$

Proof 2.4 *The proof of Theorem 2.4 is given in Appendix C.* □

3. Numerical simulations

In this section, we perform numerical simulations for the model (1) with particular distribution functions $f_i(s)$, $i = 1, 2, 3, 4$ as : $f_1(s) = f_2(s) = \delta(s - s_1)$, $f_3(s) = \delta(s - s_3)$ and $f_4(s) = \delta(s - s_4)$, where $\delta(\cdot)$ is the dirac delta function, s_i, $i = 1, 2, 3, 4$ are positive constants. Then, we can see that $\eta_1 = \eta_2 = e^{-\mu_1 s_1}$, $\eta_3 = e^{-\mu_2 s_3}$ and $\eta_4 = e^{-\mu_3 s_4}$. We examine the behavior of the infected steady state E^* using data sets that are commonly used in the literature [4, 6, 7]. Values of parameters are defined as : $b = 10$, $\delta = 0.01$, $\beta_1 = 2e - 6$, $\beta_2 = 3e - 4$, $\mu_1 = 0.1$, $a = 3e - 2$, $q = 2e - 4$, $k = 100$, $\alpha = 0.02$, $c = 3.2$, $\lambda = 1$, $\mu_2 = 0.5$ and $\mu_3 = 0.1$. By simple computing, the persistence of the infection when $\mathcal{R}_0 > 1$ as demonstrated in Theorem 2.4 is numerically shown on Figure 1.

3.1. Effect of CTLs constant production rate

In order to investigate the effect of CTLs production rate, we carry out some numerical simulations to show the contribution of CTLs constant production rate during the whole infection. We set the production rate λ as 0.5, 1, 1.5, 2. We choose $s_1 = s_2 = 3$, $s_3 = 8$ and $s_4 = 2.5$. From the four figures of Figure 1, we can observe that uninfected and CTLs cells reach a higher peak level as λ increases. While, the peak level of infected cells and viruses decreases as λ increases. If we interpret the constant rate $\lambda > 0$ as an inflow of antiviral drugs, one can observe from Figure 1 that the entry of antiviral drugs into the host is important as a control parameter in order to reduce the viral load.

3.2. Effect of intracellular delays and immune response delay on the stability of steady states

In this case, we choose $s_4 = 2.5$ and without loss of generality, we let $S = s_1 = s_2$. Figure 2 plots the chronic infection equilibrium E^* when S varies and $s_3 = 5$ is fixed (left column), and when s_3 is varied and $S = 3$ is fixed (right column). This figure demonstrates that the chronic equilibrium destabilizes as S and s_3 decreases. Therefore, an increase in the intracellular delay S or the immune response delay can stabilize the infected steady state E^*. In the instabilities cases, one observe oscillation patterns where a larger viral peak is generated before the viral load and the infected cells dynamics are "trapped" by the invariant plan $T_i = V = 0$. These transient viral peaks strongly resemble viral load blips clinically observed in HIV-infected patients, and they provide an alternative interpretation of these phenomena. This result is consistent with the study in [4].

4. Conclusion

In this paper, we have investigated the dynamical properties of a delayed HIV-1 infection model with both virus-to-cell and cell-to-cell transmissions, and CTL immune response delay. This model extends some previous models and also take into account of a rate of CTLs cells exported from thymus. We have derived the basic reproductive number, $\mathcal{R}_0$, which depends on $\frac{q\lambda}{\alpha}$ (it is the rate at which infected cells with virus are eliminated by the CTLs response), that can contribute to the control of viral infection. When the basic reproductive number $\mathcal{R}_0$ is less than unity, we have proved the global asymptotic stability of the disease free equilibrium E_0. When the basic reproductive number $\mathcal{R}_0$ is greater than unity, the persistence of the chronic infection equilibrium E^* has been obtained. It is challenging to analyze model (1) for the joint effect of four delays theoretically. So, numerical simulations were used to further investigate the infected steady state and the existence of the Hopf bifurcation when $s_i > 0$, $i = 1, 2, 3, 4$. Notice that the existence of the Hopf bifurcation contributes at the emergence of viral load blips which is clinically observed in HIV-infected patients. It was found that the intracellular delays and immune response delay can stabilize and/or destabilize the chronic infection equilibrium (see Figure 2).

5. Bibliographie

[1] J. K. Hale, P. Waltman, Persistence in infinite-dimensional systems, *SIAM J. Math. Anal.*, **20** (1989), 388–395.

[2] A.M. Elaiw, A.A. Raezah, A.S Alofi, Stability of delay-distributed virus dynamics model with cell-to-cell transmission and CTL immune response, *J. Comp. Anal. Appl.*, **25** (2018), 1518–1531.

[3] X. Lai, X. Zou, Modeling HIV-1 virus dynamics with both virus-to-cell infection and cell-to-cell transmission, *SIAM J. Math. Anal.*, **74** (2014), 898–917.

[4] D.F. Nkoa Onana, B. Mewoli, D.A. Ouattara, Excitability in the host-pathogen interactions of HIV infection and emergence of viral load blips, *J. theor. Biol.*, **317** (2013), 407–417.

[5] C. Vargas-De-Leon, Global properties for a virus dynamics model with lytic and non-lytic immune responses, and nonlinear immune attack rates, *J. Biol. Syst*, **22** (2014), 1–14.

[6] J. Wang et al., Threshold dynamics of HIV-1 virus model with cell-to-cell transmission, cell-mediated immune responses and distributed delay, *Appl. Math. Comp.*, **291** (2016), 149–161.

[7] Y. Yang, L. Zou, S. Ruan, Global dynamics of a delayed within-host viral infection model with both virus-to-cell and cell-to-cell transmissions, *Math. Biosci.*, **270** (2015), 183–191.

[8] Z. Yuan, X. Zou, Global threshold dynamics in an HIV virus model with nonlinear infection rate and distributed invasion and production delays, *Math. Biosc. Eng.*, **10** (2013), 483–498.

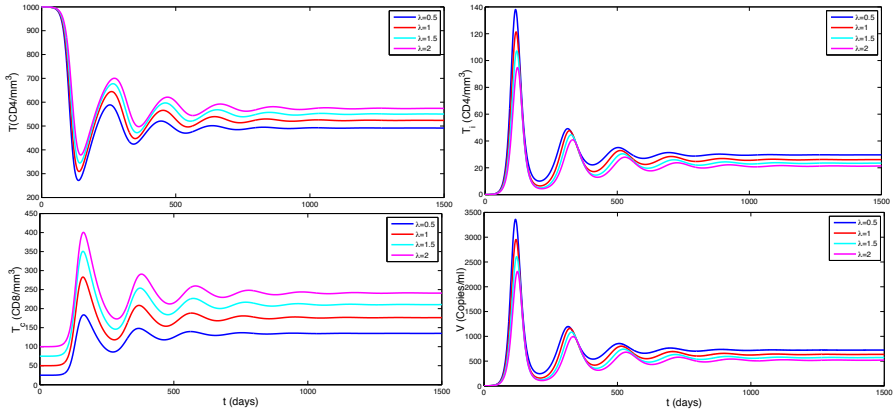


Figure 1 – Simulation results showing the effect of λ on the dynamics of the model with $s_1 = s_2 = 3$, $s_3 = 8$ and $s_4 = 2.5$.

Appendix A : Proof of Theorem 2.1

Let $m(t) = \delta + \beta_1 V(t) + \beta_2 T_i(t)$ and $d(t) = \mu_1 + qT_c(t)$. Let $r(t)$ be the sum of the two integral terms in the second equation of system (1) and $n(t)$ be the integral term in the fourth equation of system (1). From the first equation in (1), we then have $T(t) = T(0)e^{-\int_0^t m(\xi)d\xi} + \int_0^t e^{-\int_\xi^t m(\theta)d\theta} bd\xi > 0$ for $t \geq 0$. From the third equation in (1), it follows that $\lim\limits_{t\to\infty} \inf T_c(t) \geq \frac{\lambda}{\alpha} > 0$. From the second and fourth equation in (1), we then have $T_i(t) = T_i(0)e^{-\int_0^t d(\xi)d\xi} + \int_0^t r(\xi)e^{-\int_\xi^t d(\theta)d\theta}d\xi$ and $V(t) = \left[V(0) + \int_0^t n(\xi)e^{c\xi}d\xi\right] e^{-ct}$, which yield that $T_i(t) > 0$, $V(t) > 0$ for small $t > 0$. Now we prove that $T_i(t) > 0$ and $V(t) > 0$ for all $t > 0$. Otherwise, there exists $t_1 > 0$ such that $\min\{T_i(t_1), V(t_1)\} = 0$. If $T_i(t_1) = 0$, $T_i(t) > 0$ for $0 \leq t < t_1$, and $V(t) > 0$ for $0 \leq t < t_1$, then we have $\frac{dT_i(t_1)}{dt} > 0$. This contradicts $T_i(t_1) = 0$ and $T_i(t) > 0$ for $0 \leq t < t_1$. If $V(t_1) = 0$, $V(t) > 0$ for $0 \leq t < t_1$, and $T_i(t) > 0$ for $0 \leq t < t_1$, then we obtain $\frac{dV(t_1)}{dt} > 0$, which is also a contradiction. Hence, $T_i(t) > 0$ and $V(t) > 0$ for all $t > 0$.

To prove boundedness, first by the positivity of solutions we have $\frac{dT(t)}{dt} < b - \delta T(t)$. It follows that $\lim_{t\longrightarrow\infty} \sup T(t) \leq \frac{b}{\delta}$, implying $T_s(t)$ is bounded. Let $G_1(t) = \int_0^\infty f_1(s)e^{-\mu_1 s}T(t-s)ds + \int_0^\infty f_2(s)e^{-\mu_1 s}T(t-s)ds + T_i(t)$. Since $T(t)$ is bounded and $\int_0^\infty f(u)du$ is convergent, the integral in $G(t)$ is well defined and differentiable

with respect to t. Moreover , when taking the time derivative of $G(t)$, the order of the differentiation and integration can be switched. Thus, we have

$$\begin{aligned}\dot{G}_1(t) &= b(\eta_1+\eta_2) - \delta\int_0^\infty f_1(s)e^{-\mu_1 s}T(t-s)ds - \delta\int_0^\infty f_2(s)e^{-\mu_1 s}T(t-s)ds \\ &\quad -\mu_1 T_i - qT_iT_c, \\ &\leq b(\eta_1+\eta_2) - \delta\int_0^\infty f_1(s)e^{-\mu_1 s}T(t-s)ds - \delta\int_0^\infty f_2(s)e^{-\mu_1 s}T(t-s)ds \\ &\quad -\left(\mu_1+\frac{q\lambda}{\alpha}\right)T_i(t) \;\leq\; b(\eta_1+\eta_2) - d_1G_1(t),\end{aligned}$$

where $d_1 = \min\left\{\delta, \mu_1 + \frac{q\lambda}{\alpha}\right\}$. Therefore, $\lim\limits_{t\to\infty}\sup G_1(t) \leq \frac{b(\eta_1+\eta_2)}{d_1} := M_1$, implying that $\lim\limits_{t\to\infty}\sup T_i(t) \leq M_1$. Then, from the fourth equation of system (1), we have

$$\dot{V}(t) = k\int_0^\infty e^{-\mu_4 s}f_4(s)T_i(t-s)ds - cV \;\leq\; kM_1\eta_4 - cV.$$

Thus, $\lim\limits_{t\to\infty}\sup V(t) \leq \frac{kM_1\eta_4}{c} := M_2$. Now determine the upper bound of $T_c(t)$. Let $G_2(t) = \int_0^\infty f_3(s)e^{-\mu_3 s}T_i(t-s)ds + \frac{q}{a}T_c(t)$. Thus, we have

$$\begin{aligned}\dot{G}_2(t) &= \int_0^\infty f_3(s)e^{-\mu_3 s}r(t-s)ds - \mu_1\int_0^\infty f_3(s)e^{-\mu_3 s}T_i(t-s)ds + \frac{q\lambda}{a} - \alpha\frac{q}{a}T_c(t), \\ &\leq \frac{b\eta_3}{\delta}(\beta_1\eta_1M_2 + \beta_2\eta_2M_1) + \frac{q\lambda}{a} - \mu_1\int_0^\infty f_3(s)e^{-\mu_3 s}T_i(t-s)ds - \alpha\frac{q}{a}T_c(t), \\ &\leq d_2 - d_3G_2(t),\end{aligned}$$

where $d_2 = \frac{b\eta_3}{\delta}(\beta_1\eta_1M_2 + \beta_2\eta_2M_1) + \frac{q\lambda}{a}$ and $d_3 = \{\alpha, \mu_1\}$. Hence, $\lim\limits_{t\to\infty}\sup G_2(t) \leq \frac{d_2}{d_3} := M_3$, implying that $\lim\limits_{t\to\infty}\sup T_c(t) \leq \frac{a}{q}M_3$. Thus, $T(t)$, $T_i(t)$, $T_c(t)$ and $V(t)$ are uniformly bounded. □

Appendix B : Proof of Theorem 2.3

We define a Lyapunov function as follows :

$$\begin{aligned}L(t) &= T_i + \frac{b\beta_1\eta_1}{c\delta}V + \int_0^\infty f_1(s)e^{-\mu_1 s}\int_{t-s}^t \beta_1T(\tau)V(\tau)d\tau ds + \int_0^\infty f_2(s)e^{-\mu_1 s} \\ &\quad \int_{t-s}^t \beta_2T(\tau)T_i(\tau)d\tau ds + \frac{b\beta_1\eta_1}{c\delta}\int_0^\infty f_4(s)e^{-\mu_3 s}\int_{t-s}^t kT_i(\tau)d\tau ds.\end{aligned}$$

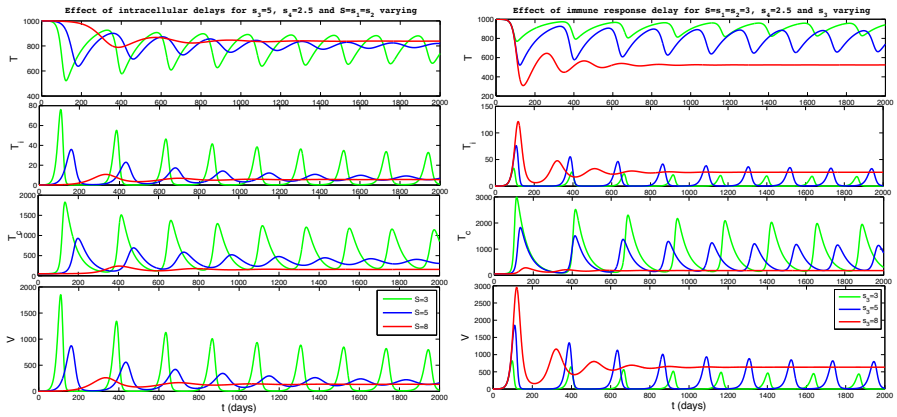


Figure 2 – Simulation results showing the effect of S and s_3 on the dynamics of the model.

Then the time derivative of $L(t)$ along solutions of system (1) satisfies

$$\frac{dL(t)}{dt} = \beta_1\eta_1 TV + \beta_2\eta_2 TT_i + \frac{kb\beta_1\eta_1\eta_4}{c\delta}T_i - \mu_1 T_i - qT_iT_c - \frac{b\beta_1\eta_1}{\delta}V.$$

Since $T \leq \frac{b}{\delta}$ and $T_c \geq \frac{\lambda}{\alpha}$, we have

$$\frac{dL(t)}{dt} \leq \left[\frac{b\beta_2\eta_2}{\delta} + \frac{kb\beta_1\eta_1\eta_4}{c\delta} - \left(\mu_1 + \frac{q\lambda}{\alpha}\right)\right]T_i = \left(\mu_1 + \frac{q\lambda}{\alpha}\right)(\mathcal{R}_0 - 1)T_i.$$

$\frac{dL(t)}{dt} \leq 0$ whenever $\mathcal{R}_0 \leq 1$. Moreover, $\frac{dL(t)}{dt} = 0 \Leftrightarrow T_i = V = 0$ or $T = \frac{b}{\delta}, T_c = \frac{\lambda}{\alpha}$ and $\mathcal{R}_0 = 1$. Thus, the largest invariant set $\mathcal{H}$ such as $\mathcal{H} \subset \left\{(T, T_i, T_c, V) \in \mathbb{R}^4_+ / \frac{dL(t)}{dt} = 0\right\}$ is the singleton $\{E_0\}$. By LaSalle's Principle, E_0 is globally asymptotically stable in Ω, completing the proof.

Appendix C : Proof of Theorem 2.4

Let $\mathcal{D}^0 = \{\phi = (\phi_1, \phi_2, \phi_3, \phi_4) \in X : \phi_2(\theta) > 0 \text{ or } \phi_4(\theta) > 0, \text{ for all } \theta \in (-\infty, \theta]\}$ and $\mathcal{D}_0 = X \setminus \mathcal{D}^0$. We just need to verify the conditions $(i) - (vii)$ of Theorem 4.2 in [1]. It is easy to verify that $X = \mathcal{D}^0 \cup \mathcal{D}_0$, $\mathcal{D}^0 \cap \mathcal{D}_0 = \emptyset$, and $S(t)\mathcal{D}^0 \subset \mathcal{D}^0$, $S(t)\mathcal{D}_0 \subset \mathcal{D}_0$ for all $t > 0$. Furthermore, from Theorem 2.1, we know that $S(t)$ is point dissipative in X. Notice that the boundedness of each component does not depend on the initial condition (2). Thus, for any bounded set Y in X, the positive orbit $\gamma^+(Y) = \bigcup_{t>0} S(t)(Y)$ through $Y \subset X$ is bounded in X. In view of this property, $S(t)$ is asymptotically smooth, that is, for any nonempty bounded set $Y \subset X$ with $S(t)Y \subset Y$, there is a compact set $Y_0 \subset Y$

such that Y_0 attracts Y. Let A_0 be the global attractor of $S(t)$ restricted to $\mathcal{D}_0$. We have $\mathcal{A} = \bigcup_{x\in A_0} w(x) = E_0$. $\{E_0\}$ is a compact and isolated invariant set. Thus, the covering is simply $\{E_0\}$, which is acyclic because no orbit connects E_0 to itself in $\mathcal{D}_0$.

Next, we will verify that $W^s(E_0)\cap\mathcal{D}^0 = \emptyset$. To this end, we suppose the opposite, that is, there exists a solution $u_t \in \mathcal{D}^0$ such that $\lim_{t\to\infty} T(t) = \frac{b}{\delta}$, $\lim_{t\to\infty} T_i(t) = 0$, $\lim_{t\to\infty} T_c(t) = \frac{\lambda}{\alpha}$, $\lim_{t\to\infty} V(t) = 0$. Note that $\mathcal{R}_0 > 1$ is equivalent to $\frac{b}{\delta}\left[\frac{k\beta_1\eta_1\eta_4}{c} + \beta_2\eta_2\right] > \mu_1 + \frac{q\lambda}{\alpha}$. For a small enough $\epsilon > 0$, we have $\left(\frac{b}{\delta} - \epsilon\right)\left[\frac{k\beta_1\eta_1\eta_4}{c} + \beta_2\eta_2\right] > \mu_1 + \frac{q\lambda}{\alpha}$ $(\star\star)$. For this ϵ, there exists a $\tau_0 > 0$ such that $T(t) > \frac{b}{\delta} - \epsilon$ for all $t > \tau_0$. Truncating the integral of η_1, η_2 and η_4 in $(\star\star)$, there is another $\tau_1 > 0$ such that

$$\left(\frac{b}{\delta} - \epsilon\right)\left[\frac{k\beta_1\widetilde{\eta}_1\widetilde{\eta}_4}{c} + \beta_2\widetilde{\eta}_2\right] > \mu_1 + \frac{q\lambda}{\alpha}, \tag{7}$$

where $\widetilde{\eta}_i = \int_0^{\tau_1} e^{-\mu_1 s} f_i(s)ds$, $i = 1, 2$, and $\widetilde{\eta}_4 = \int_0^{\tau_1} f_4(s)e^{-\mu_3 s}ds$. Let $\tau_2 = \tau_0 + \tau_1$, Then, for $t \geq \tau_2$, we have

$$\begin{aligned}
&\tfrac{dT_i}{dt} \geq \int_0^{\tau_1} \beta_1 T(t-s)V(t-s)f_1(s)e^{-\mu_1 s}ds + \int_0^{\tau_1} \beta_2 T(t-s)T_i(t-s)f_2(s)e^{-\mu_1 s}ds \\
&- \left(\mu_1 + \tfrac{q\lambda}{\alpha}\right) T_i \\
&\geq \left(\tfrac{b}{\delta} - \epsilon\right)\left[\int_0^{\tau_1} \beta_1 V(t-s)f_1(s)e^{-\mu_1 s}ds + \int_0^{\tau_1} \beta_2 T_i(t-s)f_2(s)e^{-\mu_1 s}ds\right] - \left(\mu_1 + \tfrac{q\lambda}{\alpha}\right) T_i.
\end{aligned}$$

This suggests the following comparison system for $(T_i(t), V(t))$:

$$\begin{cases}
\dot{u}_1(t) = \left(\frac{b}{\delta} - \epsilon\right)\left[\int_0^{\tau_1} \beta_1 u_2(t-s)f_1(s)e^{-\mu_1 s}ds + \int_0^{\tau_1} \beta_2 u_1(t-s)f_2(s)e^{-\mu_1 s}ds\right] \\
\qquad\quad - \left(\mu_1 + \frac{q\lambda}{\alpha}\right) u_1(t), \\
\dot{u}_2(t) = k\int_0^{\tau_1} u_1(t-s)f_4(s)e^{-\mu_3 s}ds - cu_2(t), \quad \text{for} \quad t \geq \tau_2.
\end{cases} \tag{8}$$

Notice that this is a monotone system, and by the comparison theorem and the equations $\lim_{t\to\infty} T_i(t) = 0$ and $\lim_{t\to\infty} V(t) = 0$, one should have $\lim_{t\to\infty}(u_1(t), u_2(t)) = (0, 0)$. On the other hand, the two equations for $u_1(t)$ and $u_2(t)$ are in the same forms of the second and fourth equations in system (1). Repeating the same argument for proving the instability of E_0 in Theorem 2.2 and replacing the condition $\mathcal{R}_0 > 1$ by (7), we conclude that the characteristic equation of system (8) has a positive real eigenvalue, which is a contradiction to $\lim_{t\to\infty}(u_1(t), u_2(t)) = (0, 0)$. Thus, we have $W^s(E_0)\cap\mathcal{D}^0 = \emptyset$. By Theorem 4.2 in [1], we know that there exists a value $\sigma > 0$ such that $\lim_{t\to\infty} \inf d(S(t)\phi, \mathcal{D}_0) \geq \sigma, \forall\phi \in \mathcal{D}_0$, which means that each component of the solution with the initial condition (2) satisfies

$$\lim_{t\to\infty} \inf T(t) \geq \sigma, \quad \lim_{t\to\infty} \inf T_i(t) \geq \sigma, \quad \lim_{t\to\infty} \inf T_c(t) \geq \sigma, \quad \lim_{t\to\infty} \inf V(t) \geq \sigma.$$

Arrhythmia Classification and Prediction

Electrocardiograms patterns analysis using Artificial Neural Network and non-linear regression

Abdoul Dalibou ABDOU ** — Ndeye Fatou NGOM* — Oumar NIANG *

* Laboratoire Traitement de l'Information et Systémes Intelligents (LTISI)
Département de Génie Informatique et télécommunication (GIT)
Ecole Polytechnique de Thies (EPT)
Thiès, Sénégal
fngom@ept.sn,oniang@ept.sn
** École Doctorale Développement Durable et Société (ED 2DS)
Université de Thiès
Thiès, Sénégal
abdould.abdou@univ-thies.sn

RÉSUMÉ. Les techniques d'intelligence artificielle sont trés performantes pour l'identification et l'extraction d'informations pertinentes à partir de données biomédicales. Cependant, pour les maladies du coeur, il existe toujours des difficultés à trouver des solutions basées sur l'apprentissage automatique efficaces pour l'aide à la prise de décision lors de diagnostic d'arythmies. Dans ce papier, nous proposons un système automatisé de classification et de prédiction basé sur un réseau neuronal artificiel pour l'arythmie cardiaque à l'aide d'électrocardiogrammes (ECG). Une analyse adaptative basée sur la décomposition modale empirique (EMD) est d'abord effectuée pour le débruitage du signal et la détection des principaux attributs d'un Ecg. Ces attributs sont ensuite utilisés en entrée du réseau neuronal afin de classer l'arythmie. les résultats de la classification sont combinés avec le rythme cardiaque pour effectuer une prédiction d'arrythmies bas'ee sur la régression non linéaire. Les modèles sont testés avec la base de données MIT BIH Arrhythmia et les résultats comparés à d'autres études. Une amélioration de 5.56% et 6.67% a été noté respectivement pour la classification et la prédiction.

ABSTRACT. Artificial intelligence techniques have been proven useful for the identification and the extraction of relevant information from biomedical data. However for hearth diseases, there still have difficulties in delivering efficient machine learning based on methods to be applied in arrhythmia diagnostic decision supports. In this paper we propose an automatic artificial neural network (ANN) based on classification and prediction system for cardiac arrhythmia using heartbeat recordings. An adaptive analysis based on an Empirical Mode Decomposition (EMD) is first carried out to perform signal denoising and the detection of main Ecg patterns. The ECG pattern are then used as input for an ANN to classify arrhythmia. The classification results are combined with hearth rhythms to perform non-linear regression based prediction of arrhythmia. The models are prepared and tested with the MIT-BIH database. An improvement of 5.56% and 6.67% was noted respectively for classification and prediction.

MOTS-CLÉS : Électrocardiogramme, Décomposition modole empirique, Réseau de neurones, Classification, Modèle prédictif, Fréquence cardiaque, Arythmie

KEYWORDS : Electrocardiogram, Empirical Mode Decomposition, Neural Network, Classification, Predictive model, Hearth rate, Arrhythmia

1. Introduction

Hearth disease is one of the leading cause of death around the world. Therefore, the understanding of hearth anomalies have become a main research topics in the field of cardiac care. An anomaly is an abnormality that occurs when the behavior of the system is unusual and significantly different from previous normal behavior [1]. Today with the advances of computer sciences for signal interpretation, electrocardiogram (ecg) analysis one of the most promising cardiac diagnostic approach that can provides valuable information about heart condition. The main steps involved in ecg analysis are abnormalities detection, classification and prediction. Anomalies detection methods are based on adaptive sampling [2] and ECG feature extraction using adaptive methods such as wavelet transform [3] or empirical mode decomposition [4]. Classification is usually performed with K nearest neighbor [5], super vector machine [6] and Neural Network models [4, 7, 8]. Unlike detection and classification methods approaches, predictive models gives indicators for possible abnormalities before the symptoms occur from historical data and an intelligent system. Few studies aimed arrhythmia prediction and most of them are based on linear model[9, 10]. In this paper, artificial neural network and empirical mode decomposition are first used for hearth anomalies detection and classification. Then the ANN outputs is used as input of non linear regression model as input for arrhythmia arrhythmia scheme. The main contributions are the ECGs morphological and frequency properties taken as input during the classification, and the predictive model based on non linear hearth rate frequency analysis. The proposed approach is illustrated using MIT-BIH database, compared to other studies and discussed.
The body of the paper is organized as follows. Section 2 presents the architecture of our methodology, the basics of the empirical mode decomposition and the neural network classifier. Section 3 describes the filter parameter extraction. Section 4 presents the classification method. Section 5 describes the predictive model. Section 6 shows and discusses about the results obtained with our methodology. Section 7 draws conclusions and perspectives of work.

2. Model Presentation

The step processing for the proposed approach is presented in the Figure 1. The inputs of the system are ECGs. For the MIT-BIH database, each ECG includes three components : time of samples, MLII signal and V5 one. For the classification, we first extract the V5 signal, then denoise the signal through filtering and compute input parameters (Negative form, maximum amplitude, minimum amplitude, maximum width maximum, minimum width, hearth rate and hearthbeat) for the neural network classifier. The outputs of the anomalies detector are then used during the arrhythmia classification and prediction steps. For this purpose, we first compute the 10 previous hearth rate, then we estimate the hearth rate at t (prediction horizon) and we predict the existence or not of arrhythmia.

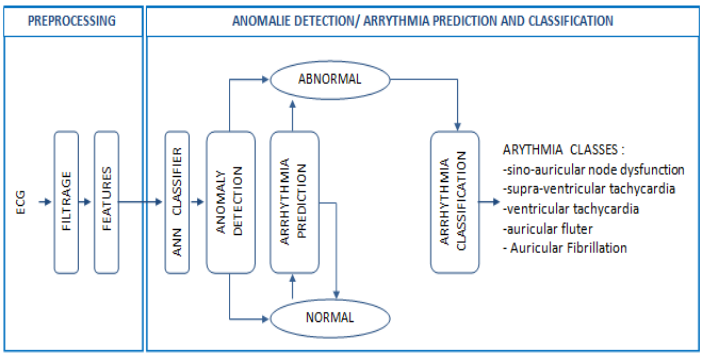


Figure 1. *Chart Flow of the proposed classification and predictive approach.*

2.1. Preprocessing

2.1.1. Empirical mode decomposition

In this study, the ecgs denoising process is done using empirical mode decomposition (EMD). EMD decomposes iteratively a complex signal s(n) into elementary components AM-FM Types, called Intrinsic Mode Functions (IMFs) [11].

$$s(n) = r_k(n) + \sum_{k=1}^{K} imf_k(n) \qquad [1]$$

Where imf_k is the k^{th} mode or IMF of the signal and r_k is the residual trend. The sifting procedure generates a finite number of IMFs. The underlying principle of the EMD is to identify locally in the signal, the fastest oscillations defined as the waveform interpolating the local maxima and minima. To do this, these last points are interpolated with a cubic spline to produce the upper and lower envelopes. The average envelope is subtracted from the initial signal and the same interpolation scheme is reiterated.

2.1.2. Filtering

Usually a real ECG signal faces muscular noise, motion artifacts, and baseline drifts changes. Butterworth filter is often used for noise smoothing[12]. Removing the first IMFs, after EMD decomposition, filter out ecg noise while preserving QRS content [13, 14]. In this work, we have combined the advantages of EMD filtering (one run approach for low and high frequency noise) and 6th order band-pass Butterworth filter (noise smoothing). We subtract the first IMF (IMF_1) to remove the high frequency and apply the Butterworth filter to smooth the signal.

2.2. Neural Network

A neural network is a mathematical function, see the picture 2[4].In this paper, we propose a neural network (figure 3) composed of sixteen nodes. To set up a neural network, there must be defined the input data, the activation function and the thresholds of the nodes. Each data is associated with a weight.

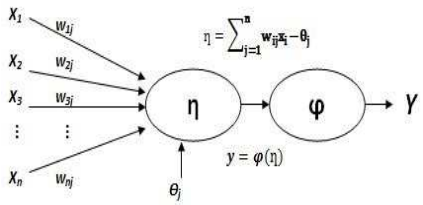


Figure 2. *Representation of an artificial neuron [4]. Inputs are multiplied by their weight. The products are added to give the weighted sum. The threshold of the node is subtracted from the weighted sum to determine the output of the node.*

3. ECG patterns detection

Intrinsic parameters are used for ECG patterns detection that can lead the classification and prediction that will be further performed. Indeed, statistical properties(mean, variance, standard deviation, energy and power) are often used as input parameters for classification [4]. However these parameters are global descriptors of data. Unlike statistical properties, morphological and frequency attributes allow local analysis. Thus, in this work, we used a set of morphological and frequency properties (negative form, maximum amplitude, minimum amplitude, maximum width, minimum width, hearth rate, and hearth rhythm) of the ECG and the Heaviside activation function for classifying and predicting the arrhythmia.

3.1. Parameters Vector

The parameters vector is used as the neural network entry. It is composed of an electrocardiogram's properties.The parameters processing are performing as follows :

1) Input : ECG
2) detect the negative form
3) compute QRS complex width and amplitude
4) apply min and max function on width and amplitude

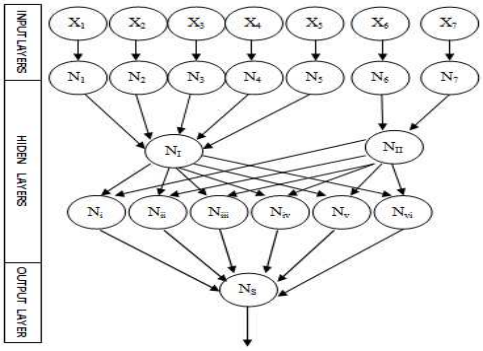


Figure 3. *Architecture of the neural network. The variables X_1 ... X_7 represent the morphological and frequency properties, respectively the negative form, the maximum amplitude, the minimum amplitude, the maximum width, the minimum width, the hearth rate and the hearthbeat. The variables N_1 .. N_{vi} represent the intermediate nodes. The Node N_s determines the arrhythmia type.*

5) compute hearth rate and hearth rhythm

6) Output : return a parameters vector

Detailed description of the QRS complex detection, width and amplitude computations were presented in [4]. The hearth rhythm regularity detection requires a certain number of steps :

1) identify the R-waves

2) count the R-R intervals

3) compute the R-R regularity rate

4. Arrhythmia Classification

The classification involves two functions : the network function and the classifier function. The implemented neural network uses H(x),the step activation function with a threshold (s) for each parameter : $H(x) = 0\ if\ x < s\ and\ H(x) = 1\ if\ x \geq s$. The parameters are negative form, minimum width, maximum width, minimum amplitude, maximum amplitude, hearth rate and hearth rhythm which are respectively associated to the following thresholds 0, 0.06, 0.10, 0.5, 2.5, 50, 110 and 10 [10]. The network function is composed of sixteen neurons. It takes as input a parameter vector and returns one of these six classes : Class 0 : normal,Class 1 : sino-auricular node dysfunction,Class 2 : supra-ventricular tachycardia,Class 3 : ventricular tachycardia,Class 4 : auricular flutter and Class 5 : Auricular Fibrillation.

5. Arrythmias Prediction

From our knowledge, the few works that have been done for hearth rhythm prediction are based on linear model. However real world phenomena are not linear. This work is one of the first step for the development of an efficient model for ECG predictive analysis. The proposed approach is based on the exponential non-linear regression model. This estimation is done through the following equation

$$y = b_1 + b_2 * x^{b_3} \qquad [2]$$

where, y is the estimated frequency, x the prediction horizon and $b_{1,2,3}$ are the model coefficients. The frequency prediction is done such as :

1) computation of the 10 previous hearth rates,

2) estimation of the hearth rate,

3) classification of the estimated cardiac frequency,

4) prediction of an arrhythmia.

The prediction base is constructed by extracting the samples from last 10 minutes and computing the corresponding hearth rates. The cardiac frequency is estimated using the algorithm 1.

Tableau 1. *ECG classified*

CLASS	ECG
Sino-auricular node dysfonction	103, 105,107, 108, 114,115,119,121,123,201,202
	208,210,213,220,222,223,228,231,233,234
Supra-ventricular tachycardia	102,111,215
Ventricular tachycardia	101
Auricular fluter	204,209
Auricular fibrillation	113, 118, 217

Algorithm 1 Prediction Method

1: function predire(f,t,th)
2: $modelfun \leftarrow @(b,x)b(1)+b(2)*x(:,1).^{b}(3)$
3: $beta0 \leftarrow [111]$
4: $mdl \leftarrow Estimate(t,f,modelfun,beta0)$
5: $fp \leftarrow predict(mdl,th)$
6: return fp

The function (algorithm 1) takes ten previous frequencies(f), time samples (t) and th (horizon prediction). It estimates the hearth rate at th (prediction horizon).

6. Results and discussion

To illustrate the classification and prediction of arrhythmias, we have used the MIT-BIH ECG database [4, 15]. MIT-BIH Arrhythmia is a waveform and a class completed references databases of physionet.org composed of 48 signals recorded on a half-hour that can be downloaded from physionet.org.

6.1. Classification

We first use dual filtering based on the EMD and the Butterworth filter during the pre-processing step. Then, we compute the parameters for anomalies detection and arrhythmia classification, and estimate the hearth rate for the arrhythmia prediction. We have the following results : 8 Abnormal ECG was detected as abnormal and 1 abnormal ECG was detected as normal(detection error) ; 4 normal ECG were detected as abnormal (detection error) and 2 ECG normal ECG were detected as normal.The detection method achieved a performance rate of 88.89% with an error rate of 11.11%. The performance indices are : Accuracy (66.67%), Sensitivity (66.67%), Specificity (88.89 %) and Positive predictive (33.34%). The performance of anomalies detection is represented by the specificity. After the anomalies detection, if an anomaly is detected then the system continues with the arrhythmia classification else it continues with arrhythmia prediction. The classification result is presented in the table 1.

Compared to the results presented in [4], there is an improvement in performance of 5.56% (83.33 % to 88.89%). The same trend is noted for the back propagation method (83.40%) proposed in [16] and ANN based classification (87.50%) proposed in [17]. The

noted improvement is due to the use of the morphological and frequency parameters. Besides, these parameters allow local analysis with defined thresholds.

6.2. Prediction

If an anomaly isn't detected then the system continues with the proposed non linear regression based arrhythmia prediction (algorithm 1). An estimated hearth rate at prediction horizon $t = 11$ is used to predict an arrhythmia. To validate our model, the data were tested with the linear model and compared with the proposed non linear model. The image 4 shows the results of the prediction with the two models in relation to the calculated heart rates.

The linear prediction predicts correctly twelve EGC and the non-linear prediction pre-

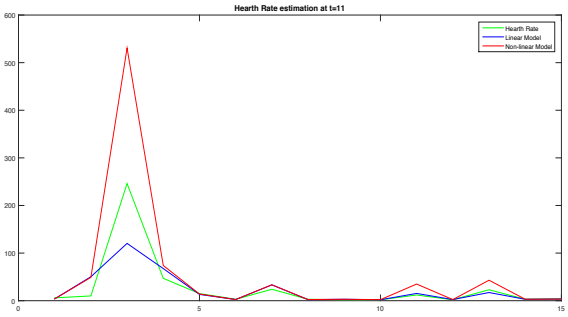


Figure 4. *Hearth Rate estimation*

dicts correctly thirteen ECG. The performance rate of the linear and non-linear model are respectively 80.00% and 86.67%. Thus there is an improvement in the prediction accuracy using the proposed model.

7. Conclusion

In this work, we proposed an approach based on empirical mode decomposition(EMD), the neural network, and non-linear regression for classification and prediction of arrhythmia. The main contributions are the ECGs morphological and frequency properties taken as input during the classification and the predictive model based on hearth rate analysis. The output of our approach gives promising results for the classification and prediction of arrhythmia. Future works will focus on modeling the neural network with a filter bank and the implementation of a secure online system for classification and prediction that can be used by practitioners as help for decision support.

8. Bibliographie

[1] SOTT PURDY, ZUHA AGHA, SABUTAI AHMAD, ALESADER LAVIN, « Unsupervised real-time anomaly detection for streaming data », *Neurocomputing*,262 :134-147, 2017.

[2] L. MESIN, « Heartbeat monitoring from adaptively down samples electrocardiogram », *Computers in Biology and Medecine*,84 :217-225, 2017.

[3] S.SAHOO, B. KANUNGO, S. BEHERA, S.SABUT, « Multiresolution wavelet transform based feature extraction and ECG classification to detect cardiac abnormalities », *Measurment*,108 :55-66, 2017.

[4] A-DALIBOU ABDOU, NDEYE FATOU NGOM, SAMBA SIDIBE, OUMAR NIANG, ABDOULAYE THIOUNE, CHEIKH H.T.C NDIAYE, « Neural Networks for biomedical signals classification based on empirical mode decomposition and principal component analysis », *Lecture Notes of the Institute for Computer Sciences, Social Informaticq and Telecommunications Engineering (LNICST). ISSN 1867-8211. pp. 267-278*, 2017.

[5] V.GUPTA, M. MITTAL, « KNN and PCA classifier with autoregressive modeling during different ECG signal interpretation », *Proc. Comp. Scienc*,125 :18-24, 2018.

[6] K.N. RAJESH, R. DHULI, « Classification of ECG hearthbeats using nonlinear decomposition methods and support vector machine », *Computers in biology and medecine*,87 :271-284, 2017.

[7] K.M ATOUF, A. ISSAM, A. MOHAMED, B. ABDELLATIF, « ECG image classification in real time based on harr like features and artificial neural networks », *Procedia Computer Science* ,73 :32-39, 2015.

[8] J.A. GUTIERREZ-GNECCHI, R. MORFIN-MAGANA , D. LORIAZ-ESPINOZ , A.C. TELLEZ-ANGUIANO, E. REYES-ARCHUNDIA, A. MENDEZ-PATINO, R. CASTANEDA-MIRANDA, « DSP based arrythmia classification using wavelet transform and probabilistic neural network », *Biomedical Signal Processing and Control*,32 :44-56, 2017.

[9] M.L. TALBI, P. RAVIER, « Detection of PVC in ECG signal using fractional linear prediction », *Biomedical Signal Processing and Control*,23 :42-51, 2016.

[10] R. DUBOIS, « Application des nouvelles méthodes d'apprentissage pour la détection précoce d'anomalies cardiaques en électrocardiographie », *PHD Université Pierre et Marie Curie - Paris VI*, 2004.

[11] S. PAL, M. MITRA, « Empirical mode decomposition based on ECG enhencement and QRS detection », *Computers in biology and medecine*,12 :83-92, 2012.

[12] H. ZHANG, « An improved QRS wave group detection algorithm and matlab implementation », *Physics procedia*,25 :1010-1016, vol. n° , 2012.

[13] R. RODRIGUEZ, A. MEXICANO, J. BILA, S. CERVANTES, R. PONCE, « features extraction of electrocardiogram signals by applying adaptative threshold and principal component analysis », *Journal of applyied research and technology*,13 :261-269, 2015.

[14] Z.H. SLIMANE, A. NAIT ALI, « QRS Complex detection using empirical mode decomposition », *Digital signal processing*, 20 :1221-1228, 2010.

[15] GB MOODY, RG MARK, « The impact of mit-bih arrhythmia database », *IEE Eng in Med and Biol*, 20 :45-50, 2001.

[16] J.FENG, « ECG Classification based on feature-extraction and neural network », *Phd, Sichuan Normal University*, 2005.

[17] C.ALEXAKIS, H.O NYONGESA, R. SAATCHI, N.D. HARRIS, C. DAVIES, C. EMERY, R.H IRELAND, S.R HELLER, « Feature extraction and classification of electrocardiogram (ECG)signals related to hypoglycaemia », *Proc. Comput. Cardiol.*,30 :537-540, 2003.

A spatio-temporal model for phenomena dynamics based on 2D Diffusion equations

Samuel .I Billong IV
Département d'informatique
ENS Polytechnique
Université de Yaoundé I
B.P : 8390 Yaoundé
Cameroun
Samuelismael4@gmail.com

G.E Kouamou
Département d'informatique
ENS Polytechnique
Université de Yaoundé I
B.P: 8390 Yaoundé
Cameroun
Georges_edouard@yahoo.com

T. Bouetou
Département d'informatique
ENS Polytechnique
Université de Yaoundé I
B.P : 8390 Yaoundé
Cameroun
tbouetou@gmail.com

ABSTRACT. This study highlights the relevance of the dynamics of population in an environment in the prediction of phenomena. A hybrid model compatible with a 2D diffusion equation is proposed. It is based on the balanced method coupled with the models of the dynamics of the populations. The resulting equations, since they are a kind of conservation equations, are discretized using the finite volume method. This equation is strongly linked to a probabilistic diffusion coefficient which highlights the random moving of mobile entities. It also represents the influence of the neighbors of each site in the dynamics of mobile entities, within a closed environment. This approach is illustrated on the well-known SIR epidemiological model to produce a variant which consider the spatio-temporal aspect of the spread.

KEYWORDS: Population Dynamics, Prediction, Diffusion, Finite Volume, Random, Closed Environment

1. Introduction

The observation of the phenomena is a major research activity which deals with the prediction based on models and theories sometimes existing or created. Today, with the development of new technologies, migrations, wars, the mixing of populations, terrorism etc. The prediction becomes decision support tool. Given the complexity of this field, there are several approaches due to the random behavior of mobile entities in one hand and geographic constraints on the other. It would be therefore possible to circumvent these difficulties by considering for each phenomenon a theory or a model that adapts to it as well as possible.

In this study, we work in a closed environment divided into several distinct sites. In this environment the mobile entities move from one site to another in a random manner, thus changing the size of the population.

The most researches often put emphasis on the phenomenological meaning around a theme. The contrast between ideal and reality, the expected and lived brings us to think alike in prediction domain by studying a phenomenon in all its outlines with all its characteristics and make that the rendering of its study is reliable, its modes of emergence or manifestation can be elaborated in the form of mathematical equation. The conclusion is clear, most prediction models take much more into account the dynamics of the phenomenon to be studied without explicitly highlighting the impact of the spatio-temporal dynamics of the entities concerned by the prediction model. Hence we have got the following ideas

- Study the behavior of a phenomenon dynamically as a function of time and space in its evolution

- Take a phenomenon into an atomic division according to the spatial-temporal behaviors and to superpose afterwards to find the global behavior

- Find for each phenomenon a spatio-temporal behavior which adapts to an existing model or to propose a model

- Dissociate the dynamics of the individuals to that of a phenomenon then to associate the different results to predict a future behavior

- To write probabilistic time equations in order to derive at a model that can integrate the random aspect of a phenomenon and its spatio-temporal evolution.

Some authors (Jianhong Wu and P. van den Driessche [4]) thought in this way. Without pretending to control the manifestation of all the contours related to a phenomenon in the case of prediction, our contribution will be much more oriented towards the impact of the spatio-temporal dynamics of the mobile entities in the prediction models. [5] [4].

In the remainder of this paper the section 2 present the related works which inspired our point of view, a particular emphasis will be put on those that allowed us to build our model. The section 3 describes the approach used to build the equations underline the proposed model. The section 4 show an illustration on the extension of the SIR epidemiological model then the paper end with conclusions and perspectives.

2. Literature review.

2.1. Particle diffusion

The idea underlying this study consist to extend the diffusion equation applied to the analysis and the prediction of phenomenon. Starting from the mobile particle balanced method based on the diffusion of particles [10], the following 1D diffusion is build

$$\frac{\partial n(x,t)}{\partial t} = -\frac{\partial j_x(x,t)}{\partial x} \quad (1)$$

Where $\vec{j}(x,t)$ is the density of moving entities and n(x,t) the number of mobile entities per volume unit. For 2D and more the equation (1) takes the following form

$$\frac{\partial n(M,t)}{\partial t} + div\vec{j} = 0 \quad (2)$$

It is proved that these equations are valid in any geometry the divergence operator can change depending on the coordinate system adopted.

2.2. Population dynamics: parabolic partial differential equation

Jimmy Garnier in his thesis [5] stutied this family of equation

$$\frac{\partial u(t,x)}{\partial t} = D(u)(t,x) + f\big(x, u(t,x)\big) \qquad t > 0\,,\ x \in R \quad (3)$$

These equations are useful in many fields such as combustion chemistry, biology or ecology. These equations are generally used to modelize the evolution of entities that interact each other and moved. f population dynamics or population genetics[7]. The quantity u (t, x) represents the population density at time t and at position x. The reaction term f (x, u) corresponds to the growth rate of the population. This term of reaction depends on the one hand on the density u and on the other hand with the medium in which the population evolves through the variable of space x.

In this large set, we will focus mainly on a single type of reaction-dispersion equation where the dispersion operator D is a second-order elliptic differential operator. [5]

$$\frac{\partial u(t,x)}{\partial t} = \frac{\partial^2 u(t,x)}{\partial x^2} + f\big(x, u(t,x)\big) \quad (4)$$

A particular attention is focused on this equation because it looks like the diffusion equation obtained from the fick law which state that the flow due to the random movement is approximately proportional to the gradient in the number of individuals.

In the same way Jianhong Wu [6] starting from the basic concepts to developed a model for the spatial spread of diseases involving hosts in random displacement during certain stages of the progression of the disease. He got a diffusion model based on the conservation laws and Fick's law. This model was applied to the study of two cases, namely the spread of rabies in continental Europe during the period 1945-1985 and the rates of spread of West Nile virus in North America.

The same approach is used in 2D closed environment to avoid solving equation (1) which has two unknowns. To underline the randomness aspect in displacement, the probabilistic diffusion coefficient will be built as Wu [6]. This coefficient will also take into account the dimensional aspect of the quantities used

3. Methodology

3.1. Basis of the model

Starting from the diffusion of the particles in 2 dimension assuming that the particles move along x and y coordinates, we obtain the fallowing 2D conservation equation

$$\frac{\partial n(M,t)}{\partial t} + div\vec{j} = 0 \quad (2)$$

In the following it is necessary:

- Find an explicit form of equations (1) and (2) as a function of time and space
- Find the equivalence of the operator div (j (M, t)) as a function of density of mobile entities n (M, t)

Fick's law mentioned in [4] states that the flow due to the random movement is approximately proportional to the gradient in the number of individuals like

$$j = -D\frac{\partial n(t,x)}{\partial x}$$

To have a diffusion equation of the form

$$\frac{\partial n(x,t)}{\partial t} = D\Delta n(x,t) \quad (4)$$ Where D is the diffusion coefficient and Δ is the Laplacian operator.

We have exploited the parabolic equation below ta ken from [5]

$\frac{\partial u(t,x)}{\partial t} = D(u)(t,x) + f\big(x, u(t,x)\big)$ (3) $\quad t > 0\ ,\ x \in R$. In the difusion form as:

$\frac{\partial u(t,x)}{\partial t} = \frac{\partial^2 u(t,x)}{\partial x^2} + f\big(x, u(t,x)\big)$ (4)

According to our analysis, it combines particle scattering and Fick's law, notwithstanding a residual coefficient D (diffusion coefficient). This went well to the form of equation sought. To continue in our positioning we made approximations on equation (4) as follows:

- We first neglect the creation factor f (x, u (t, x)) taking into account the time difference between t and t + dt which will not be at the scale of a duration that can hold significantly account for the death or birth of a new mobile entity.
- Add a coefficient D in front of the second-order elliptic differential operator to take into account the randomness of the displacement of moving entities from one site to another on the one hand and the homogeneity of the dimensional equation on the other hand.
- The coefficient D can be constant or follow a law of variation according to the complexity related to the motions inter - site of the mobile entities.

In case of 2 dimension we got the following equation

$\frac{\partial u(t,x)}{\partial t} = D(\frac{\partial^2 u(t,x)}{\partial x^2} + \frac{\partial^2 u(t,x)}{\partial y^2})$ (5)

Thus the whole difficulty of this modeling will reside on our capacity to give a form adapted to the diffusion coefficient D. Seen in this angle the diffusion coefficient D will make our model adaptive.

3.2. Model development

For a first approach we will consider our closed environment as a homogeneous site distribution

The complexity of our approach takes us to a discrete solution seen the impossibility to have an analytical solution because of the randomness of the displacement of the moving entities. For that we chose the method of the finished volumes [9].

For the diffusion coefficient D we modeled it as a transition matrix that materializes the contribution of a site x to a site y during the diffusion. So we have made assumptions that lead us to formulate the coefficient D in the following ways

- A mobile entity in a site can decide to move or not
- the probability of moving from one site to another depends on the number of neighbor’s sites

- $D_{x,y} = \omega P_{x,y}\ with\ \omega \approx \frac{\Delta x^2}{\Delta t}$

- ω depend on the configuration of the problem (as a function of the characteristic speed of the movement of the mobile entities).

- $P_{x,y}$ the probability of leaving a site x for a site y

After applying the finite volume method we obtain the following numerical scheme

$$u_{i,j,k}^{n+1} = \frac{\Delta t D_{l,k}}{\Delta x^2}\left(u_{i+1,j,x}^{n}\delta_{l,x} + u_{i-1,j,y}^{n}\delta_{l,y}\right) + \frac{\Delta t D_{l,k}}{\Delta y^2}\left(u_{i,j+1,z}^{n}\delta_{l,z} + u_{i,j-1,t}^{n}\delta_{l,t}\right) + \left(1 - 2\left(\frac{\Delta t D_{l,k}}{\Delta x^2} + \frac{\Delta t D_{l,k}}{\Delta y^2}\right)\right)u_{i,j,k}^{n}\delta_{l,k} \quad (6)$$

Where $\delta_{i,j}$ are the kronecker's symbols and $u_{i,j,k}^{n}$ the mobile entity density

It is the general formulation of the model. We take into account the boundary conditions according to the considered space geometry. In this formula n represents the temporal discretization index i and j the spatial discretization indices along the x axis and the y axis. k represents the number assigned to a site used as an index in the diffusion matrix

4. Illustration

This section, is the ways of presenting our ideas of the taking into account the spatial-temporal factors to bring a corrective term to a familiar model of prediction and to change its original form. Suppose a homogeneous population and each individual of the population can be identified by its position in a site within our enclosed environment. Take the case of a disease that acts on mobile entities and whose dynamics are modeled by the epidemiological model SIR (Susceptible, Infective, and Recovered) and combined with our probabilistic diffusion modeling approach. Given our assumptions, our modeling should lead us to ordinary differential equations at first if we do not take into account the inter-site migration described above in our modeling. Then we will have another system of differential equations that highlights the impact of the inter-site migration. That is the goal of this example modify the equations of an ordinary system by taking into account the inter-site migration in the equations of the model. In the following, we assume that it is the same disease that occurs and spreads across all the different sites involved in trade For this purpose, let α, β and γ be respectively the rates of infection, cure and return to the susceptible condition of the individuals of a site, and assuming that we have P sites, we obtain in a first time a set of 3P ordinary differential equations describing the dynamics of infection within the population of a site i. Then, in a second step, a set of 3P

complex differential equations describing the dynamics of the infection combined with the inter-site dynamic of the populations

- Characteristic equation system for the 3P differential equations without taking into account the dynamics between sites

$$(7)\begin{cases}\frac{dS(t)}{dt} = \gamma R - \alpha S \\ \frac{dI(t)}{dt} = \alpha S - \beta I \\ \frac{dR(t)}{dt} = \beta I - \gamma R\end{cases}$$

- Characteristic equation system for 3P differential equations taking into account inter-site dynamics

$$(8)\begin{cases}\frac{\partial S(t)}{\partial t} = \gamma R - \alpha S + D\Delta S \\ \frac{\partial I(t)}{\partial t} = \alpha S - \beta I + D\Delta I \\ \frac{\partial R(t)}{\partial t} = \beta I - \gamma R + D\Delta R\end{cases}$$

After having this new equation family we discretize with the finite volumes method according to the pre-established model. We make a small simulation with a dataset using a code written in python language to have a result that shows the modification brought by the spatio-temporal consideration. The simulation takes place in a closed environment with 4 sites.

The results are presented in the annexes in the form of a histogram showing the differences between the results based on the diffusive SIR model and those of the native SIR model.

These results show us that when the diffusion rate is greater than the contamination coefficients in the natural SIR model, the rate of contamination decreases because the entities disappear from a site even before the disease has time to spread.

In conclusion we can say that it will be possible to take into account the spatio-temporal aspect in the prediction of the phenomena as a function of the diffusion speed because the more important it is the more it influences the dynamic of the studied phenomenon.

5. Conclusion

The purpose of this article was to provide a corrective factor in the prediction models, by considering the spatio- temporal impact within the dynamics of a prediction model. For this purpose a particular attention was focused through several models of

prediction dealing with mobile entities. Upon certain clearly defined hypothesis, we designed a hybrid model based on diffusion equations. This approach was illustrated by modifying a naturel SIR model to have another hybrid equation system. The validation is done through a simulation with a dataset. The results show that the aspect of spatio-temporal dynamics modifies the behavior of the native SIR model.

The further work intend to bring the model closer to the reality. In this regard the following ideas will be develop

- The discretization will be done in irregular mesh
- The formalization of the method of building the probality of moving
- propose the intervals of diffusion speed depending on the nature of the problem addressed

6 Bibliographie

[1] Allard, A. (2008). *Modélisation Mathématique en épidémiologie par les réseaux de contacts : Introduction de l'Hétérogénéité dans la Transmissibilité*, M.Sc. Theses. Université de Laval, Quebec

[2]. BASILEU, C. (2011*). Modélisation structurelle des réseaux sociaux : Application à un système d'aide à la décision en cas de crise sanitaire*. PhD Thesis, Université Lyon 1

[3]. Djamila Moulay. *Modélisation et analyse mathématique de systèmes dynamiques en épidémiologie. Application au cas du Chikungunya. Mathématiques* [math]. PhD Thesis Université du Havre, 2011. Français. <tel-00633827 >

[4]. Driessche, P. v. (2008). *Spatial Structure: Patch Models*. Springer Berlin Heidelberg, Lecture Notes in Mathematics Volume 1945, 2008, pp 179189

[5]. Jimmy Garnier. *Analyse mathématique de modèles de dynamique des populations : équations aux dérivées partielles paraboliques et équations intégro-différentielles*. Equations aux dérivées partielles [math.AP]. A ix-Marseille Université, 2012. Français. <tel-00755296 >

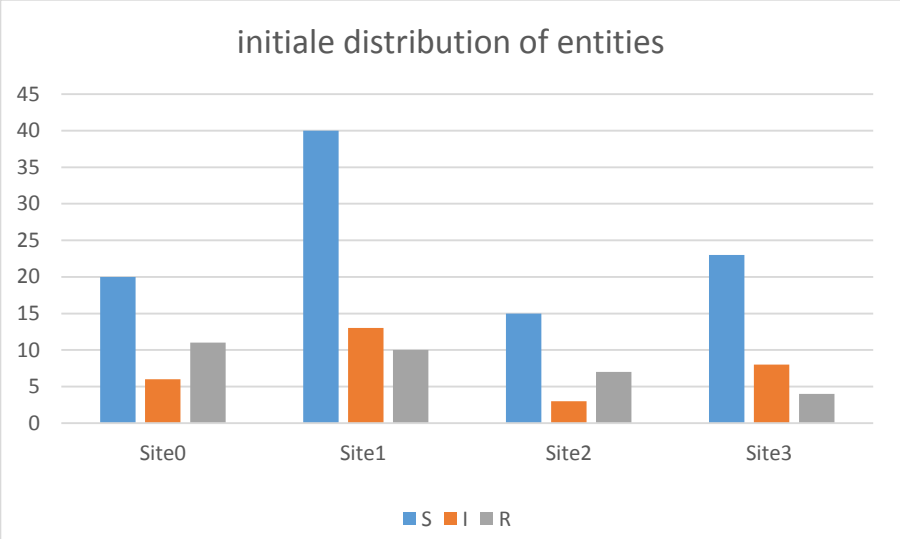
[6] Julien Arino. (2008). *Diseases in metapopulations. Centre for Disease Modelling*, Preprint 200804, University of York.

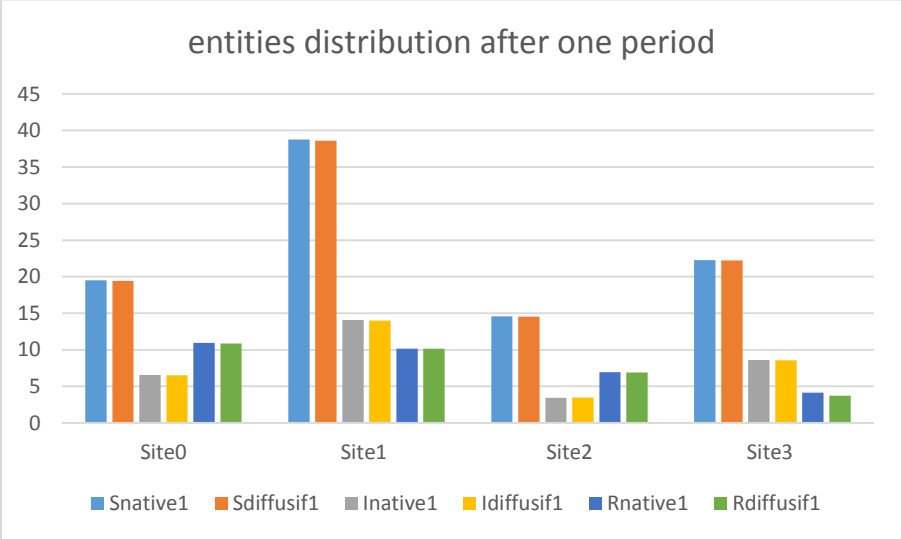
[7] N S Kolmogorov, N Petrovsky et I G Piskunov : *Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique*. Bull. Univ. Moscou, Sér A, 1:1–26, 1937.

[9] Robert Eymard, Thierry Gallouet and Raphaele Herbin *Finite Volume Methods.* January 2003 This manuscript is an update of the preprint n° 97-19 du LATP, UMR 6632, Marseille, September 1997whichappeared in Handbook of Numerical Analysis, P.G. Ciarlet, J.L. Lions eds, vol 7, pp 713-1020

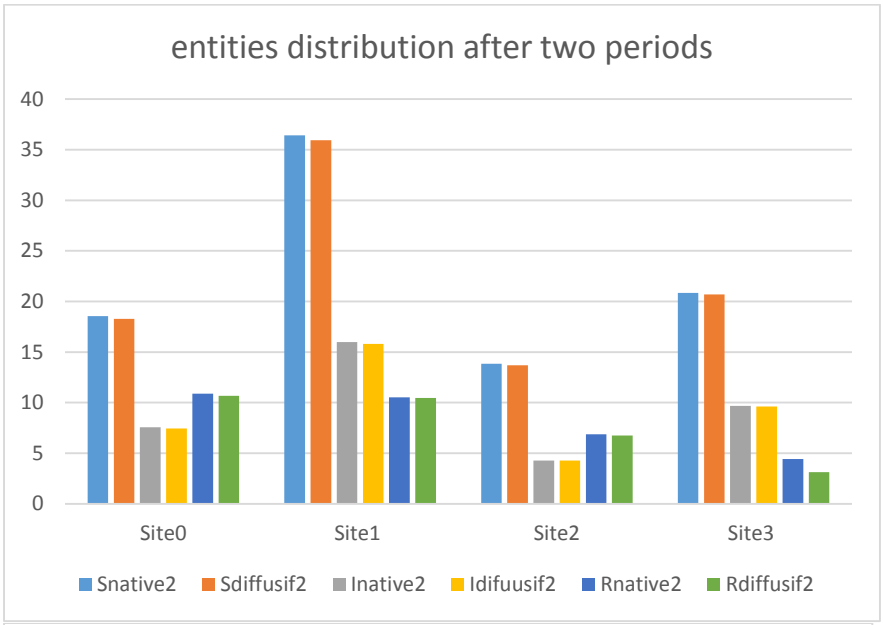
[10] Salamito, BernardCardini, Stéphane Jurine, Damien *physique tout en un* Editeur: Dunod Publication: 2013 pages: 1124 ISBN: 978-2-10-060077-9

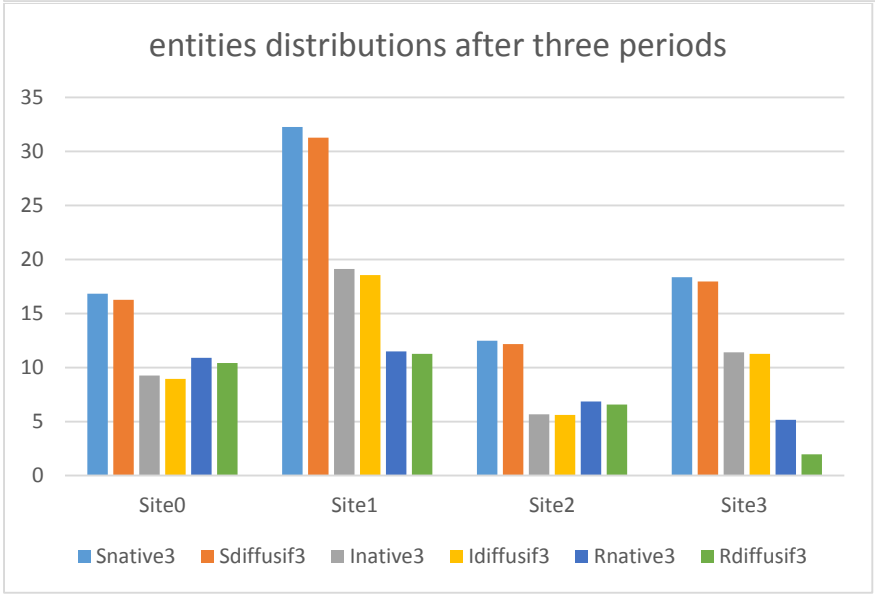
Annex





Snative = S in native SIR, **Sdiffusif** = S with diffusif model of SIR. Then the same thing is made with I and R to compare data in the same histogram


entities distribution after two periods
40
35
30
25
20
15
10
5
0
Site0
Site1
Site2
Site3
Snative2
Sdiffusif2
Inative2
Idifuusif2
Rnative2
Rdiffusif2


entities distributions after three periods
35
30
25
20
15
10
5
0
Site0
Site1
Site2
Site3
Snative3
Sdiffusif3
Inative3
Idiffusif3
Rnative3
Rdiffusif3

Validation of a Lagrangian model using trajectories of oceanographic drifters

Amemou Hilaire[a,b,d], Koné Vamara[b], Verley Philippe[c], Lett Christophe[d]

[a] Laboratoire de Physique de l'Atmosphère et Mécanique des fluides, LAPA-MF, Université Félix Houphouet Boigny, Côte d'Ivoire.

[b] Centre de Recherches Océanologiques, CRO, 29 Rue des Pêcheurs, BPV 18, Abidjan, Côte d'Ivoire.

[c] AMAP, IRD, CIRAD, CNRS, INRA, Université de Montpellier, Montpellier, France.

[d] Sorbonne Université, IRD, Unité de Modélisation Mathématique et Informatique des Systèmes Complexes, UMMISCO, F-93143, Bondy, France.

ABSTRACT. We compared trajectories of oceanographic drifters and simulations using the Lagrangian model Ichthyop. The drifters data covered the tropical Atlantic Ocean over the period 1992-2008. The model was forced firstly by interannual outputs of the Regional Ocean Modeling System (ROMS) model, and then by the Ocean Surface Current Analysis Real-time (OSCAR) remote-sensing product. We found that the relative error between data and simulations increases with time approximately linearly before leveling out. The results were close for ROMS-Ichthyop and OSCAR-Ichthyop simulations, in accordance with the spatial resolutions of the forcing products that were close ($\frac{1}{3}^\circ$ for OSCAR and $\frac{1}{5}^\circ$ for ROMS). Simulated particles traveled generally significantly lower distances than observed drifters, likely because these spatial resolutions were insufficient to resolve the meso-scale oceanic processes.

RÉSUMÉ. Nous avons comparé les trajectoires des drifters océanographiques avec celles des simulations en utilisant le modèle Lagrangien Ichthyop. Les données des drifters couvraient l'océan Atlantique tropical sur la période 1992-2008. Le modèle Ichthyop a été forcé par les sorties interannuelles du modèle Regional Ocean Modeling System (ROMS) dans un premier temps, et ensuite, par les sorties du produit Ocean Surface Current Analysis Real-time (OSCAR). Nous avons trouvé que l'erreur relative entre les données et les simulations augmentent approximativement linéairement avec le temps avant de se stabiliser. Des résultats proches ont été trouvés entre les simulations ROMS-Ichthyop et OSCAR-Ichthyop, en accord avec les résolutions spatiales des produits de forçage qui sont proches ($\frac{1}{3}^\circ$ pour OSCAR et $\frac{1}{5}^\circ$ pour ROMS). Les particules simulées ont généralement parcouru des distances significativement plus faibles que les drifters observés, probablement parce que ces résolutions spatiales étaient insuffisantes pour résoudre les processus de méso-échelle océaniques.

KEYWORDS : trajectory, drifter, currents, Lagrangian model, tropical Atlantic.

MOTS-CLÉS : trajectoire, drifter, courants, modèle Lagrangian, Atlantique tropical.

1. Introduction

The tropical Atlantic ocean is located between $65°$W and $15°$E and $10°$S and $14°$N. It is mainly dominated by the presence of strong currents such as the North Equatorial Counter Current (NECC) and the South Equatorial Current (SEC), and undercurrents such as the Equatorial Under Current (EUC). Model-simulated trajectories obtained from hydrodynamic models are increasingly used to simulate the ocean circulation (Lumpkin et al., 2002) and fish larval trajectories (Koné et al., 2017). Some models use an Eulerian approach and others a Lagrangian approach. The main advantage of the Lagrangian view is the knowledge of the "water particle history" from origin to destination. Useful applications are found in the study of marine pollution (plastic debris) and marine ecology (Lett et al., 2007). As these studies become more frequent, the need to evaluate simulated trajectories increases. The objective of this work is precisely to validate the trajectories simulated by the Lagrangian model Ichthyop using NOAA oceanographic drifters. The Ichthyop model is firstly forced by interannual outputs of the Regional Ocean Modeling System (ROMS) model over the period 1992-2008, and then by the Ocean Surface Current Analysis Real-time (OSCAR) remote-sensing product. The purpose is to make a comparative analysis of the solutions ROMS-Ichthyop and OSCAR-Ichthyop coupled models using the trajectories of in situ drifters as reference.

2. Material and methods

2.1. Oceanographic drifters

The trajectories of near-surface drifters are gathered by the National Oceanic and Atmospheric Administration (NOAA). Drifters are drogued at 15 m depth and are tracked by the Argos satellites. Their positions are given every six hours (Hansen and Poulain, 1996). The data used in this study comprise 278 drifters.

2.2. OSCAR currents

The OSCAR product provides near real-time ocean surface velocities from different satellites fields (TOPEX / Poseidon (1992- 2002) and Jason (2002-present) (Bonjean and Lagerloef, 2002). OSCAR is a product distributed by NASA's Physical Oceanography Data Center (`http://podaac.jpl.nasa.gov`). Velocities are calculated from quasi-linear equations of motion by combining geostrophy, Ekmann and Stommel formulations and a complementary term to the surface flotation gradient (Bonjean and Lagerloef, 2002). Horizontal velocities are directly estimated from the height of the sea surface, sea surface velocity, surface wind speed, and sea surface temperature. The OSCAR product is on a $\frac{1}{3}^{\circ}$ grid resolution with a 5-day interval.

2.3. ROMS hydrodynamic model

The ROMS hydrodynamic model is three-dimensional, split-explicit based on the hydrostatic balance, incompressibility and Boussinesq hypotheses for solving, primitive equations and also the free surface (Shchepetkin and McWilliams, 2005). The model configuration used here was built over the tropical Atlantic at a horizontal resolution of $\frac{1}{5}^{\circ}$ ($\simeq$ 22 km). The 45 vertical levels of the grid are discretized according to a sigma coordinate system to increase the vertical resolution near the surface. On the surface, the model

is forced with interannual winds derived from atmospheric forcings CFSR (Climate Forecast System Reanalysis). The superficial heat and fresh water fluxes introduced into the model come from the model SODA (Simple Ocean Data Assimilation) version 2. The model has three open borders (North, South and West) and a closed border (East) forced by the outputs of SODA. The outputs of the model are saved every 2 days. Model details and implementations are documented by Djakouré et al. (2014) and Koné et al. (2017).

2.4. Ichthyop Lagrangian model

Ichthyop (Lett et al., 2008) is a tool that simulates Lagrangian particle transport using current fields produced by hydrodynamic models such as ROMS, for applications in physical oceanography and / or in marine ecology. The main applications are the study of the transport processes and their effects on the variability of fish ichthyoplankton recruitment. In the present application, cloud of discrete particles (1000) without mass are released at the observed drifters location. The displacement of each particle is given by the sum of an advective component and a horizontal dispersive component (Peliz et al., 2007). The positions of particles are computed at each time step using the Runge-Kutta 4 forward scheme and saved every six hours (corresponding to the recorded periods of the drifters positions).

2.5. Statistical analysis

The mean position of the particles simulated at each time step is defined as follows:

$$\overline{x}^p(t) = \frac{1}{N}\sum_{i=1}^{N} x_i^p(t) \quad [1]$$

where t is the time, N is the total number of simulated particles and p is the index of the spatial coordinate (longitude, latitude and depth). Thus, we obtain the barycenter of the simulated particles. The standard deviation is given by

$$\sigma^p(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i^p(t) - \overline{x}^p(t))^2} \quad [2]$$

The relative error measures the separation of two particles or, equivalently, the propagation of a cloud of passive tracers. This index allows to obtain the evolution of the distance between the observation (drifter) and the simulation (barycenter) as a function of time:

$$D(t) = \sqrt{(x_b(t) - x_o(t))^2 + (y_b(t) - y_o(t))^2} \quad [3]$$

with $(x_b(t), y_b(t))$ and $(x_o(t), y_o(t))$ the coordinates of the simulated particles barycenter and observed drifter at time t, respectively.

Absolute dispersion is defined as the distance to the initial position at each time step:

$$D_0(t) = \sqrt{(x_d(t) - x_d(t_0))^2 + (y_d(t) - y_d(t_0))^2} \quad [4]$$

where $(x_d(t), y_d(t))$ are the coordinates of the simulated particles barycenter or observed drifter at time t and $(x_d(t_0), y_d(t_0))$ their initial position.

3. Results

3.1. Examples of drifters and simulated trajectories

As a first example, the ROMS-Ichthyop simulation reproduced the path followed by the drifter 55152 quite well, while with the OSCAR-Ichthyop simulation most of the particles were swept away to the east (Figure 1). Here, the initial position of the drifter is close to the divergence zone between the south and north branches of the SEC. A small difference in positioning of this zone between ROMS and OSCAR leads to, in the first case, most particles ended up close to the South American coast, in agreement with the observed trajectory, whereas in the second case, they mostly ended up close to the African coast, at the other end of the basin. As a contrasting, second example, we note that it is the OSCAR-Ichthyop simulation that follows the drifter 13513 trajectory best. With ROMS-Ichthyop, most particles are transported westwards, whereas the drifter goes eastwards (Figure 2) through the NECC. We also note that there is more variability in the trajectories simulated by the ROMS-Ichthyop simulation.

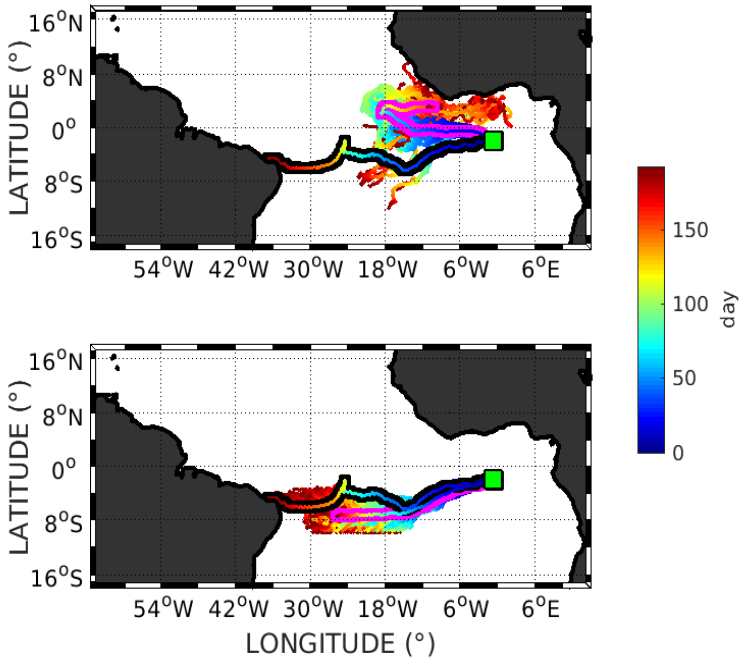


Figure 1. *Trajectories of the NOAA drifter 55152 (black), of the particles simulated using OSCAR-Ichthyop (top) and ROMS-Ichthyop (bottom), and particles barycenter (pink).*

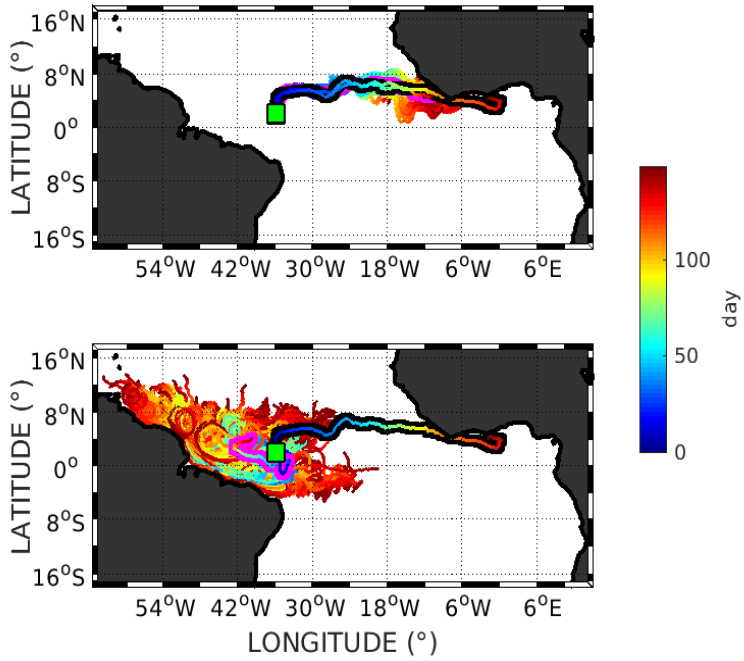


Figure 2. *Same as Figure 1 for NOAA drifter 13513.*

3.2. Comparison of all drifters and simulations

For all 278 NOAA drifters, the error distance between the observed trajectories and those simulated by Ichthyop forced by ROMS or OSCAR are close, increasing linearly with time and then stabilizing around 180 days (Figure 3). The error is slightly larger in the ROMS-Ichthyop simulation. On the other hand, the distances traveled between origin and destination are slightly closer to the observations in the case of that simulation (Figure 4). After 200 days, the drifters are on average more than 2000 km away from their release point whereas the particles in ROMS-Ichthyop are only at 1800 km and those in OSCAR-Ichthyop at 1200 km. The simulated speeds are therefore significantly lower than the observed speeds in both cases.

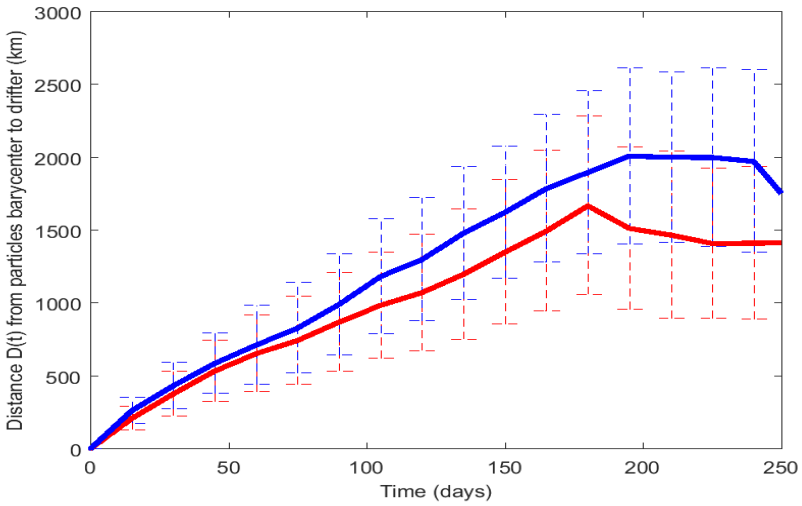


Figure 3. *Distance from the simulated particles barycenter to the observed drifter location over time, averaged for the 278 drifters found in the tropical Atlantic in the 1992-2008 period. Mean (plain) and standard deviation (dash) are in red for OSCAR-Ichthyop and in blue for ROMS-Ichthyop.*

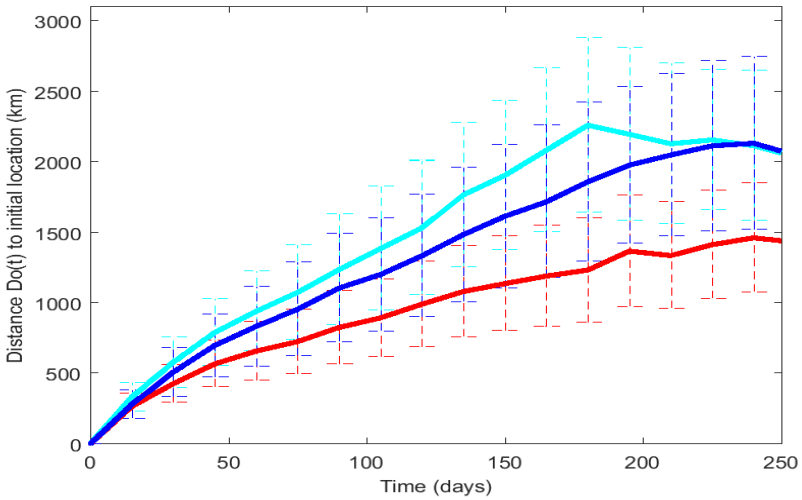


Figure 4. *Distance to initial locations of simulated particles barycenter and observed drifter over time, averaged for the 278 drifters found in the tropical Atlantic in the 1992-2008 period. Mean (plain) and standard deviation (dash) are in red for OSCAR-Ichthyop, in blue for ROMS-Ichthyop, and in cyan for observed drifters.*

4. Discusion

The trajectories simulated by the Lagrangian model Ichthyop forced by the OSCAR satellite product or by the ROMS hydrodynamic model is globally satisfactory to the extent that the main currents of the tropical Atlantic region are obtained. However, differences due to vortices, areas of divergence and variability in each of the considered regions can have significant consequences in terms of simulated drifting trajectories. Despite good results obtained by one or the other forcing product for some drifters (Figs 1 and 2), on average the simulated particles barycenter and observed drifter trajectories differed significantly. Averaged over all 278 drifters in our studied domain, the distance between particles barycenters and drifters was 1600 km for OSCAR-Ichthyop and 1800 km for ROMS-Ichthyop after 180 days (Figure 3), the same order of magnitude as the value of 229 km after 20 days obtained by Price et al. (2006). We found that this error distance increased approximately linearly with time, like in many other studies (LaCasce and Ohlmann, 2003). We also found that the velocities of simulated particles were lower than those of drifters, as in other circulation models (Doos et al., 2011), likely related to the insufficient spatial resolution used for forcings ($\frac{1}{5}^\circ$ grid for ROMS and $\frac{1}{3}^\circ$ for OSCAR). It is clear that at such resolutions, mesoscale structures such as eddies are not represented correctly. Recent works such as Jorda et al. (2014) detail the importance of including mesoscale currents as forcing factor, especially in areas of strong currents. They showed that these currents should be selected at the highest possible frequency because the changes in their frequency can have a significant impact on the modeled trajectories. Similarly, Doos et al. (2011) showed that a good agreement between the simulated speeds and drifters speeds can be obtained provided sufficiently fine resolutions in space and time are used. Regarded time, frequencies we used (2 days for ROMS and 5 days for OSCAR) may also be insufficient to account for oceanic variability. It is also often pointed that low values of simulated velocities can come from the coarse resolution of the field atmospheric winds (McClean, 2002). Indeed, atmospheric forcing is often taken at 2 m in height and not at the surface of the ocean which can cause differences at the level of modeled results. The effect of drifter slip is also poorly simulated in forcing atmospherics, leading to lower particle dynamics (Edwards et al., 2006). These limitations point to the need to develop higher resolution solutions, which will be done in future work by using the capacity of ROMS of embedding smaller but higher resolution (child) grids within larger but lower resolution (parent) domain (Debreu et al., 2012; Djakouré et al., 2014).

5. References

[1] Bonjean, F., and G. S. E. Lagerloef (2002), Diagnostic model and analysis of the surface currents in the Tropical Pacific Ocean, J. Phys. Oceanogr., 32, 2938-2954.

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

[3] Djakouré, S., P. Penven, B. Bourlès s, J. Veitch, and V. Koné (2014), Coastally trapped eddies in the north of the Gulf of Guinea, J. Geophys. Res. Oceans, 119, 6805?6819, doi :10.1002/2014JC010243.

[4] Döös, K., V. Rupolo and L. Brodeau (2011), Dispersion of surface drifters and model-simulated trajectories, Ocean Modell., 39, 301-310.

[5] Edwards, K. P., F. E. Werner, and B. O. Blanton (2006), Comparison of observed and modeled drifter trajectories in coastal regions : An improvement through adjustments for ob-

served drifter slip and errors in wind fields, J. Atmos. Oceanic Technol., 23, 1614-1620, doi : 10.1175/JTECH1933.1.
[6] Hansen, D. and P.-M. Poulain, 1996 : Quality control and interpolations of WOCETOGA drifter data. J. Atmos. Oceanic Technol., 13, 900-909.
[7] Jorda, G., E. Comerma, R. Bolanos, M. Espino (2014). Impact of forcing errors in the CAM-CAT oil spill forecasting system : A sensitivity study. Journal of Marine Systems.
[8] Koné V, Lett C, Penven P, Bourlès B, Djakouré S (2017). A biophysical model of S. aurita early life history in the northern Gulf of Guinea. Progress in Oceanography 151 : 83-96.
[9] LaCasce, J.H. and C. Ohlmann, 2003. Relative dispersion at the surface of the Gulf of Mexico. J. Mar. Res., 61, pp. 285-312.
[10] Lett, C., Penven, P., Ayón, P., and Fréon, P., 2007b. Enrichment, concentration and retention processes in relation to anchovy (Engraulis ringens) eggs and larvae distributions in the northern Humboldt upwelling ecosystem. Journal of Marine Systems, 64, 189-200.
[11] Lett C., Verley P., Mullon C., Parada C., Brochier T., Penven P., Blanke B., 2008. A Lagrangian tool for modelling ichthyoplankton dynamics. Environmental Modelling and Software, 23 :1210-1214.
[12] Lumpkin, R., Treguier, A.M., Speer, A.K., 2002. Lagrangian eddy scales in the Northern Atlantic Ocean. J. Phys. Oceanogr. 32, 2425-2440.
[13] McClean, J.L., Poulain, P.-M., Pelton, J.W., Maltrud, M.E., 2002. Eulerian and Lagrangian statistics from surface drifters and a high-resolution POP simulation in the North Atlantic. J. Phys. Oceanogr. 32, 2472-2491.
[14] Peliz, A., Marchesiello, P., Dubert, J., Marta-Almeida, M., Roy, C., Queiroga, H., 2007. A study of crab larvae dispersal on the western Iberian shelf : physical processes. Journal of Marine Systems, doi : 10.1016/j.jmarsys.2006.11.007.
[15] Price, J.M., Reed, M., Howard, M.K., Johnson, W.R., Ji, Z.G., Marshall, C.F., Guinasso, N.L., Rainey, G.B., 2006. Preliminary assessment of an oil-spill trajectory model using satellite-tracked, oil-spill-simulating drifters. Environ. Modell. Softw. 21 (2), 258-270.
[16] Shchepetkin, A., and J. McWilliams (2005), The regional oceanic modeling system (ROMS) : A split-explicit, free-surface, topographyfollowing- coordinate oceanic model, Ocean Modell., 9, 347-404.

Appendix 1

We did not proceed to model calibration in our work but performed a sensitivity analysis of our results to model parameters. As an example, in appendices 1 and 2 we show the results obtained for Figure 3 and Figure 4 when simulations are performed without the horizontal dispersive component (Figure 5 and Figure 6) and with release of particles in the neighborhood (disc of 50 km with/without diffusion and of 200 km radius R, Figure 7 and Figure 8) of observed drifters, as opposed to with horizontal dispersive component and release of particles at the exact drifters location in the main text. We essentially found the same results, showing the robustness of those presented in the main text.

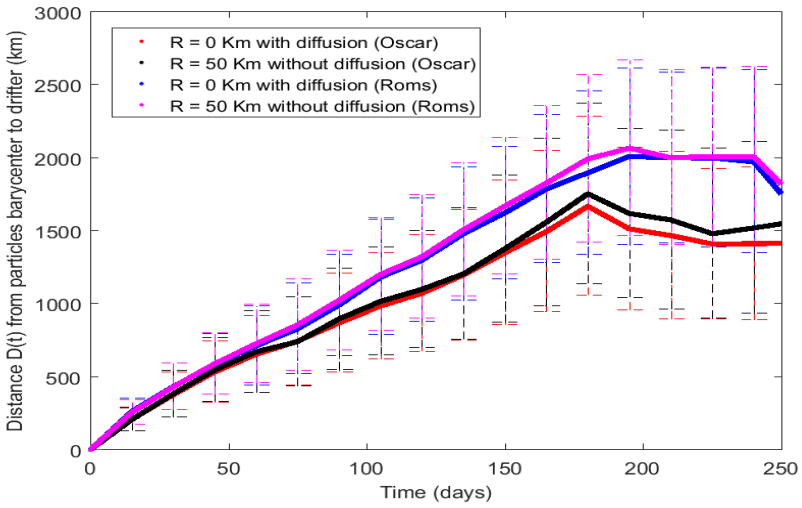


Figure 5. *Distance from the simulated particles barycenter to the observed drifter location over time, averaged for the 278 drifters found in the tropical Atlantic in the 1992-2008 period.*

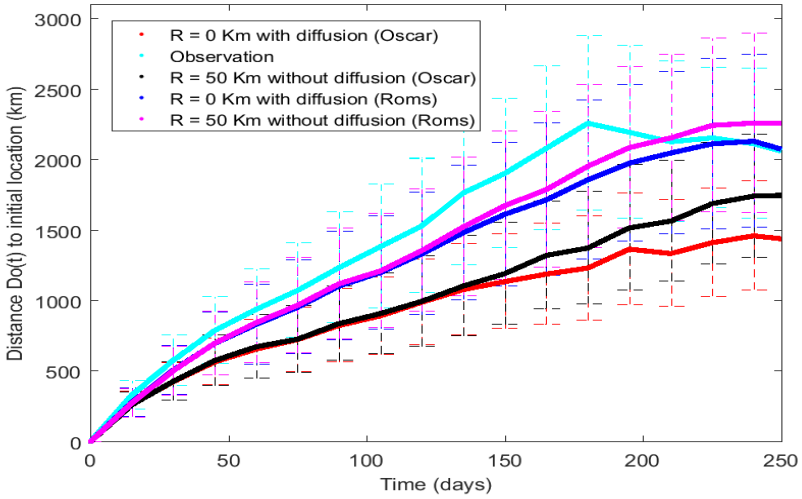


Figure 6. *Distance to initial locations of simulated particles barycenter and observed drifter over time, averaged for the 278 drifters found in the tropical Atlantic in the 1992-2008 period.*

Appendix 2

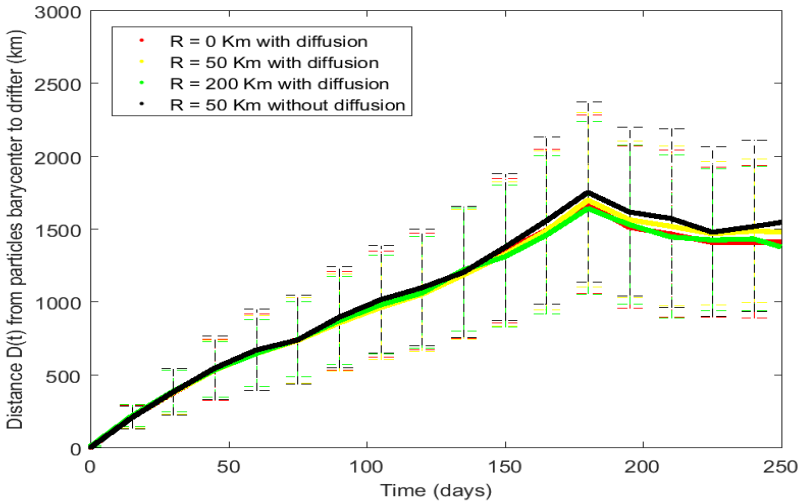


Figure 7. *Distance from the simulated particles barycenter (OSCAR) to the observed drifter location over time, averaged for the 278 drifters found in the tropical Atlantic in the 1992-2008 period.*

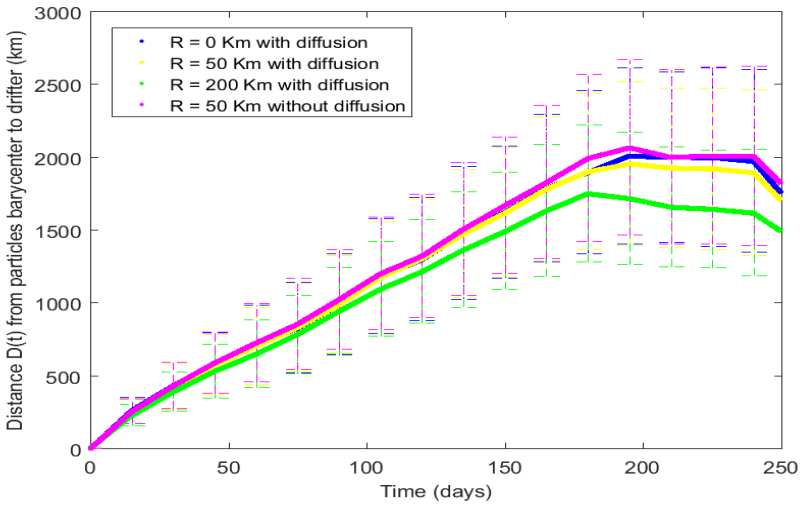


Figure 8. *Distance from the simulated particles barycenter (ROMS) to the observed drifter location over time, averaged for the 278 drifters found in the tropical Atlantic in the 1992-2008 period.*

Modelling and control of coffee berry borer infestation

Yves Fotso[1, 2, 4, 5, †] — Frédéric Grognard[2, 5] — Berge Tsanou[1, 4, 5] — Suzanne Touzeau[2, 3, 5]

[1] Department of Mathematics and Computer science, Faculty of science, University of Dschang, P. O. Box: 67 Dschang, Cameroon.
† Corresponding author: fotsofyves@yahoo.fr

[2] Université Côte d'Azur, INRIA, INRA, CNRS, UPMC Univ Paris 06, BIOCORE, France.

[3] Université Côte d'Azur, INRA, CNRS, ISA, France.

[4] UMI 209 IRD & UPMC UMMISCO, Bondy, France.

[5] INRIA, University of Douala, Dschang, Yaoundé I, LIRIMA, EPITAG, France

ABSTRACT. In this paper, we developed a mathematical model that describes the infestation dynamics of coffee berry by *Hypothenemus hampei* (CBB). This model takes into account control some integrated pest management strategies which are used by coffee growers to eradicate CBB in plantations, represented by two functions depending on time. We design these functional controls to maximize the yield of healthy berries at the end of the cropping season, while minimising the borer population for the next cropping season and the control costs. By using optimal control theory, we show that an optimal control exists for this problem and Pontryagin's maximum principle is used to characterize an optimal control. Numerical simulations are provided to illustrate our results.

RÉSUMÉ. Dans ce papier, nous développons un modèle mathématique qui décrit la dynamique d'infestation des baies du caféier par les scolytes (CBB). Ce modèle prend en compte certaines stratégies de lutte intégrée utilisées par les caféiculteurs pour éradiquer les CBB dans les plantations, représentées par deux fonctions dépendantes du temps. Nous concevons ces contrôles fonctionnels pour maximiser le rendement des baies saines à la fin de la saison, tout en minimisant la population de CBB pour la prochaine saison et les coûts de contrôle. En utilisant la théorie du contrôle optimal, nous montrons qu'un contrôle optimal existe pour ce problème et le principe du maximun de Pontryagin est utilisé pour caractériser ce contrôle optimal. Des simulations numériques sont faites pour illustrer nos résultats.

KEYWORDS : *Hypothenemus hampei*, optimal control, numerical simulations

MOTS-CLÉS : *Hypothenemus hampei*, contrôle optimal, simulations numériques

1. Introduction

Coffee plays an important role in the economic growth of many developing countries such as Brazil, Cameroon, Ethiopia, Ivory Coast, Mexico, Viet Nam and many others. Coffee production throughout the world is affected by several pests and diseases. Among these pests, the coffee berry borer (CBB), *Hypothenemus hampei*, is considered as the most important pest economically [2, 1, 9]. The CBB feeds and spends its entire development cycle in the berries, developing in the berry of the coffee tree in all maturation stages. It causes direct loss such as a reduction of coffee production and indirect losses such as a lowering of the quality of the coffee berries. Sibling mating inside the berry makes this insect quite difficult to control. The levels of infestation of CBB in coffee growing areas are estimated at 60% in Mexico, $50-90\%$ in Malaysia, 60% in Colombia, 75% in Jamaica, 80% in Uganda and 90% in Tanzania [8]. It is now present in almost all of the major coffee producing countries. It lives the greatest part of its life cycle inside the coffee berry, which involves egg laying followed by the emergence of adult females from the berry. The life cycle of the CBB is composed of four stages: eggs, larvae, nymphs and adults. Mature females are responsible for the dispersal of the population: they emerge from the berries to colonize and lay their eggs in new berries, while males and larvae stages remain inside the berries. Faced with the extent of the damages, several programs and control methods have been developed by the coffee growers, such as improved cultural practices, chemical and biological control and trapping [2, 1, 9]. In this paper, we propose an epidemiological model to describe the dynamics of infestation of coffee berries by CBB. This model takes into account several control strategies. Our aim is to design an optimal strategy, that maximizes the yield of healthy berries at the end of the cropping season, while minimizing the CBB population for the next season.

After a formulation of the model and the control problem in Section 2, we study the stability of equilibrium points in the absence of control in Section 3. In Section 4, we prove the existence of the optimal control which is later characterized in Section 5. We illustrate these analytical results by simulations in Section 6 and conclude our work.

2. The model

We propose an epidemiological model of infestation of coffee berries by CBB. We subdivide the total number of coffee berries into two compartments: the healthy coffee berries s, and the infested coffee berries i. We assume that the new coffee berries are produced at a constant rate Λ and we place ourselves during a cropping season. We assume that adult males are not limited, so we only consider female CBB, which are responsible for dispersal and host selection. We subdivide the female population in two compartments: the colonizing females or host-searching females which emerge from the berries and search for a new host denoted by y and the infesting females who can find and infested the coffee berry and denoted by z. The healthy berries are submitted to a force of infestation $\beta\frac{y}{y+d}$ by colonizing females, which compete for this resource, hence the saturation term. We denote by μ and δ the natural mortality of healthy and infested coffee berries respectively. According to the biology of CBB, we assume that the average number of new colonizing females produced is taken proportional of the number of infesting females. Let ϕ be that average number per unit of time. We denote by μ_y and μ_z the natural mortality rates of colonizing and infesting females respectively. We denote by ε

the conversion parameter from the coffee berries to CBB population, that is the number of CBB colonizing females that can infest one unit of healthy berry. Usually $\varepsilon = 1$, but if the infestation process fails, $\varepsilon < 1$. The first control u represents the efforts made to reduce the infestation of healthy coffee berries. In practice, this control function represents the biological control using entomopathogenic fungi such as *Beauveria bassiana*, that is applied to the surface of the coffee berries and that kills the colonizing females of CBB when they drill an entry hole into the coffee berries [1]. The second control v represents the efforts made to reduce the colonizing females. It consists mainly of the use of chemicals, traps and the parasitoids. We associate to these control functions, the parameters $\alpha_i \in (0,1), i = u, v$ which measure the effectiveness of control u and v respectively. The dynamics of CBB is given by the following nonlinear differential equations:

$$\begin{cases} s' = \Lambda - (1-\alpha_u u)\beta\dfrac{sy}{y+d} - \mu s \\ i' = (1-\alpha_u u)\beta\dfrac{sy}{y+d} - \delta i \\ y' = \phi z - \varepsilon\beta\dfrac{sy}{y+d} - (\mu_y + \alpha_v v)\, y \\ z' = (1-\alpha_u u)\varepsilon\beta\dfrac{sy}{y+d} - \mu_z z \end{cases} \tag{1}$$

Since variable i does not interact with the other variables of system (1), its dynamics can be decoupled from the system. We will focus on the dynamics of the other three variables. Thus we obtain the following system:

$$\begin{cases} s' = \Lambda - (1-\alpha_u u)\beta\dfrac{sy}{y+d} - \mu s \\ y' = \phi z - \varepsilon\beta\dfrac{sy}{y+d} - (\mu_y + \alpha_v v)\, y \\ z' = (1-\alpha_u u)\varepsilon\beta\dfrac{sy}{y+d} - \mu_z z \end{cases} \tag{2}$$

The goal of coffee farmers is the production of high quality coffee at the best market price produced at lowest cost. So our problem consists in maximizing the yield at the end of the cropping season, while minimizing the coffee berry borer population for the next season. Since all these methods of control are expensive and require a lot of energy in their implementation. We propose the following objective function:

$$\mathcal{J}(u,v) = \int_0^{t_f} \frac{1}{2}\left[C_u u^2(t) + C_v v^2(t)\right] dt - D_s s(t_f) + D_y y(t_f) \tag{3}$$

where t_f represents the time at the end of the cropping season and the parameters, C_u and C_v measure the relative cost of interventions associated with controls u and v respectively; D_s and D_y represent the weights of healthy coffee berries and of colonizing females at the end of the season respectively. The set of admissible controls is defined as follows

$$\mathcal{U} = \left\{u, v \in L^1(0,t_f)/(u,v) \in [0,1]\times[0,1], \forall t \in [0,t_f]\right\} \tag{4}$$

Table 1. *Biological meaning and value of parameters (with s in number of berries and y, z in number of females).*

symbol	Description	value
Λ	Production rate of new coffee berries	1200 berries day^{-1}
μ	Natural mortality rate of healthy coffee berries	0.01 day^{-1}
ϕ	Emergence of new colonizing females	2 day^{-1}
β	Infestation rate	0.0125 day^{-1}
d	Saturation constant	2 females
ε	Conversion rate from coffee berries to CBB	1 female berry^{-1}
μ_y	Natural mortality rate of colonizing females	1/81 day^{-1}
μ_z	Natural mortality rate of infesting females	1/28 day^{-1}
α_u	Effectiveness rate of control $u(t)$	0.62
α_v	Effectiveness rate of control $v(t)$	0.31

The problem now is to find the control pair $(u^\star, v^\star)$ satisfying:

$$\mathcal{J}(u^\star, v^\star) = \min_{(u,v)\in\mathcal{U}} \mathcal{J}(u, v) \tag{5}$$

3. Basic properties

For model (2) to be biologically acceptable, it is important to show that all these variables are always positive when time evolves.

Theorem 1 *If the initial condition $(s(0), y(0), z(0)) \in \mathbb{R}^3_+$, then the solution $(s(t), y(t), z(t))$ of system (2) are non negative for all time $t > 0$ and bounded. Moreover, the compact set*

$$\Omega = \left\{(s, y, z) \in \mathbb{R}^3_+ / s \leq \frac{\Lambda}{\mu},\ \varepsilon s + z \leq \frac{\varepsilon\Lambda}{\xi},\ y \leq \frac{\varepsilon\phi\Lambda}{\xi\mu_y}\right\}$$

where $\xi = \min\{\mu, \mu_z\}$, is positively invariant for the model system (2).

Proof: See Appendix A.

In the absence of controls ($u = v = 0$), system (2) has one trivial equilibrium $\mathcal{E}^0 = (s^0, 0, 0)$ where $s^0 = \frac{\Lambda}{\mu}$, which corresponds to a plantation without infestation. Thereafter, we will define the basic offspring number which is the average number of new females originated from a single infesting female in the healthy coffee berries in plantation. The basic offspring number is defined by

$$\mathcal{N} = \frac{\varepsilon\phi\beta\frac{s^0}{d}}{\mu_z\left(\varepsilon\beta\frac{s^0}{d} + \mu_y\right)}. \tag{6}$$

Lemma 1 *There exists another coexistence equilibrium $\mathcal{E}^* = (s^*, y^*, z^*)$ which is biologically realistic when $\mathcal{N} > 1$:*

$$s^* = \frac{\Lambda + \frac{\mu_y d}{\varepsilon\mathcal{T}}}{\beta + \mu}, \quad y^* = \frac{\mu_z}{\mu_y}\mathcal{T}z^*, \quad z^* = \frac{\mu d\left(\frac{\varepsilon\beta s^0}{d} + \mu_y\right)}{\mu_z(\beta + \mu)\mathcal{T}}(\mathcal{N} - 1) \quad with \quad \mathcal{T} = \frac{\phi}{\mu_z} - 1.$$

It is easy to prove that $\mathcal{N} > 1$ implies $\mathcal{T} > 0$.

The long term behavior of model system (2) without controls is given by:

Proposition 1 *1) The trivial equilibrium $\mathcal{E}^0$ is locally asymptotically stable whenever $\mathcal{N} < 1$, and unstable otherwise.*

2) The coexistence equilibrium $\mathcal{E}^*$ exists and is locally asymptotically stable whenever $\mathcal{N} > 1$.

Proof: See Appendix B.

The aim of our control problem is to prove the existence of the optimal control and uniqueness of the optimality system and the characterisation of the optimal control.

4. Existence of an optimal control

The existence of an optimal control is obtained by the theorem of Fleming and Richer [4].

Theorem 2 *There exists an optimal control pair $(u^\star, v^\star)$ and a corresponding solution $(s^\star, y^\star, z^\star)$ of the initial value problem (2) that minimizes the cost function $\mathcal{J}$ in $\mathcal{U}$ such that*

$$\mathcal{J}(u^\star, v^\star) = \min_{(u,v)\in\mathcal{U}} \mathcal{J}(u, v) \tag{7}$$

Proof: we use Theorem 4.1 in Fleming and Ricker [4] which gives the conditions of existence of optimal control for the optimal system (2), which we recall here for self-containeness:

(i) the set of controls and corresponding state variables is non-empty;

(ii) the control set $\mathcal{U}$ is convex and closed;

(iii) the right hand side of the state system (2) is bounded by a linear function in the state and control variable;

(iv) there exist constants $\zeta_1, \zeta_2 > 0$ and $\beta > 1$ such that the integrand function define by f^0 of the objective functional satisfies $f^0(t, \tilde{x}, \tilde{u}) \geq \zeta_1 \|\tilde{u}\|^\beta - \zeta_2$ for all $t \in [0, t_f]$.

The existence of the solution of system (2) is obtained in using the result from Lukes [5](Theorem 9.2.1), since system (2) has bounded coefficients and any solution is bounded on the finite interval time $[0, t_f]$, so condition (i) is satisfied. By definition, the control set $\mathcal{U}$ is convex and closed, so condition (ii) is satisfied. The right hand side of the state system satisfies condition (iii) since we have a linear dependence of the state equations on controls u and v. Finally, the integrand function f^0 of the objective functional is clearly convex in the controls since it is quadratic. Moreover, since all states are bounded, it is easy to find $\zeta > 0$ such that we have $f^0(t, \tilde{x}, \tilde{u}) = \frac{1}{2}(C_u u^2 + C_v v^2) \geq \frac{1}{2}\min\{C_u, C_v\}(u^2 + v^2) - \zeta \geq \frac{1}{2}\min\{C_u, C_v\}\|\tilde{u}\|^2 - \zeta$ which proves property (iv). We conclude that there exists an optimal control pair $(u^\star, v^\star)$ that minimizes the cost function $\mathcal{J}$ in $\mathcal{U}$.

5. Characterization of the optimal control

Since an optimal control minimizing the objective function (3) exists, we use Pontryagin's principle [7] to have the necessary conditions for the optimal control $u^\star$ and

$v^\star$ of our control problem . Let $\tilde{x} = (s, y, z)$ and $\tilde{u} = (u, v) \in \mathcal{U}$. According to this principle, there exists a nontrivial absolutely continuous mapping $\lambda : [0, t_f] \longrightarrow \mathbb{R}^3$, $t \longmapsto \lambda(t) = (\lambda_1(t), \lambda_2(t), \lambda_3(t))$ called the adjoint vector containing the adjoint variables. We define the Hamiltonian by

$$\begin{aligned}\mathcal{H}(\tilde{x}, \lambda, \tilde{u}) &= \frac{1}{2}\left[C_u u^2 + C_v v^2\right] + \lambda_1\left[\Lambda - (1-\alpha_u u)\beta\frac{sy}{y+d} - \mu s\right] \\ &+ \lambda_2\left[\phi z - \varepsilon\beta\frac{sy}{y+d} - (\mu_y + \alpha_v v)y\right] + \lambda_3\left[(1-\alpha_u u)\varepsilon\beta\frac{sy}{y+d} - \mu_z z\right].\end{aligned}$$

Theorem 3 *Given an optimal control $(u^\star, v^\star)$ and corresponding solutions $(x^\star, y^\star, z^\star)$, there exist adjoint variables $\lambda_i(t)$ for $i = 1, 2, 3$ satisfying the following system of linear differential equation*

$$\begin{cases} \lambda_1' = \left[(\lambda_1 - \varepsilon\lambda_3)(1-\alpha_u u) + \varepsilon\lambda_2\right]\beta\dfrac{y}{y+d} + \mu\lambda_1 \\ \\ \lambda_2' = \left[(\lambda_1 - \varepsilon\lambda_3)(1-\alpha_u u) + \varepsilon\lambda_2\right]\beta\dfrac{ds}{(y+d)^2} + \lambda_2(\mu_y + \alpha_v v) \\ \\ \lambda_3' = -\lambda_2\phi + \mu_z\lambda_3 \end{cases} \tag{8}$$

for almost all $t \in [0, t_f]$, with transversality conditions $\lambda_1(t_f) = -D_s$, $\lambda_2(t_f) = D_y$ and $\lambda_3(t_f) = 0$. Furthermore, we can characterize the optimal control pair by

$$\begin{aligned} u^\star(t) &= \min\left\{\max\left\{0, \frac{1}{C_u}(\varepsilon\lambda_3 - \lambda_1)\alpha_u\beta\frac{sy}{y+d}\right\}, 1\right\}; \\ v^\star(t) &= \min\left\{\max\left\{0, \frac{\alpha_v}{C_v}\lambda_2 y\right\}, 1\right\}. \end{aligned}$$

Proof: We use the direct application of Pontryagin's maximum principle for bounded control [7]. The differential equations governing these adjoint variables $(\lambda_i)_{i=\{1,2,3\}}$ are obtained by differentiation of the Hamiltonian (8), evaluated at the optimal control:

$$\lambda_1' = -\frac{\partial\mathcal{H}}{\partial s}, \quad \lambda_2' = -\frac{\partial\mathcal{H}}{\partial y}, \quad and \quad \lambda_3' = -\frac{\partial\mathcal{H}}{\partial z}$$

and the traversality conditions are obtained by $\lambda_1(t_f) = \left[\frac{\partial\Theta}{\partial s}\right]_{t=t_f}$, $\lambda_2(t_f) = \left[\frac{\partial\Theta}{\partial y}\right]_{t=t_f}$ and $\lambda_3(t_f) = \left[\frac{\partial\Theta}{\partial z}\right]_{t=t_f}$ with the function $\Theta(\tilde{x}) = -D_s s + D_y y$. To determine an explicit expression for the optimal control $(u^\star, v^\star)$, we use the standard optimality technique given in [6]. On the set $I_1 = \{t \in [0, t_f] : 0 < u^\star(t) < 1 \qquad 0 < v^\star(t) < 1\}$; The minimum condition is

$$\frac{\partial\mathcal{H}}{\partial u} = C_u u + (\lambda_1 - \varepsilon\lambda_3)\alpha_u\beta\frac{sy}{y+d} = 0, \qquad \frac{\partial\mathcal{H}}{\partial v} = C_v v - \alpha_v\lambda_2 y = 0;$$

Thus, the controls have the explicit expression given by:

$$u^\star(t) = \frac{1}{C_u}(\varepsilon\lambda_3 - \lambda_1)\alpha_u\beta\frac{sy}{y+d} \quad and \quad v^\star(t) = \frac{\alpha_v}{C_v}\lambda_2 y.$$

On the set $I_2 = \{t \in [0, t_f] : u^\star(t) = 0\}$, the minimum condition of control $u^\star$ is given by $\frac{\partial \mathcal{H}}{\partial u} \geq 0$, which implies that $\frac{1}{C_u}(\varepsilon\lambda_3 - \lambda_1)\alpha_u\beta\frac{sy}{y+d} \leq 0$.

On the set $I_3 = \{t \in [0, t_f] : u^\star(t) = 1\}$, then the minimum condition of control $u^\star$ is $\frac{\partial \mathcal{H}}{\partial u} \leq 0$, which implies that $\frac{1}{C_u}(\varepsilon\lambda_3 - \lambda_1)\alpha_u\beta\frac{sy}{y+d} \geq 0$. All these criteria on the control $u^\star$ can be written in the compact form given in the theorem. The characterization of control $v^\star$ is obtained in a similar way.

6. Numerical simulations

In this section, we present the numerical solution of our control problem and compare it with the solution in the absence of controls. We use the forward-backward sweep method to solve numerically our optimal model[6]. The process begins by using an initial guess on the control variable, then the state variables are solved simultaneously forward in time with a semi-implicit finite difference method developed and the adjoint equations are solved using the backward semi-implicit finite difference method. The controls are updated by inserting the new values of state and adjoint variables into its characterization. We assume that the implementation costs of these controls are C_u=3 day^{-1} and C_v= 1 day^{-1} and we use the following weights, D_s = 1 berry^{-1} and D_y=1 female^{-1} with initial values $(s(0), y(0), z(0)) = (0, 20, 0)$. The other parameter values are given in Table 1. Since coffee berries become mature after 8–9 months, we simulate the system (2) over a period $t_f = 250$ days.

The simulations plots are given in figure 1. We compare the cases with (dashed red curves) and without (plain blue curves) controls in the presence of pests. We observe that the control $u(t)$ (lower-right panel: black curve) significantly reduces infestation (lower-left panel) and increases at its maximum value at almost mid-season, while control $v(t)$ (lower-right panel: magenta curve) greatly reduces colonizing females at the end of season (upper-right panel: 6×10^4 females instead of 2.5×10^6 at t_f). The fairly long and high application of these controls, especially of $u(t)$, is due to the relatively low costs of the controls. The yield increases with the controls (upper-left) but, due to the limited effectiveness of the controls (parameters α_u and α_v), it remains below its value in the pest-free case (plain green curve).

7. Conclusion

In this paper, we formulate a deterministic epidemiological control model that describes the infestation of coffee berries by the CBB. We have designed an optimal control problem that consists in maximizing the yield of healthy berries at the end of the cropping season, while minimizing the CBB population for the next season. We have computed the basic offspring number and investigated the existence and stability of equilibria in the absence of controls. We have showed that an optimal control exists and that it can be characterized using the Pontryagin's maximum principle. Furthermore, we have solved numerically the system to assess the role of controls on dynamics of CBB population. This numerical result shows that, the application of these controls reduce the CBB population and increase the healthy berries at the end of the cropping season.

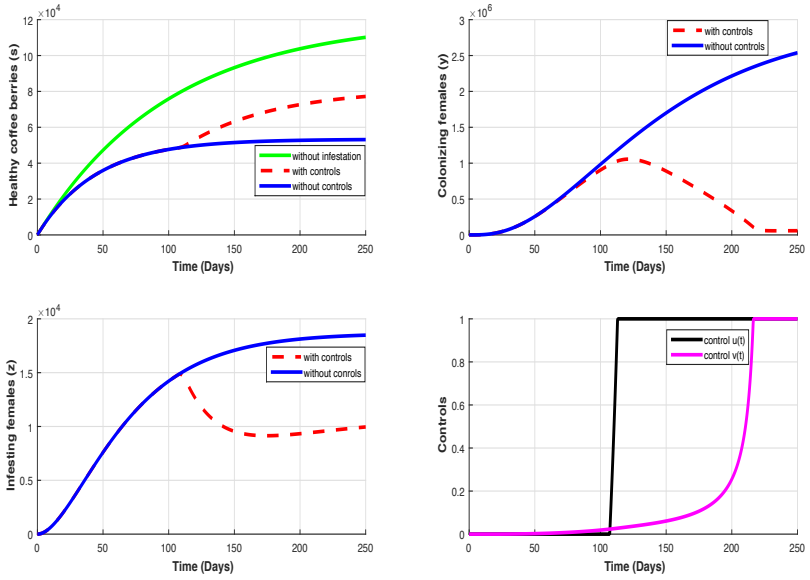


Figure 1. *Simulations of system (2) with (plain blue curves) or without (dashed red curves) controls. Upper-left panel: healthy coffee berries s (plain green curve: pest-free case); Upper-right: colonizing females y; Lower-left: infesting females z; Lower-right: evolution of controls $u(t)$ (plain black curve) and $v(t)$ (plain magenta curve).*

8. References

[1] L. F. ARISTIZÁBAL, A. E. BUSTILLO, S. P. ARTHURS, "Integration pest management of coffee berry borer: strategies from latin america that could be useful for coffee farmers in hawaii", *Insects*, 2016, 7,6.

[2] A. DAMON, "A review of the biology and control of the coffee berry borer, hypothenemus hampei (coleoptera: Scolytidae)", *Bulletin of entomological research*, (2000) 90, 453 - 46.

[3] P. DRIESSCHE, J. WATMOUGH, "Reproduction numbers and subthreshold endemic equilibria for compartment models of disease tranmission", *Math Biosci*, 2002.

[4] W. H. FLEMING, *R. W. Rishel*, "Determinstc and Stochastic optimal control", Springer-Verlag, New York, 1975.

[5] D. L. LUCKES, "*Differential Equations: Classical to controlled*", Academic Press, New York, 1982.

[6] S. LENHART, J. T. WORKMAN, "*Optimal control applied to biological models*", Taylor & Francis Group, 2007.

[7] L. S. PONTRYAGIN, V. G. BOLTYANSKII, E. F. GAMKRELIZE, E. F. MISHCHENKO, "*The Mathematical theory of optimal processes*", Wiley, New York, 1962.

[8] F. E. VEGA, "Coffee berry borer *Hypothenemus hampei* (Ferrari)(Coleoptera: Scolytidae). in capinera ", *J.L.(Ed.), Encyclopedia of Entamology*, 2004, 575-576.

[9] F. E. VEGA, F. INFANTE, A. CASTILLO, J. JARAMILLO, "The coffee berry borer, *Hypothenemus hampei* (ferrari) (coleoptera: curculionidae): a short review, with recent findings and future research directions", *Terrestrial Arthropod Reviews 2(2)*, (2009), 129 - 147.

Appendix. A:

For any initial condition $(s(0), y(0), z(0)) \in \mathbb{R}^3_+$, the corresponding solution $(s(t), y(t), z(t))$ of system (1) lies in $\mathbb{R}^3_+$, since we have $s'_{|_{s=0}} = \Lambda > 0$, $y'_{|_{y=0}} = \phi z \geq 0$ and $z'_{|_{z=0}} = (1-\alpha_u u)\varepsilon\beta\frac{sy}{y+d} \geq 0$. Therefore, all solutions of system (2) with initial positive condition stay in first quadrant. Hence, $\mathbb{R}^3_+$ is positively invariant.

Since all variables are nonnegative for all $t > 0$, then $s'(t) \leq \Lambda - \mu s(t)$. It can be shown that using a standard comparison principle, that

$$s(t) \leq s(0)e^{-\mu t} + \frac{\Lambda}{\mu}\left(1 - e^{-\mu t}\right) \tag{9}$$

from which we deduce that $s \leq \frac{\Lambda}{\mu}$ if $s(0) \leq \frac{\Lambda}{\mu}$ and we have $\overline{\lim}\, s(t) \leq \frac{\Lambda}{\mu_S}$.

Let $\tilde{z}(t) = \varepsilon s(t) + z(t)$ and $\xi = \min\{\mu, \mu_z\}$, then adding the first and third equation of model system (2), we obtain

$$\tilde{z}(t) = \varepsilon\Lambda - \varepsilon\mu y - \mu_z z \leq \varepsilon\Lambda - \xi\tilde{z}(t). \tag{10}$$

In particular $\tilde{z}(t) \leq \frac{\varepsilon\Lambda}{\xi}$ if $z(0) \leq \frac{\varepsilon\Lambda}{\xi}$ and we have $\limsup_{t\to\infty} \tilde{z}(t) \leq \frac{\varepsilon\Lambda}{\xi}$. Now, using the second equation of model system (2), we have

$$y'(t) \leq \phi z(t) - \mu_y y \leq \frac{\phi\varepsilon\Lambda}{\xi} - \mu_y y. \tag{11}$$

proceeding in the same way as previously, then $y(t) \leq \frac{\phi\varepsilon\Lambda}{\xi\mu_y}$ if $y(0) \leq \frac{\phi\varepsilon\Lambda}{\xi\mu_y}$ and we have $\limsup_{t\to\infty} y(t) \leq \frac{\phi\varepsilon\Lambda}{\xi\mu_y}$.

Appendix. B:

A method for computing the basic reproduction number in epidemiological models which corresponds to the number of secondary infections produced by a single infectious individual in a susceptible population was developed in [3]. We use the same technique to compute the basic offspring number for model system (2 in absence of controls.

Let $\tilde{x} = (s, y, z)$ be the set of state variables. The system (2) can be rewritten as $\frac{d\tilde{x}_i}{dt} = F_i(\tilde{x}) - V_i(\tilde{x})$, where F_i is the rate of new recruits (birth of new colonizing females) in compartment i, $V_i = V_i^- - V_i^+$, where V_i^+ representing the rate of transfer into a compartment i by all other means, and V_i^- is the rate of transfer out the compartment i. For this model, F and V are given by

$$F = \begin{bmatrix} 0 \\ \phi z \\ 0 \end{bmatrix}; \qquad V = \begin{bmatrix} -\Lambda + \beta\frac{sy}{y+d} + \mu s \\ \varepsilon\beta\frac{sy}{y+d} + \mu_y y \\ -\varepsilon\beta\frac{sy}{y+d} + \mu_z z \end{bmatrix}.$$

To obtain the next generation matrix, we compute the Jacobian matrices of F and V denoted by $R = \mathcal{J}_F(\mathcal{E}^0)$ and $T = \mathcal{J}_V(\mathcal{E}^0)$. Here, we have

$$R = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \phi \\ 0 & 0 & 0 \end{pmatrix}; \qquad T = \begin{pmatrix} \mu & \beta\frac{s^0}{d} & 0 \\ 0 & \varepsilon\beta\frac{s^0}{d} + \mu_y & 0 \\ 0 & -\varepsilon\beta\frac{s^0}{d} & \mu_z \end{pmatrix}.$$

The basic offspring is obtained by computing the spectral radius of the next generation matrix RT^{-1}:

$$\mathcal{N} = \rho(RT^{-1}) = \frac{\varepsilon\phi\beta\frac{s^0}{d}}{\mu_z\left(\varepsilon\beta\frac{s^0}{d} + \mu_y\right)}. \tag{12}$$

The immediate consequence of the next generation method is that, the equilibrium $\mathcal{E}_0$ is locally asymptotically stable if $\mathcal{N} < 1$ and unstable otherwise.

The Jacobian matrix associated with system (2) at equilibrium point $\mathcal{E}^*$ is given by:

$$\mathcal{J} = \begin{pmatrix} -\beta\frac{y^*}{y^*+d} - \mu & -\beta\frac{s^*d}{(y^*+d)^2} & 0 \\ -\varepsilon\beta\frac{y^*}{y^*+d} & -\varepsilon\beta\frac{s^*d}{(y^*+d)^2} - \mu_y & \phi \\ \varepsilon\beta\frac{y^*}{y^*+d} & \varepsilon\beta\frac{s^*d}{(y^*+d)^2} & -\mu_z \end{pmatrix}.$$

The characteristic equation of which is

$$\lambda^3 + a_2\lambda^2 + a_1\lambda + a_0 = 0.$$

where

$$\begin{aligned} a_0 &= \mu_y\mu_z\beta\frac{y^*}{y^*+d} + \mu\mu_y\mu_z\left(1 - \frac{d}{y^*+d}\right); \\ a_1 &= \left(\mu + \frac{\beta y^*}{y^*+d}\right)(\mu_z + \mu_y) + \mu\varepsilon\frac{\beta s^* d}{(y^*+d)^2} + \mu_y\mu_z\left(1 - \frac{d}{y^*+d}\right); \\ a_2 &= \mu_z + \mu + \mu_y + \frac{\beta y^*}{y^*+d} + \frac{\varepsilon\beta s^* d}{(y^*+d)^2}. \end{aligned}$$

Since a_2, a_1 and a_0 are positive, The Routh-Hurwitz criterion for stability only imposes the $a_2a_1 - a_0 > 0$ needs to be positives, which can easily be shown. This implies that $\mathcal{E}^*$ exists and is asymptotically stable if and only if $\mathcal{N} > 1$.

Operating diagram of a flocculation model 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. The objective of this study is to analyze a model of the chemostat involving the attachment and detachment dynamics of planktonic and aggregated biomass in the presence of a single resource. Considering the mortality of species, we give a complete analysis for the existence and local stability of all steady states for general monotonic growth rates. Moreover, we determine the operating diagram which depicts the asymptotic behavior of the system with respect to control parameters. We show that the model exhibits a rich set of behaviors with a multiplicity of coexistence steady states, bi-stability, and occurrence of stable limit cycles.

RÉSUMÉ. L'objectif de cette étude est d'analyser un modèle du chémostat impliquant la dynamique d'attachement et de détachement de la biomasse planctonique et agrégée en présence d'une seule ressource. En considérant la mortalité des espèces, nous donnons une analyse complète de l'existence et de la stabilité locale de tous les équilibres pour des taux de croissance monotones. De plus, nous déterminons le diagramme opératoire qui décrit le comportement asymptotique du système par rapport aux paramètres de contrôle. Nous montrons que le modèle présente un ensemble riche de comportements avec multiplicité d'équilibres de coexistence, bi-stabilité et apparition des cycles limites stables.

KEYWORDS : Bi-stability, Chemostat, Flocculation, Limit cycles, Operating diagram

MOTS-CLÉS : Bi-stabilité, Chémostat, Floculation, Cycles limites, Diagramme opératoire

1. Introduction

In the culture of microorganisms, the processes of attachment and detachment of bacteria are well known and frequently observed. This phenomenon is manifested either by a fixation of microorganisms on a support as in the growth of biofilms or simply by an aggregation such as the formation of flocs or granules [9, 15]. In fact, the formation of flocs has a direct impact on growth dynamics, since the access to the substrate is limited for microorganisms within such structures. Nevertheless, it is only recently that they have been explicitly taken into account in mathematical models based on the chemostat (see the monograph [7]).

This flocculation mechanism may explain the coexistence of microbial species when the most competitive species inhibits its own growth by the formation of flocs [8]. In fact, these bacteria in flocs consume less substrate than planktonic bacteria since the attached bacteria have less access to the substrate, given that this access to the substrate is proportional to the outside surface of flocs. An extension of the model [8] has been studied in [4] when the growth rate of isolated bacteria of the most competitive species exhibits an inhibition. In this case, there may be coexistence around a stable limit cycle. The interested reader can refer to [3, 4] for a review of the different specific attachment and detachment rates used in the literature.

In this work, we consider the flocculation model of one species introduced in [3]. This model, which has been studied also in [2, 7, 11, 12], is written as follows:

$$\left\{\begin{array}{lcl} \dot{S} & = & D(S_{in}-S)-f(S)u-g(S)v \\ \dot{u} & = & [f(S)-D_u]u-a(u+v)u+bv \\ \dot{v} & = & [g(S)-D_v]v+a(u+v)u-bv \end{array}\right. \tag{1}$$

where $S(t)$ is the concentration of the substrate at time t; $u(t)$ and $v(t)$ are, respectively, the concentrations of planktonic and attached bacteria at time t; $f(S)$ and $g(S)$ represent, respectively, the growth rates of isolated and attached bacteria; D and S_{in} are, respectively, the dilution rate and the concentration of the substrate in the feed device; D_u and D_v represent, respectively, the disappearance rates of planktonic and attached bacteria.

We assume that isolated bacteria can aggregate with isolated bacteria or flocs to form new flocs with a rate $a(u+v)u$, where a is a positive constant, proportional to both the density of isolated bacteria u and the total biomass density $u+v$. Furthermore, the flocs can split and liberate isolated bacteria with rate bv, where b is a positive constant, proportional to their density v.

The study of this model (1) has been limited to the biologically interesting case $D_v \leq D_u \leq D$, where $D_u = \alpha D$ and $D_v = \beta D$, α and β belong to $[0,1]$ and represent, respectively, the fraction of planktonic and attached bacteria leaving the reactor. This case was proposed by [1] to model a reactor with biomass attached to the support or to decouple the residence time of solids and the hydraulic residence time $(1/D)$.

In this work, we study the model (1) where D_u and D_v can be modeled as in [10, 13] by:

$$D_u = \alpha D + m_u, \quad D_v = \beta D + m_v$$

where the non-negative parameters m_u and m_v representing mortality (or maintenance) rate are taken into consideration. Therefore, our study will not be restricted to the cases $D_v \leq D_u \leq D$, as in [2, 3, 7, 11, 12], and the cases $D < D_u$, $D < D_v$ or $D_u < D_v$, which are also of biological interest, will be investigated.

Thus, our main objective in this article is to give a complete analysis of model (1) and to study its operating diagram in order to illustrate the behavior of the system according to the control parameters D and S_{in}.

This paper is organized as follows. First, we present in Section 2 some general hypotheses about the growth functions of flocculation model (1). Then, we analyze the existence and the local stability of steady states according to the dilution rate and the disappearance rates of planktonic and attached bacteria. In Section 3, we present the operating diagram in order to show the regions of emergence of multiplicity of positive equilibria according to the control parameters. Finally, conclusions are drawn in the last Section 4.

2. Hypotheses and model analysis

We use the following general hypotheses for growth functions $f(S)$ and $g(S)$:

(H1) $f(0) = g(0) = 0$ and $f'(S) > 0$ and $g'(S) > 0$ for all $S > 0$.

(H1) $f(S) > g(S)$ for all $S > 0$.

Assumption (H1) means that the growth can take place if and only if the substrate is present. In addition, the growth rates of isolated and attached bacteria increase with the concentration of substrate. Assumption (H2) means that bacteria in flocs consume less substrate than isolated bacteria.

The following result shows that our model (1) preserves the biological meaning.

Proposition 2.1 *For any non-negative initial condition, the solutions of system (1) remain non-negative and positively bounded. In addition, the set*

$$\Omega = \left\{ (S,u,v) \in \mathbb{R}^3_+ : S + u + v \leq \frac{D}{D_{\min}} S_{in} \right\}, \quad \textit{where} \quad D_{\min} = \min(D, D_u, D_v),$$

is positively invariant and is a global attractor for the dynamics (1).

The proofs of all results of this section are detailed in [5]. In the following, we use the following notations:

$$\varphi(S) = f(S) - D_u \quad \text{and} \quad \psi(S) = g(S) - D_v, \tag{2}$$

$$U(S) := \frac{\varphi(S)\,(\psi(S) - b)}{a\,[\psi(S) - \varphi(S)]} \quad \text{and} \quad V(S) := -\frac{\varphi^2(S)\,(\psi(S) - b)}{a\,[\psi(S) - \varphi(S)]\,\psi(S)}, \tag{3}$$

$$H(S) := f(S)U(S) + g(S)V(S). \tag{4}$$

From (H1), when equations $f(S) = D_u$, $g(S) = D_v$ and $\psi(S) = b$ have solutions, they are unique and we define the usual *break-even concentrations*

$$\lambda_u = f^{-1}(D_u), \quad \lambda_v = g^{-1}(D_v) \quad \text{and} \quad \lambda_b = \psi^{-1}(b).$$

From (H2), if in addition $D_v \geq D_u$, then $\lambda_v > \lambda_u$. When equations $f(S) = D_u$ or $g(S) = D_v$ or $\psi(S) = b$ have no solution, we put $\lambda_u = \infty$ or $\lambda_v = \infty$ or $\lambda_b = \infty$.

2.1. Existence of steady states

In order to study the existence of equilibria of model (1), we define the interval I by:

$$I = \begin{cases}]\lambda_u, \lambda_v[& \text{if} \quad \lambda_u < \lambda_v \\ ]\lambda_v, \min(\lambda_u, \lambda_b)[& \text{if} \quad \lambda_u > \lambda_v. \end{cases} \tag{5}$$

We can state the following result:

Lemme 2.1 *Under the assumptions (H1-H2), system (1) has the following steady states:*

1) the washout $E_0 = (S_{in}, 0, 0)$, that always exists,

2) a positive steady state, $E_1 = (S^*, u^*, v^*)$ with S^* solution of equation

$$D(S_{in} - S^*) = H(S^*) \tag{6}$$

where H is given by (4), $u^* = U(S^*)$ and $v^* = V(S^*)$, where U and V are given by (3). This coexistence steady state exists if and only if $S^* \in I$ where I is defined by (5).

The following proposition presents the number of positive steady states of (1).

Proposition 2.2

– When $D_u \leq D_v$, then the positive steady state $E_1 = (S^*, u^*, v^*)$ exists if and only if $S_{in} > \lambda_u$. If it exists, it is unique.

– When $D_u > D_v$, then there exists at least one positive steady state in the case $\lambda_u < \min(\lambda_v, S_{in})$ or $\lambda_v < \min(\lambda_u, \lambda_b) < S_{in}$. Generically, the system can have generically an odd number of positive steady states. When $S_{in} < \min(\lambda_u, \lambda_b)$ and $\lambda_v < \lambda_u$, then generically the system has no positive steady state or an even number of positive steady states.

2.2. Stability of steady states

In this section, we study the local asymptotic stability of each steady state of system (1). Let J be the Jacobian matrix of (1) at (S, u, v), that is given by

$$J = \begin{bmatrix} -D - f'(S)u - g'(S)v & -f(S) & -g(S) \\ f'(S)u & \varphi(S) - a(2u + v) & -au + b \\ g'(S)v & a(2u + v) & \psi(S) + au - b \end{bmatrix}. \tag{7}$$

The stability of the washout steady state is given as follows:

Proposition 2.3 *E_0 is Locally Exponentially Stable (LES) if and only if $S_{in} < \lambda_u$ and $S_{in} < \lambda_b$.*

In the following, we analyze the stability of positive steady states. At $E_1 = (S^*, u^*, v^*)$, the Jacobian matrix is given by

$$J_1 = \begin{bmatrix} -m_{11} & -m_{12} & -m_{13} \\ m_{21} & -m_{22} & a_{23} \\ m_{31} & m_{32} & -m_{33} \end{bmatrix}$$

where

$$\begin{cases} m_{11} = D + f'(S^*)u^* + g'(S^*)v^*, \quad m_{12} = f(S^*), \quad m_{13} = g(S^*), \\ m_{21} = f'(S^*)u^*, \quad m_{22} = a(2u^* + v^*) - \varphi(S^*), \quad a_{23} = b - au^*, \\ m_{31} = g'(S^*)v^*, \quad m_{32} = a(2u^* + v^*) \quad \text{and} \quad m_{33} = b - au^* - \psi(S^*). \end{cases} \tag{8}$$

The characteristic polynomial is given by

$$P(\lambda) = \lambda^3 + c_1\lambda^2 + c_2\lambda + c_3,$$

$$\begin{gathered} c_1 = m_{11} + m_{22} + m_{33}, \\ c_2 = m_{12}m_{21} + m_{13}m_{31} - m_{32}a_{23} + m_{11}m_{22} + m_{11}m_{33} + m_{22}m_{33}, \\ c_3 = m_{11}(m_{22}m_{33} - m_{32}a_{23}) + m_{21}(m_{12}m_{33} + m_{32}m_{13}) + m_{31}(m_{12}a_{23} + m_{13}m_{22}). \end{gathered} \tag{9}$$

According to Routh–Hurwitz criterion, E_1 is LES if and only if

$$c_1 > 0, \quad c_3 > 0 \quad \text{and} \quad c_4 = c_1c_2 - c_3 > 0. \tag{10}$$

We have the following results:

Lemme 2.2 *All m_{ij} are positive for all $i, j = 1, \ldots, 3$ with $(i, j) \neq (2, 3)$ and we have $c_1 > 0$.*

The next lemma shows that the sign of c_3 is given by the position of the curve of the function $H(\cdot)$ with respect to the line of equation $y = D(S_{in} - S)$. More precisely, we give the link between the determinant of the Jacobian matrix J_1 at $E_1 = (S^*, u^*, v^*)$ and $D + H'(S^*)$.

Proposition 2.4 *One has $c_3 = -\det(J_1) = -\varphi(S^*)(\psi(S^*) - b)(D + H'(S^*))$.*

Since the condition $c_4 > 0$ of the Routh–Hurwitz criterion (10) could be unfulfilled, we will study the behavior of flocculation model (1) according to the dilution rate and the disappearance rates of planktonic and attached bacteria. In fact, there exist four cases that must be distinguished:

$$\begin{array}{ll} \text{Case 1:} \quad D_u \leq D_v \leq D, & \text{Case 2:} \quad D_v < D_u \leq D, \\ \text{Case 3:} \quad D_v < D_u \text{ and } D < D_u, & \text{Case 4:} \quad D_u \leq D_v \text{ and } D < D_v. \end{array} \tag{11}$$

To determine the local stability of the positive steady state in the first and second cases of (11), we will have need of the following.

Proposition 2.5 *In the cases 1 and 2 ($D_u \leq D$ and $D_v \leq D$), we have $c_4 > 0$.*

It was shown in [7], see also [11, 12] that if $D_u = D_v = D$ then the positive steady E_1 exists and is unique and LES if and only if $S_{in} > \lambda_u$. Actually, this result holds in case 1.

Proposition 2.6 *In the case 1 ($D_u \leq D_v \leq D$), the positive steady state $E_1 = (S^*, u^*, v^*)$ exists if and only if $S_{in} > \lambda_u$. If it exists, it is unique and LES.*

The case 2 was solved in [2] where it was shown that the stability depends only on the relative position of the curve of function $y = H(S)$ and the straight line of equation $y = D(S_{in} - S)$ that is to say, on the sign of $D + H'(S^*)$. More precisely, we have:

Proposition 2.7 *Let $E_1 = (S^*, u^*, v^*)$ be a positive steady state. Assume that case 2 holds.*

1) If $\lambda_u < \lambda_v$: E_1 is LES if $H'(S^*) > -D$ and is unstable if $H'(S^*) < -D$.

2) If $\lambda_u > \lambda_v$: E_1 is LES if $H'(S^*) < -D$ and is unstable if $H'(S^*) > -D$.

In the case 3 of (11), when

$$D < D_v \leq D_u \quad \text{or} \quad D_v < D \leq D_u$$

c_4 can change sign by varying the control parameter S_{in} such that the positive steady state E_1 could change its behavior without any collision with another steady state [5]. In fact, numerical simulations show the emergence of stable limit cycles by Hopf bifurcations.

In the case 4 of (11), we always have $\lambda_u < \lambda_v$ and $H'(S) > 0$. Therefore, from Prop. 2.4, it is deduced that in the case 4 of (11) we always have $c_3 > 0$. We were not able to find a set of parameters for which $c_4 < 0$, as in the case 3 of (11) and we conjecture that in this case the positive steady state E_1 which is unique as soon as it exists, is also LES as soon as it exists.

3. Operating diagram

The operating diagram shows how the system behaves when we vary the two control parameters S_{in} and D in (1). All other parameters in (1) are fixed, such as growth functions and specific attachment and detachment velocities. In fact, they depend on the nature of the organisms and the substrate introduced into the chemostat. Note that this operating diagram has not been studied in the existing literature in the generic case where the disappearance rates are distinct.

If $m_u \geq f(+\infty)$ then equation

$$f(S) = \alpha D + m_u \tag{12}$$

has no solution. We assume that $m_u < f(+\infty)$. The equation (12) is equivalent to $D = \tilde{f}(S) := \frac{f(S)-m_u}{\alpha}$. Since f is increasing, then there exists a unique increasing function

$$\begin{array}{rccl} F_0 : [0, & \bar{D}_u[& \longrightarrow & [f^{-1}(m_u), +\infty[\\ & D & \longrightarrow & F_0(D) = \tilde{f}^{-1}(D) \end{array}$$

solution of equation (12) where $\bar{D}_u = \frac{f(+\infty)-m_u}{\alpha}$. Note that if $D \geq \bar{D}_u$, then equation (12) has no solution and we put $F_0(D) = +\infty$. If $m_v \geq g(+\infty)$ then equation

$$g(S) = \beta D + m_v \tag{13}$$

has no solution. We assume that $m_v < g(+\infty)$. The equation (13) is equivalent to $D = \tilde{g}(S) := \frac{g(S)-m_v}{\beta}$. Since g is increasing, then there exists a unique increasing function

$$\begin{array}{rccl} F_1 : [0, & \bar{D}_v[& \longrightarrow & [g^{-1}(m_v), +\infty[\\ & D & \longrightarrow & F_1(D) = \tilde{g}^{-1}(D) \end{array}$$

solution of equation (13) where $\bar{D}_v = \frac{g(+\infty)-m_v}{\beta}$. Note that if $D \geq \bar{D}_v$, then equation (13) has no solution and we put $F_1(D) = +\infty$. If $m_v + b \geq g(+\infty)$ then equation

$$g(S) = \beta D + m_v + b \tag{14}$$

has no solution. We assume that $m_v + b < g(+\infty)$. The equation (14) is equivalent to $D = \tilde{g}_b(S) := \frac{g(S)-m_v-b}{\beta}$. Since g is increasing, then there exists a unique increasing function

$$\begin{array}{rccl} F_2 : [0, & \bar{D}_b[& \longrightarrow & [g^{-1}(m_v + b), +\infty[\\ & D & \longrightarrow & F_2(D) = \tilde{g}_b^{-1}(D) \end{array}$$

solution of equation (14) where $\bar{D}_b = \frac{g(+\infty)-m_v-b}{\beta}$. Note that if $D \geq \bar{D}_b$, then equation (14) has no solution and we put $F_2(D) = +\infty$.

In the following, we show the emergence of the bi-stability region with multiplicity of positive steady states in the case

$$F_1(D) < F_0(D) < F_2(D) \quad \text{for all} \quad D \in [0, \min(\bar{D}_u, \bar{D}_v, \bar{D}_b)[.$$

In this case, the function H is defined and decreasing on the interval $I =]\lambda_v, \lambda_u[$. It vanishes at λ_u and tends to infinity as S tends to λ_v (see Fig. 1(c)). Assume that H is convex. Thus, equation $H'(S) = -D$ has a unique solution $\tilde{S}(D) \in I =]\lambda_v, \lambda_u[$ if and only if $H'(\lambda_u)+D > 0$, or also $D > \bar{D}$ with $\bar{D}$ solution of equation $H'(F_0(D))+D = 0$. More precisely, since the function H' is increasing, then there exists a unique decreasing function

$$\begin{array}{rccl} \tilde{S} : [\bar{D}, & \bar{D}_v[& \longrightarrow &]\lambda_v, \lambda_u[\\ & D & \longrightarrow & \tilde{S}(D) = \tilde{H}^{-1}(D) \end{array}$$

solution of equation $H'(S) = -D$ with $\tilde{H}(S) = -H'(S)$. Thus, we define the curve Γ_3 of equation

$$S_{in} = F_3(D) := \frac{1}{D} H(\tilde{S}(D)) + \tilde{S}(D)$$

which corresponds to the saddle-node bifurcation with the appearance of two positive steady states. In order to illustrate the operating diagram, we considered the parameter values provided in Table 2 with the growth rates f and g of Monod-type:

$$f(S) = \frac{m_1 S}{k_1 + S} \quad \text{and} \quad g(S) = \frac{m_2 S}{k_2 + S}, \tag{15}$$

where m_i denotes the maximum growth rate and k_i the Michaelis-Menten constant, $i = 1, 2$. Table 1 shows the existence and local stability of steady states E_0, E_1 and E_2 in the regions $\mathcal{I}_k$, $k = 0, 1, 2$, of the operating diagram shown in figure 1 (b). The letter S (resp. U) means stable (resp. unstable). Absence of letter means that the corresponding steady state does not exist. Let Γ_i, $i = 0, \ldots, 3$, be the respective curves of equations

Condition	Region	E_0	E_1	E_2
$S_{in} < F_3(D)$	$(S_{in}, D) \in \mathcal{I}_0$	S		
$F_3(D) < S_{in} < F_0(D)$	$(S_{in}, D) \in \mathcal{I}_1$	S	S	I
$F_0(D) < S_{in}$	$(S_{in}, D) \in \mathcal{I}_2$	I	S	

Table 1. *Existence and local stability of steady states according to the regions in the operating diagram of figure 1.*

$S_{in} = F_i(D)$ (see Fig. 1(a)). Γ_0 and Γ_3 separate the operative plan (D, S_{in}) at most in three regions, denoted $\mathcal{I}_k$, $k = 0, 1, 2$ (see Fig. 1(b)). The transition from the region $\mathcal{I}_0$ to the region $\mathcal{I}_1$ by the curve Γ_3 (in magenta) corresponds to a saddle-node bifurcation with the appearance of two positive equilibria E_1 which is LES and E_2 which is unstable. The transition from the region $\mathcal{I}_1$ to the region $\mathcal{I}_2$ by the curve Γ_0 (in red) corresponds to a transcritical bifurcation when the unstable steady state E_2 disappears and E_0 becomes unstable.

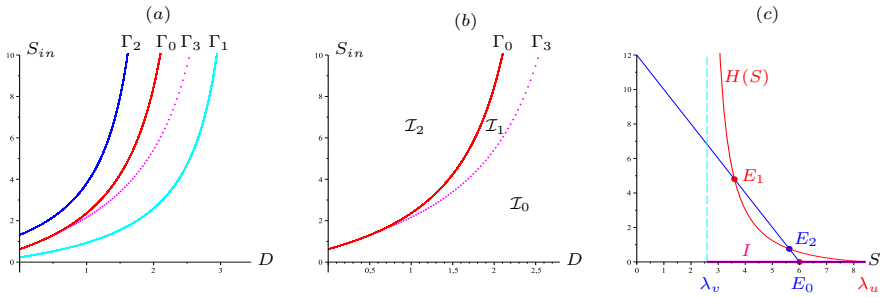


Figure 1. *(a) The case* $F_1(D) < F_0(D) < F_2(D)$. *(b) Operating diagram of (15). (c) Bi-stability and multiplicity of positive steady states when* $(D, S_{in}) = (2, 6) \in \mathcal{I}_1$.

For this set of parameters mentioned in Table 2, the numerical simulations show that the condition $c_4 > 0$ of the Routh–Hurwitz criterion is satisfied in the region $\mathcal{I}_1$, that is, the steady state E_1 is LES as long as it exists. However, this condition may not be satisfied for another set of parameters where the positive steady state can change behavior by a Hopf bifurcation with the emergence of stable limit cycle. In this case, the analysis of the operating diagram is the subject of on-going investigations.

4. Conclusion

In this work, we have analyzed mathematically and through numerical simulations a model of the chemostat where one species is present in two forms, isolated and attached with the presence of a single resource. The new feature was that maintenance terms are added to removal rates in order to give a complete analysis of the flocculation model (1). The operating diagram shows the occurrence of the bi-stability region with multiplicity of coexistence steady states that can bifurcate through saddle-node bifurcations or trans-critical bifurcations. However, the bi-stability could occur in the classic chemostat model [14] only when the growth rate is non-monotonic.

Acknowledgments. We thank the financial support of the PHC UTIQUE project No. 13G1120 and of the Euro-Mediterranean research network TREASURE (`http://www.inra.fr/treasure`)

A. Parameters used in numerical simulations

Parameter	m_1	k_1	m_2	k_2	a	b	α	β	m_u	m_v
Figure 1	3.5	2.5	3	1.5	1	1	1	0.75	0.7	0.4

Table 2. *Parameter values used for (1) when the growth rates f and g are given by (15).*

B. References

[1] O. BERNARD, Z. HADJ-SADOK, D. DOCHAIN, A. GENOVESI, J-P. STEYER, "Dynamical model development and parameter identification for an anaerobic wastewater treatment process", *Biotechnol. Bioeng.*, vol. 75, 2001, 424–438.

[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, A. RAPAPORT, T. SARI, "Emergence of coexistence and limit cycles in the chemostat model with flocculation for a general class of functional responses", *Appl. Math. Modell.*, vol. 40, 2016, 7656–7677.

[5] R. FEKIH-SALEM, T. SARI, "Properties of the chemostat model with aggregated biomass and distinct dilution rates", *Preprints Siam SIADS*, 2018, (`https://hal.inria.fr/hal-01722448v1`.

[6] 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.

[7] J. HARMAND, C. LOBRY, A. RAPAPORT, T. SARI, "The Chemostat: Mathematical Theory of Microorganism Cultures", *Wiley, Chemical Engineering Series, Chemostat and Bioprocesses Set*, 2017.

[8] B. HAEGEMAN, A. RAPAPORT, "How flocculation can explain coexistence in the chemostat", *J. Biol. Dyn.*, vol. 2, 2008, 1–13.

[9] IWA TASK GROUP ON BIOFILM MODELING, "Mathematical modeling of biofilms", *IWA publishing*, 2006.

[10] S. MARSILI-LIBELLI, S. BENI, "Shock load modelling in the anaerobic digestion process", *Ecol. Model.*, vol. 84, 1996, 215–232.

[11] A. RAPAPORT, P. WALTMAN, "Properties of the chemostat model with aggregated biomass", *to appear in Euro. J. of Appl. Math.*, 2018.

[12] T. SARI, R. FEKIH-SALEM, "Analysis of a model of flocculation in the chemostatguilf, *Proceedings of the 8th conference on Trends in Applied Mathematics in Tunisia, Algeria, Morocco*, 2017, 75–80.

[13] S. SHEN, G. C. PREMIER, A. GUWY, R. DINSDALE, "Bifurcation and stability analysis of an anaerobic digestion model", *Nonlinear Dyn.*, vol. 48, 2007, 391–408.

[14] H.L. SMITH, P. WALTMAN, "The Theory of the Chemostat: Dynamics of Microbial Competition", *Cambridge University Press*, 1995.

[15] D.N. THOMAS, S.J. JUDD, N. FAWCETT, "Flocculation modelling: a review", *Water Res.*, vol. 33, 1999, 1579–1592.

How do variations in water levels affect Predator-prey interactions

K. Belkhodja, A. Moussaoui

Department of Mathematics
University of Tlemcen
Tlemcen
Algeria
moussaouidz@yahoo.fr

RÉSUMÉ. Le niveau d'eau dans les rivières, les lacs et les réservoirs a une grande influence sur les interactions entre les proies et les prédateurs. En effet, l'augmentation du volume d'eau réduit la capture de la proie par le prédateur. Le même raisonnement s'applique lorsqu'il y a une diminution du volume d'eau, favorisant la capture de la proie par le prédateur. L'objectif de cet article est d'étudier les propriétés dynamiques d'un modèle prédateur-proie avec des prélèvements à taux constant non nul et soumis à des fluctuations du niveau d'eau dans un lac. Cette étude est importante pour comprendre le comportement et la dépendance des espèces à une variation saisonnière du niveau de l'eau. Des conditions ont été établies pour la coexistence et l'extinction des espèces. Les principaux résultats ont été illustrés par des simulations numériques. Les résultats de cette étude démontrent comment les variations du niveau d'eau peuvent affecter la répartition des espèces de poissons

ABSTRACT. Water level in rivers, lakes and reservoirs has great influence on the interactions between prey and predator fish. Indeed, the increase of the water volume hinders the capture of the prey by the predator. The same reasoning applies when there is a decrease in the volume of water, favoring the capture of the prey by the predator. The objective of this paper is to study the dynamical properties of a predator-prey model with nonzero constant rate prey harvesting and subject to fluctuating water level in a lake. This investigation is important to understand the behavior and dependence of species on a seasonal variation of water level. Conditions have been derived for for the coexistence and extinction of species. Main results have been illustrated using numerical simulations. The results of this study demonstrate how water level variations can affect the distribution of fish species

MOTS-CLÉS : Modèle prédateur-proie, stabilité, lac de Pareloup

KEYWORDS : Predator-prey model, stability, Pareloup lac

1. Introduction.

Lac de Pareloup is a lake in Aveyron, France. It lies on the Lévézou plateau, 25 km south east of Rodez. This is the fifth largest hydroelectric reservoir by area in France having an area of 1260 hectares. Two interdependent fish species account as the most important species living this lake, They are the Roach species as prey (Gardon in French) and Pike species as predator (Brochet in French). The water level of Pareloup lake is regulated, mainly for hydroelectric purposes. The water level is lowered by increasing discharge in winter, when the consumption of electricity is highest. In the spring, snow melts refilling the lakes with the aid of the reduced discharge and the water level is usually kept quite constant over the summer until late summer. The management of this lake is of considerable ecological importance. Significant variations of the water level of the lake can have a strong impact on the persistence of some species. Indeed, when the water level is low, in winter, the contact between the predator and the prey is more frequent, and the predation increases. Conversely, when the water level is high, in the spring, its more difficult for the predator to find a prey and the predation decreases. In [4], the authors examine how seasonal variations in water level affect the outcome of a predator-prey interactions in Pareloup Lake. More recently, in [6] the authors assume that both species are subjected to harvesting and discuss the effects of water level and harvesting on the survival of the two species. All these studies demonstrate that the dynamics of the systems depends heavily on the fluctuation of the water level and give some valuable suggestions for saving the species and regulating populations when the ecological and environmental parameters are affected by periodic factors.

In this paper, we assume that the predator is not of commercial importance. The prey is continuously being harvested at a constant rate by a harvesting agency. The harvesting activity does not affect the predator population directly. It is obvious that the harvesting activity does reduce the predator population indirectly by reducing the availability of the prey to the predator. Let $G(t)$ and $B(t)$ are respectively the densities of the prey and predator at time t. We make the following assumptions :

(A1) In the absence of predator, prey growing logistically with a growth rate γ_G.

(A2) In the absence of prey, predator population declines exponentially.

(A3) The predator need a quantity γ_B for his food, but he has access to a quantity of food depending on the water level equal

$$\frac{r}{H}\frac{G}{B+D},$$

where r is a positive constant, D measures the other causes of mortality outside the metabolism and predation and H is the water level of the lake. The minimum value of H is reached in autumn and the maximum value is attained during the spring. If

$$\frac{r}{H}\frac{G}{B+D} \geq \gamma_B,$$

then the predator will be satisfied with the quantity γ_B for his food. Otherwise, i.e if

$$\frac{r}{H}\frac{G}{B+D} \leq \gamma_B,$$

the predator will content himself with

$$\frac{r}{H}\frac{G}{B+D}.$$

Consequently, the quantity of food received by the predator is

$$\min\Big(\frac{r}{H}\frac{G}{B+D},\gamma_B\Big).$$

Considering the above basic assumptions we can now write the following dynamical system :

$$\begin{cases} \dfrac{dG}{dt} = G(t)\left(\gamma_G - m_G G(t)\right) - \min\left(\dfrac{r}{H}\dfrac{G(t)}{B(t)+D},\gamma_B\right)B(t) - Q, \\ \dfrac{dB}{dt} = -m_B B(t) + \min\left(\dfrac{r}{H}\dfrac{G(t)}{B(t)+D},\gamma_B\right)B(t). \end{cases} \quad [1]$$

where e is the conversion rate and Q represents the rate of harvesting ($Q > 0$).
The objective of this paper is to study the dynamical properties of the predator-prey model with constant harvesting. It will be better for us to determine how the water level and constant harvesting affect the dynamics of system (1).
Let $B_0,\ G_0$ be respectively the initial density of the predator and prey with $B_0 > 0$ and $G_0 > 0$. We denote by

$$H_0 = \max\left(\frac{r}{\gamma_B}\frac{G_0}{(B_0+D)}, \frac{r\left(\left(\gamma_G+m_B\right)^2 - 4m_G Q\right)}{4m_G m_B \gamma_B D}\right),$$
$$H_1 = \frac{er}{2m_G m_B D}\Big(\gamma_G - \sqrt{\gamma_G^2 - 4m_G Q}\Big),$$
$$H_2 = \frac{er}{2m_G m_B D}\Big(\gamma_G + \sqrt{\gamma_G^2 - 4m_G Q}\Big).$$

and we assume :

$$\frac{\gamma_G^2}{2m_G}\Big(\frac{B_0+D}{B_0}\Big) < Q < min\Big(\frac{m_B D}{2e}, \frac{\gamma_G^2}{4m_G}\Big), \quad [2]$$

$$H > H_0 \quad [3]$$

2. Mathematical analysis and main result

Proposition 1 *All the solutions of system (1) which initiate in R_+^2 are uniformly bounded.*

Proof. See Appendix A.

To simplify our analysis, we rewrite system (1) in a simpler form. We prove the following result.

Proposition 2 *Under hypothesis (3), we have* $\dfrac{r}{H}G(t) < \gamma_B(B(t)+D),\ \forall t \geq 0.$

Proof. See Appendix B.

Consequently system (1) is reduced to the simple form

$$\begin{cases} \dfrac{dG}{dt} = G(t)\left(\gamma_G - m_G G(t)\right) - \dfrac{r}{H}\dfrac{G(t)B(t)}{B(t)+D} - Q, \\ \dfrac{dB}{dt} = e\dfrac{r}{H}\dfrac{G(t)B(t)}{B(t)+D} - m_B B(t). \end{cases} \quad [4]$$

3. Local stability analysis of the steady states.

We now explore the existence and stability of boundary and positive equilibria of system (4)

Proposition 3 *System (4) has the following equilibria :*

– $P_1 = (G_1, 0)$, *where* $G_1 = \dfrac{\gamma_G - \sqrt{\gamma_G^2 - 4m_G Q}}{2m_G}$.

– $P_2 = (G_2, 0)$, *where* $G_2 = \dfrac{\gamma_G + \sqrt{\gamma_G^2 - 4m_G Q}}{2m_G}$.

and an interior equilibrium $P^* = (G^*, B^*)$*, where*

$$G^* = \frac{\left(\gamma_G - \frac{r}{H}\right) + \sqrt{\left(\gamma_G - \frac{r}{H}\right)^2 + 4m_G\left(\frac{m_B D}{e} - Q\right)}}{2m_G},\ B^* = \frac{er}{m_B H}G^* - D.$$

It is easy to see that a necessary and sufficient condition for the existence of the interior equilibrium P^* *is :*

$$H_1 < H < H_2. \quad [5]$$

3.1. Stability Analysis

Now we study the nature of these equilibria. The Jacobian matrix associated to (4) is given by

$$J(G,B) = \begin{pmatrix} \gamma_G - 2m_G G - \dfrac{r}{H}\dfrac{B}{B+D} & \dfrac{-r}{H}\dfrac{GD}{(B+D)^2} \\ \dfrac{er}{H}\dfrac{B}{B+D} & -m_B + \dfrac{er}{H}\dfrac{GD}{(B+D)^2} \end{pmatrix}.$$

We obtain the following results

Proposition 4 • *The axial equilibrium point* P_1 *is always unstable.*

• *The axial equilibrium point* P_2 *is stable if* $H > H_2$ *otherwise, it is a saddle point.*

The proof is trivial and we omit it.

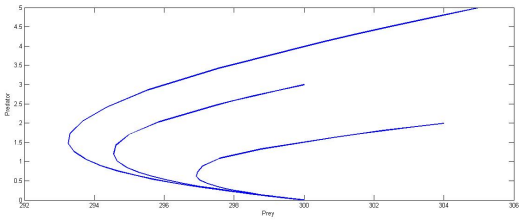


Figure 1. *Extinction of the predator when* $H > H_2 = 40$. *The parameter values are :* $\gamma_G = 7; m_G = 0.02; m_B = 7.5; e = 0.2; r = 50; D = 10; H = 60; Q = 300$.

Proposition 5 *If condition (5) holds, and*

$$H_* < H_1 < H_2,$$

where $H_* = \dfrac{r}{\gamma_G}$, *then the coexistence equilibrium* P^* *when it exists, it is locally asymptotically stable.*

Proof. See Appendix C.

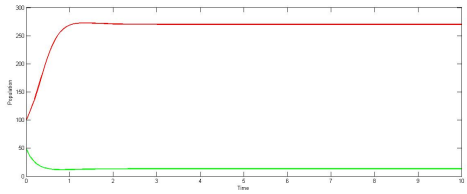


Figure 2. *The densities of each species plotted against time when* $H_1 = 2.85 < H < H_2 = 8.93$. *The figure demonstrates the stability of the system (4) around the equilibrium* $(270.25, 13.24)$. *The parameter values are :* $\gamma_G = 7; m_G = 0.02; m_B = 5; e = 0.2; r = 20; D = 30; H = 5; Q = 100$.

Proposition 6 *If condition (5) holds, and if*

$$H_1 < H_* < H_2,$$

then there exists $\tilde{H}$ *such that,*

$$H_1 < \tilde{H} < H_2,$$

and

- *when* $H_1 < H < \tilde{H}$, P^* *is unstable.*
- *when* $\tilde{H} < H < H_2$, P^* *is locally asymptotically stable.*

Proof. See Appendix D.

4. Global stability of P^*

To investigate the global behavior of system (4) we first prove that system (4) around P^* has no nontrivial periodic solutions. The proof is based on an application of a divergence criterion [3]. Let $F(G,B) = \frac{1}{GB}$ Obviously $F(G,B) > 0$ if $G > 0$, $B > 0$
We define :

$$f_1(G,B) = G(t)\left(\gamma_G - m_G G(t)\right) - \frac{r}{H}\frac{G(t)B(t)}{B(t)+D} - Q,$$

$$f_2(G,B) = e\frac{r}{H}\frac{G(t)B(t)}{B(t)+D} - m_B B(t),$$

and

$$\Delta(G,B) = \frac{\partial(Ff_1)}{\partial G} + \frac{\partial(Ff_2)}{\partial B}$$

We find that

$$\Delta(G,B) = \frac{1}{G^2 B}\left[-m_G G^2 + Q - \frac{er}{H}\frac{G^2 B}{(B+D)^2}\right] = \frac{1}{G^2 B} Tr J^*,$$

which is less than zero when the interior equilibrium is locally stable for all $G > 0$, $B > 0$. As the solution is bounded, then by Bendixson-Dulac criterion, there will be no limit cycle in the first quadrant.

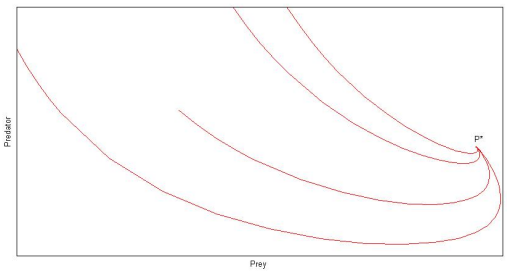


Figure 3. *phase space trajectories corresponding to different initial levels.*

Now, we are in a position to prove the following theorem.

Proposition 7 *Existence and local stability of a positive interior equilibrium ensure that system (4) around P^* is globally asymptotically stable .*

Proof. The proof is based on the following arguments :
(a)-System (4) is bounded.
(b)-The axial equilibrium P_1 is always an unstable saddle point and existence of positive equilibrium confirms that the axial equilibrium P_2 is also an unstable saddle point.
(c)-Positive equilibrium P^* is LAS when $H_1 < H < H_2$.
(d)-System (4) around P^* has no non-trivial periodic solutions.

Biological Implications :
Proposition 4 implies that Model (4) can have the-only prey population being locally stable at its equilibrium P_2 when the water level is high (Figure 1).

Proposition 5 implies that Model (4) can coexist at the equilibrium P^* if the water level is between two thresholds (Figures 2 and 3).

5. Existence of cycle limit

Proposition 8 *If $H_1 < H < \tilde{H}$, then system (4) has at least one limit cycle.*

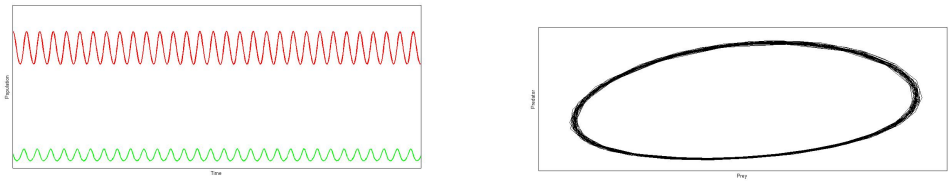


Figure 4. *There is a limit cycle arrounding the unstable interior equilibrium point.*

Proof. We have shown that all solutions are bounded and if $H_1 < H < \tilde{H}$, there are no asymptotically stable equilibrium point, we can deduce by Poincaré-Bendixson theorem [3] that there exists at least one periodic orbit (Figure 4).

6. Extinction of species

In this section, we prove a result on the extinction of the prey.

Proposition 9 *If $H < H_1$, then the population of prey disappears.*

Proof. From the first equation of system (4), we have

$$\frac{dG}{dt} = G(t)\left(\gamma_G - m_G G(t)\right) - \frac{r}{H}\frac{G(t)B(t)}{B(t)+D} - Q \leq G(t)\left(\gamma_G - m_G G(t)\right) - \frac{r}{H_1}\frac{G(t)B_0}{B_0+D} - Q$$

Hence

$$\frac{dG}{dt} \leq -m_G G^2 + \left(\gamma_G - \frac{2m_G Q}{\gamma_G}\frac{B_0}{B_0+D}\right)G - Q \quad .$$

Using condition (2) we get

$$\frac{dG}{dt} < 0,$$

and this leads to the extinction of prey and subsequently that of predators.

7. Conclusion.

Based on the results of this work, it can be concluded that changes in water level have an impact on the distribution of species. By making some assumptions about biological parameters, we have reduced our model to a simple form. The boundedness of the system is established, which, in turn, implies that the system is biologically well posed. The mathematical analysis presented here, shows that if H is below the level H_1, we will have the extinction of the species and beyond H_2, we will have the extinction of the predators. It remains the level between H_1 and H_2. Here we found two cases, the first when the level H_* is below H_1, in this case the interior equilibrium point if it exist, it is locally asymptotically stable. The second case is when the H_* is between H_1 and H_2, in which case we have shown numerically that there exists a $\tilde{H}$ which changes the sign of the trace and hence the nature of P^* will change. Indeed, below $\tilde{H}$ the trace is positive and P^* is unstable. Using Poincaré Bendixon's theorem we have proved the existence of at least one limit cycle around P^*. Above $\tilde{H}$, the trace becomes negative and therefore P^* is stable.

8. Bibliographie

G. BIRKHOFF AND G. C. ROTA, « Ordinary differential Equations », *Ginn Boston*, 1982.

H. COOPS, M. BEKLIOGLU AND T.L. CRISMAN, « The role of water-level fluctuations in shallow lake ecosystems workshop conclusions ». *Hydrobiologia*, 23-27, 2003

J, HALE, « Theory of Functional Differential Equation with applications in Population Dynamics ». *Academie Press, New York*, 1993.

N. CHIBOUB FELLAH, S.M. BOUGUIMA, A. MOUSSAOUI, « The effect of water level in a prey-predator interaction : A nonlinear analysis study, *Chaos, Solitons and Fractals* ». 45, 205-212, 2012.

A.MOUSSAOUI, « A reaction-diffusion equations modelling the effect of fluctuating water levels on prey-predator interactions ». *Appl Math Comput* ; 268, 1110-1121, 2015.

M.A. MENOUER, A. MOUSSAOUI, « Effects of consecutive water level fluctuations and harvesting on predator-prey interactions ». *Chaos, Solitons and Fractals*, 91, 434-442, 2016.

J.H. WLOSINSKI, E.R. KOLJORD, « Effects of Water Levels on Ecosystems, an Annotated Bibliography, Long Term Resource Monitoring Program ». *Technical Report 96-T007*, 1996.

Appendice A.

Proof of Proposition 1
We define a function

$$w = eG + B. \qquad [6]$$

The time derivative of (6) along the solutions of (1) is

$$\frac{dw}{dt} = e\frac{dG}{dt} + \frac{dB}{dt} = eG(\gamma_G - m_G G(t)) - Q - m_B B$$

then

$$\frac{dw}{dt} + m_B w \leq eG(\gamma_G + m_B - m_G G(t)),$$

This implies that

$$\frac{dw}{dt} + m_B w \leq \mu,$$

where $\mu = \dfrac{(\gamma_G + m_B)^2}{4m_G}$.
Applying the theory of differential inequalities [1], we obtain

$$0 \leq w(G,B) \leq \frac{\mu}{m_B} + e^{-m_B t}\left\{w(G(0),B(0)) - \frac{\mu}{m_B}\right\},$$

and for $t \to \infty$, we have $0 \leq w \leq \dfrac{\mu}{m_B}$.

Hence all the solutions of (1) which initiate in R^2_+ are eventually confined in the region :

$$B = \left\{(G,B) \in R^+_2 : w = \frac{\mu}{m_B} + \varepsilon, \forall \varepsilon > 0\right\}.$$

Appendix B

Proof of Proposition 2
Let

$$u(t) = \frac{r}{H}G(t) - \gamma_B(B(t) + D).$$

Note that $u(0) < 0$ by condition (3). It is claimed that $u(t) < 0$ for all t. If this were not the case, there exists $t_0 > 0$ such that : $u(t_0) = 0\ and\ \dfrac{du}{dt}(t_0) \geq 0$.
The condition $u(t_0) = 0$ implies that $B(t_0) = \dfrac{r}{\gamma_B H}G(t_0) - D$.
From (1), we get

$$\frac{du}{dt}(t_0) = \frac{r}{H}\frac{dG}{dt}(t_0) - \gamma_B\frac{dB}{dt}(t_0),$$

and

$$\frac{du}{dt}(t_0) = -\frac{r}{H}\left[\frac{r}{H} + e\gamma_B\right]\frac{B(t_0)}{B(t_0)+D}G(t_0) - \frac{r}{H}m_G\left(G(t_0)\right)^2 + \frac{r}{H}\left[\gamma_G + m_B\right]x(t_0) - \frac{r}{H}Q - m_B\gamma_B D.$$

It follows that

$$\frac{du}{dt}(t_0) \leq -\frac{rm_G}{H}\left(G(t_0)\right)^2 + \frac{r}{H}\left[\gamma_G + m_B\right]G(t_0) - \frac{r}{H}Q - m_B\gamma_B D,$$

Condition (3) implies that $\frac{du}{dt}(t_0) < 0$ and we obtain a contradiction. This implies that $u(t) < 0$ for all $t \geq 0$.

Appendice C.

Proof of Proposition 5
The Jacobian matrix of (4) evaluated at the equilibrium P^*, is given by

$$J^* = \begin{pmatrix} -m_G G^* + \dfrac{Q}{G^*} & -\frac{r}{H}\frac{G^* D}{(B^*+D)^2} \\ \dfrac{er}{H}\dfrac{B^*}{B^*+D} & -\dfrac{er}{H}\dfrac{G^* B^*}{(B^*+D)^2} \end{pmatrix}.$$

Let $DetJ^*$ and TrJ^* be respectively the determinant and the trace associated to J^*, then

$$\begin{aligned} DetJ^* &= \frac{er}{H}\frac{B^*}{(B^*+D)^2}\left[m_G G^{2*} - Q + \frac{r}{H}\frac{G^* D}{B^*+D}\right], \\ &= \frac{er}{H}\frac{B^*}{(B^*+D)^2}\left[m_G G^{2*} - Q + \frac{m_B D}{e}\right], \end{aligned}$$

which is positive from the above conditions, and

$$\begin{aligned} TrJ^* &= \frac{1}{G^*}\left(-m_G\left(G^*\right)^2 + Q - \frac{er}{H}\frac{\left(G^*\right)^2 B^*}{\left(B^*+D\right)^2}\right), \\ &= \frac{1}{G^*}\Big(2Q - \frac{m_B D}{e} - \Big(\gamma_G - \frac{r}{H}\Big)G^* - \frac{er}{H}\frac{\left(G^*\right)^2 B^*}{\left(B^*+D\right)^2}\Big). \end{aligned}$$

Condition (2) and $H > H^*$ give that $TrJ^* < 0$. Hence the equilibrium is locally asymptotically stable when $H_1 < H < H_2$.

Appendice D.

Proof of Proposition 6

Because of the difficulty that face us in searching the sign of the trace of J^*, we treat it numerically. We present in the figure 5 the graph of the trace as a function of the water level H where H is between the levels H_1 and H_2.

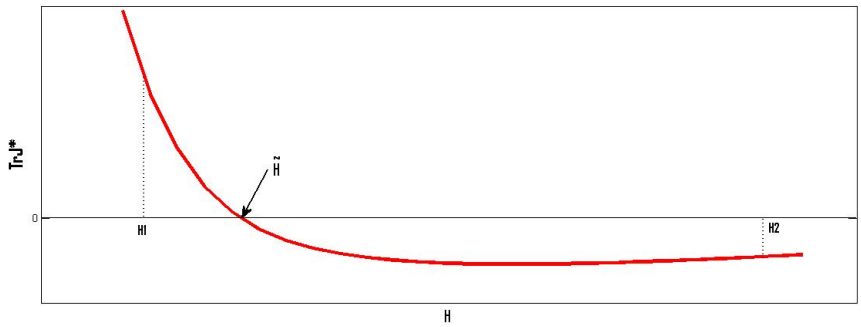


Figure 5. *The trace against the level water H*

According to this presentation, the trace at the point H_1 is positive and at the point H_2 it is negative. Moreover, between the levels H_1 and H_2, it is decreasing. Thus, According to the Intermediate Value Theorem, there exists a $\tilde{H}$ between H_1 and H_2 which annuls the trace and changes its sign.

We conclude that if $H_1 < H < \tilde{H}$, the trace is positive and therefore the interior equilibrium point P^* is unstable. Otherwise, if $\tilde{H} < H < H_2$, the trace changes sign and becomes negative, and then P^* is locally asymptotically stable.

Time series homogenization

Case of monthly temperature series of the northern part of Madagascar

Ralahady Bruno Bakys* — Totohasina André**

* Department of Mathematics and Computer Science
ENSET, University of Antsiranana -B.P. 0
ralahadybru@yahoo.fr

** andre.totohasina@gmail.com

ABSTRACT. The statistical technique for detecting jumps in the temperature series based on the regression model is favorable for homogenizing the climate data of the northern part of Madagascar. Thus, we will present the results of the homogenization of the series of maximum and minimum temperatures corresponding to the Antsiranana climate station. The homogenization of the temperature series is carried out at the monthly and daily scales.

RÉSUMÉ. La technique statistique pour la détection des sauts dans les séries de températures basée sur le modèle de régression est favorable pour homogénéiser les données climatiques de la partie Nord de Madagascar. Ainsi, nous allons présenté les résultats de l'homogénéisation des séries des températures maximales et minimales correspondant au station climatiques d'Antsiranana. L'homogénéisation des séries de températures est réalisée aux échelles mensuelle et quotidienne.

KEYWORDS : homogenization, hydrology, climatology, tests, trend, jumps.

MOTS-CLÉS : homogénéisation, hydrologie, climatologie, tests, tendance, sauts.

1. Motivations and introduction

The statistical characteristics of the recordings in a measuring station can undergo all kinds of artificial disturbances which do not reflect the real variations of the climate: displacement of stations, replacement of measuring instruments, change of hours of observations or modification of the immediate environment of the measuring instrument. As a result, decisions may be made based on data that contains errors. Meteorological network data are used in most climate variability research. The reliability of these data should be verified before using them in this area. Indeed, the need for long series of reliable climate data is increasingly felt in various areas. For example, climate change studies require the creation of comprehensive databases with which the climate signal can be adequately analyzed, tracked over time and predict future changes with minimal uncertainty of error. It is also very important to find robust techniques for detecting these artificial biases so that the data used is as close as possible to the observations that would have been made without disturbing the measurement conditions. The process of detection and correction of non-climatic breaks is called homogenization.

2. Data

The data used for this study come from the General Directorate of Meteorology, series of temperatures (minimum and maximum) at the time step per day from January 1st, 1950 to December 31st, 2008 from the five weather stations located in the Northern region of Madagasca (table: 1 and the tables:4 to 8)

Country	Number ID	Station Name		Longitude	Beginning	End
Madagascar	10111	Antsiranana	12° 21'04" S	49° 17'39" E	1950	2007
Madagascar	21011	Antalaha	14° 59'56" S	50° 19'12" E	1991	2004
Madagascar	20511	Sambava	14° 16'43" S	50° 10'29" E	1950	2008
Madagascar	30511	Nossy Be	13° 19'05" S	48° 18'33" E	1950	2007
Madagascar	67295	Vohemar	13° 22'22" S	49° 59'56" E	1950	2006

Table 1. *Characteristics of the base station and its neighbors*

3. Methodology

Various homogenization techniques([2], [3], [4], [6]) have been developed to accommodate different types of factors such as the variable to be homogenized, the spatial and temporal variability of the data depending on where the stations are located, the length of the series and the number of missing data (Aguilar et al., [1]).

The method we used in this study is based on linear regression models that look for both jumps and trends. These are models in which the conditions of normality, independence and heteroskedasticity of the data are assumed, and which are solved by standard least squares techniques. Multiple regression is based on the application of several regression models to homogenize temperature series (Vincent, [11]). When the residues are independent, the applied model fits the data well. If not, adjust with another model. The discontinuity determination in the basic series is identified with the following model:

$$y_i = \begin{cases} \tau + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + e_i + & i = 1 \ldots p \\ \tau + \delta + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + e_i & i = p+1 \ldots n \end{cases}$$

where y_i is the value of the base series at time i, x_{ik} is the value of the reference series k at time i. There are n observations and k reference series. The jump location, p, is determined by adjusting the model for all possible positions and selecting the one with the smallest sum of residual squares. The choice of the jump position is valid according to the Fisher test. The estimate of the jump amplitude is given by δ and its significant threshold is calculated according to the Student statistic (Vincent, [11]). A two-phase regression model can detect a change in mean and / or trend in a series (Solow, [10]). Either the adjusted model represents a series in which there is a one-point discontinuity p:

$$y_i = \begin{cases} \tau_1 + \lambda_1 i + e_i + & i = 1 \ldots p \\ \tau_2 + \lambda_2 i + e_i + & i = p+1 \ldots n \end{cases}$$

where y_i is the value of the base series at time i, τ_1 and τ_2 are the means before and after the change, λ_1 and λ_2 are the trends before and after the change and p is the position of the change. The model residuals are represented by e_i. The location of the jump is determined by least squares. Several changes have been made:

– Easterling and Peterson [6] apply the technique iteratively to detect several jumps and evaluate the significant thresholds by a multiple permutation procedure [8];

– Lund and Reeves [7] provide a revised Fisher statistic;

– Wang [13] proposes a model in which the slopes are equal before and after the break. We retained the improved version by Xiaolan Wang based on the maxima t test with penalty [14] and the maximum F test with penalty [16], nested in a recursive test algorithm [15].

This method has been used successfully in many studies on the analysis of extremes of precipitation and temperature around the world (Vincent et al., [12], Aguilar et al., [5]; Meehl et al, [9])

4. Homogenization process

It is almost impossible to be 100% sure of the quality of the past data, an assessment of homogeneity is always recommended. The best recommended technique is to go through

the following four steps below:

– Metadata analysis and quality control.

– Creating a series of reference times.

– Detection of change points (jumps).

– Data adjustment.

4.1. Quality control

It is applied to detect and identify errors made in the process of recording, handling, formatting, transmitting and archiving data.[1]

4.2. Quantiles-Match (QM) adjustment

It aims to adjust the series so that the empirical distribution of all segments of the trend-removed basic series match (Wang [15]); the value of the adjustment then depends on the empirical frequency of the values to be adjusted. As a result, the shape of the distribution is often adjusted, although the tests are supposed to detect jumps in averages; and the QM adjustment takes into account the seasonality of the change. Also, the annual cycle, the delay autocorrelation of 1, and the linear trend of the base series were estimated while explaining all the identified hops (Wang [15]); and the predicted trend for the base series is preserved in the QM adjustment algorithm.

The homogenization of the monthly temperatures is performed by adjusting the monthly temperature data observed before the date of the jump by correction factors. These correction factors are calculated taking into account the position of the jumps and their amplitudes obtained from the QM algorithm. The figures (figure: 1) and (figure: 2) represent the raw monthly temperature data observed and the linear trend jumps by multiphase regression model from 1960 to 2007 .

4.3. Homogenization of monthly series

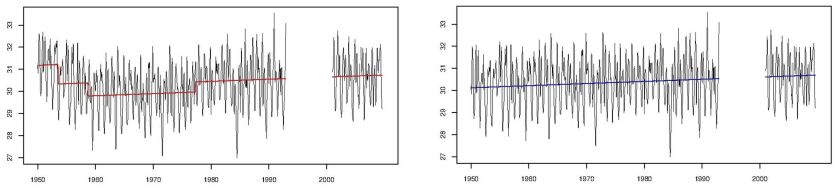


Figure 1. *Non-homogeneous and homogeneous series of maximum monthly temperatures, Antsiranana station*

The process of homogenization allowed us to retain three jumps in the series of maximum temperatures (figure: 1) and four jumps in the series of to one minimum temperatures (figure: 2).

Seg	Date	Amplitude	correction factors
1	1953 07	-0.88	-0.9892
2	1958 09	-0.5979	-0.1025
3	1977 04	0.4595	0.4281

Table 2. *Maximum temperature correction factors*

The tables (table: 2) and (array: 3) present respectively the estimated parameters of the regression model correction to n phase(s) corresponding to the detected jump in the series of monthly maximum and minimum temperatures.

The figures (figure: 5) and (figure: 4) represent the distribution of the QM (Quantile-Match) adjustments of each segment applied, respectively, to the maximum monthly temperature series and minimum.

The analysis of the figures (figure: 1) and (figure: 2) also shows that the smallest correction ($0 < A_m \leq 0,1^{\circ}C$) made to the monthly temperature series is performed during all periods, while the largest correction ($A_m \geq 1,9^{\circ}C$).

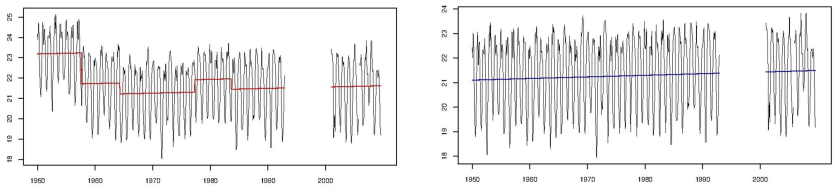


Figure 2. *Non-homogeneous and homogeneous series of minimum monthly temperatures, Antsiranana station*

Seg	Date	Amplitude	Correction factors
1	1957 08	-1.5242	-1.9416
2	1964 04	-0.5371	-0.5075
3	1977 04	0.6148	0.1057
4	1983 08	-0.5114	-0.4250

Table 3. *Minimum temperature correction factors*

The homogenization of the monthly series provides the date and the amplitude of the breaks detected. Although it is not possible to apply the correcting coefficients to the

daily data, the dates of the breaks nevertheless make it possible to determine supposedly homogeneous periods..

5. Discussion

The use of the QM adjustment takes into account the seasonality of the change; it is possible for the winter and summer temperatures to be adjusted differently because they belong to different quantiles of the distribution. This is a strong point of the temperature homogenization method used in this article. In fact, the anthropogenic influence of the measurement process at the climate station does not have the same impact on the measurement carried out during the different periods of the year.

We calculated the annual average from the homogenized maximum and minimum daily temperature series. These series are subsequently compared with the series of homogenized monthly temperatures obtained during the treatment. This comparison allows us to verify the consistency between the homogenization of annual temperature series and the homogenization of the monthly temperature series.

For the comparison, we calculated the average temperature per station, during the period 1950 to 2007, using the two homogenized monthly series. The analysis of the result clearly shows the coherence between the homogenization of annual temperatures and monthly temperatures.

6. Conclusion

The process of homogenization allowed us to retain three jumps in the series of maximum temperatures (1953 with an amplitude of $-0.88^{\circ}C$, 1958 with an amplitude of $-0.60^{\circ}C$ and 1977 with an amplitude of $0.46^{\circ}C$) and four jumps in the series of minimum temperatures (1957 with an amplitude of $-1.53^{\circ}C$, 1964 with an amplitude of $-0.54^{\circ}C$, 1974 with an amplitude of $0.61^{\circ}C$ and 1983 with an amplitude of$-0.51^{\circ}C$).

From the positions of jumps and their amplitudes, we found the monthly correction factors corresponding to the 12 months of the year. The tables 2 and 3 present the values of these factors for the three hops identified from the annual minimum temperature series. The table analysis 2 and 3 shows that the monthly correction factors are not distributed according to a uniform law.Thus, the corrections made to the monthly temperatures differ from one month to another. This constitutes a strong point of the method of homogenization of the temperatures used. In fact, the anthropogenic influence of the measurement process at the climate station does not have the same impact on the measurement made during the different periods of the year.

The monthly correction factors were calculated for all the minimum and maximum monthly temperature series corresponding to the 5 stations selected in this study. We also

adjusted the correction factors when the value of the mean absolute error is not equal to zero.

7. References

[1] E. Aguilar, Auer I., M. Brunet, Peterson T. C., and Wieringa J. Guidelines on climate metadata and homogenization. World Meteorological Organization, Geneva, Switzerland, 3 :WMO-TD 1186, 2003.

[2] H. Alexandersson. A homogeneity test applied to precipitation data. J. Climate, 6 :661-675., 1986.

[3] H. Alexandersson and A. Moberg. Homogenization of swedish temperature data. part i : A homogeneity test for linear trends. Int. J. Climatol., 17 :25-34, 1997.

[4] T. A. Buishand. Some methods for testing the homogeneity of rainfall records. J. Hydrol, 58 :11-27, 1982.

[5] Ramrez Obando Frutos R Retana J. A Solera M Soley J Gonzalez Garcia I Araujo R. M Rosa Santos A Valle V. E Brunet M Aguilar L lvarez L. A Bautista M Castanon C Herrera L Ruano E Sinay J. J Sanchez E Hernandez G I Oviedo Obed F Salgado J. E Vazquez J. L Baca M Gutierrez M Centella C Espinosa J E Aguilar,Peterson T. C. Changes in precipitation and temperature extremes in central america and northern south america, 1961-2003. Journal of geophysical research, VOL. 110, D23107, doi :10.1029/2005JD006119, 2005.

[6] T. C. Easterling, D. R. Peterson. A new method for detecting undocumented discontinuities in climatological time series. Int. J. Climatol., 15 :369-377, 1995.

[7] R. Lund and J. Reeves. Detection of undocumented changepoints : A revision of the two-phase regression model. J. Climate, 15 :2547 - 2554, 2002.

[8] Berry K. J. Brier G. W. Mielke, P. W. Application of multi-response permutation procedures for examining seasonnal changes in monthly mean sea-level pressure patterns. mon. Weath. Rev., 109 :120-126, 1981.

[9] T. Karl D.R. Easterling S. Changnon R. Pielke Jr. D.Changnon J. Evans P.Ya. Groisman T.R. Knutson K.E. Knukel L.O. Mearns C. Parmesan R. Pulwarty T. Root R.T. Sylves P.Whetton Meehl, G.A. and F. Zwiers. An introduction to trends in extreme weather and climate events :observations, socioeconomic impacts, terrestrial ecological impacts, and model projections. Bull. Am. Met.Soc., 81 :413-416, 2000.

[10] A. R. Solow. Testing for climate change : an application of the two-phase regression model. J. Clim. Appl. Met. 26, ., 26 :1401-1405, 1987.

[11] L. Vincent. A technique for the identification of inhomogeneities in canadian temperature series. J. Climate, 11 :1094-1104., 1998.

[12] Mekis É Vincent, L.A. and. Changes in daily and extreme temperature and precipitation indices for canada over the twentieth century. Atmosphere-Ocean, 44(2) :177-193, 2006.

[13] X. L. Wang. Comments on detection of undocumented changepoints : a revision of the two-phase regression model. J. Climate, 16 :3383-3385, 2003.

[14] X.L. Wang. Climatology and trends in some adverse and fair weather conditions in canada, 1953-2004. J. Geophys. Res., 111 :D09105, doi :10.1029/2005JD006155., 2006.

[15] X.L.Wang. Accounting for autocorrelation in detecting mean shifts in climate data series using the penalized maximal t or f test. J. App. Meteor. Climatol., 47 :2423-2444. DOI :10.1175/2008JAMC1741.1., 2008.

[16] X.L. Wang. Penalized maximal f test for detecting undocumented mean shift without trend change. J. Atmos. Oceanic Technol., 25 :368-384. DOI :10.1175/2007/JTECHA982.1, 2008.

Appendix 1

Extracts from station processing

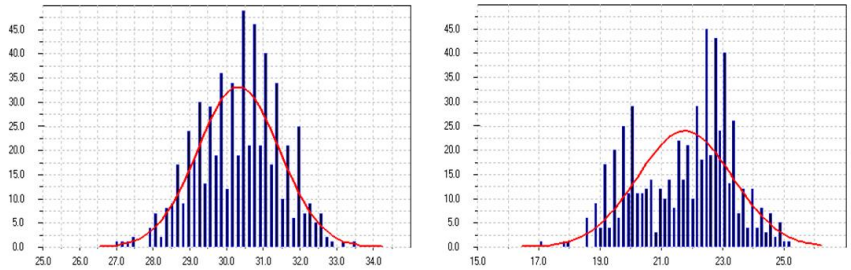


Figure 3. *Normal distribution of maximum and minimum temperature at* $[a; b] = 0.15\,^{\circ}C$ *from Antsiranana station*

Antsiranana Aero	Maximum temperature	Minimum temperature
Period	1950-2009	1950-2009
Length of Series	619	619
missing values	102	101
Arithmetic average	30.32	21.77
Standard deviation	1.11	1.55
Variance	1.24	2.39
Variance Coefficiente	3.67%	7.10%
Coefficient of Skew	-0.14	-0.32
Coefficient of Kurtosis	-0.36	-0.74
Maximum value	33.50 (1990.92)	25.10 (1953.17)
Minimal value	27 (1984.50)	17 (2007.58)
1st Quartile (25%)	29.50	20.40
Median	30.40	22.20
3rd Quartile (75%)	31.10	22.90
Kolmogorov-Smirnov test	$D = 0.05 (p = 0.09, O.K.)$	$D = 0.11 (p = 0.00, Non)$
Linear Regression Model	$y = 30.13 + 0.00 \times x$	$y = 22.30 - 0.00 \times x$
Coefficient of T-test b1	$T = 2.77 < 1.96 (95\%)$	$T = -5.44 > -1.96 (95\%)$
Trend / 10 years	$0.01 (Nom)$	$-0.02\,(Non)$
Determination Index (Correlation)	$0.01 (0.11)$	$0.05 (0.21)$
Variance (Residual + Estimate = Total)	$1.22 + 0.02 = 1.24$	$2.27 + 0.11 = 2.38$
Correlation Coefficient Series	$r1 = 0.65 < r1 (Tg_{95\%}) = 0.06\,(Non)$	$r1 = 0.81 < r1 (Tg_{95\%}) = 0.06\,(Non)$
Report by Von Neumann	$V = 0.70 > V (Tg_{95\%}) = 1.87\,(Non)$	$V = 0.37 > V (Tg_{95\%}) = 1.87\,(Non)$
Statistics of Rank Spearman	$rs = 0.09, t = 2.35 < Tkrit_{97.5\%} = 1.96\,(Non)$	$rs = -0.21, t = -5.46 < Tkrit_{97.5\%} = 1.96\,(Non)$
Statistics of Rang Mann-Kendall	$t = 0.04 < Tkrit_{95\%} = 0.05\,(O.K.)$	$t = -0.16 < Tkrit_{95\%} = 0.05\,(Non)$
Confidence Interval of Arithmetic Mean.	$(30.24, 30.41)$	$(21.65, 21.89)$

Table 4. *Statistical parameters of maximum and minimum temperature of Antsiranana stationa*

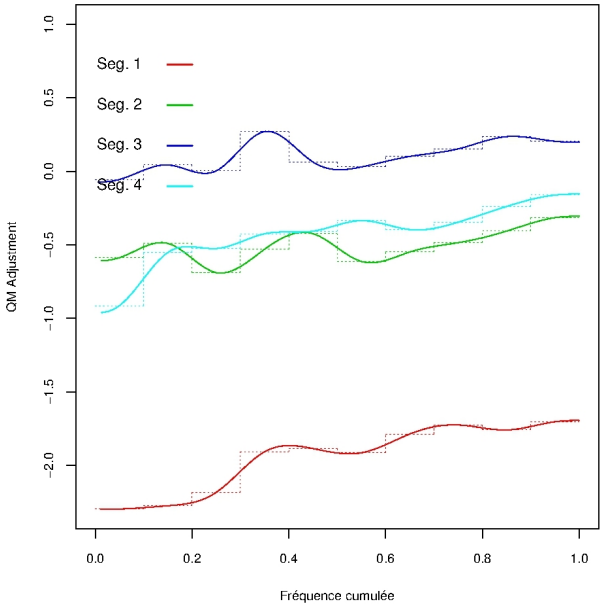


Figure 4. *Distribution of QM adjustments of minimum temperatures*

Vohemar Aero	Maximum temperature	Minimum temperature
Period	1950-2009	1950-2009
Length of Series	456	456
missing values	273	264
Arithmetic average	28.78	21.74
Standard deviation	1.64	1.52
Variance	2.69	2.32
Variance Coefficiente	5.70%	7.01%
Coefficient of Skew	-0.22	-0.35
Coefficient of Kurtosis	-1.04	-1.18
Maximum value	32.30 (1987.08,2001)	24.30 (1987.08)
Minimal value	25.10 (1984.50)	18.20 (1953.58)
1st Quartile (25%)	27.40	20.30
Median	29	22.10
3rd Quartile (75%)	30.20	23.05
Kolmogorov-Smirnov test	$D = 0.08 (p = 0.00, Non)$	$D = 0.15 (p = 0, Non)$
Linear Regression Model	$y = 28.56 + 0.00 \times x$	$y = 21.29 + 0.00 \times x$
Coefficient of T-test b1	$T = 1.77 < 1.97 (95\%)$	$T = 4.13 < 1.97 (95\%)$
Trend / 10 years	0.01	0.02 (out)
Determination Index (Correlation)	0.01 (0.08)	0.04 (0.19)
Variance (Residual + Estimate = Total)	$2.66 + 0.02 = 2.68$	$2.23 + 0.08 = 2.32$
Correlation Coefficient Series	$r1 = 0.78 < r1 (Tg_{95\%}) = 0.08 (Non)$	$r1 = 0.81 < r1 (Tg_{95\%}) = 0.07 (Non)$
Report by Von Neumann	$V = 0.44 > V (Tg_{95\%}) = 1.85 (Non)$	$V = 0.38 > V (Tg_{95\%}) = 1.85 (Non)$
Statistics of Rank Spearman	$rs = 0.06, t = 1.18 < Tkrit_{97.5\%} = 1.97 (O.K.)$	$rs = 0.19, t = 4.22 < Tkrit_{97.5\%} = 1.97 (Non)$
Statistics of Rang Mann-Kendall	$t = 0.02 < Tkrit_{95\%} = 0.06 (O.K.)$	$t = 0.11 < Tkrit_{95\%} = 0.06 (Non)$
Confidence Interval of Arithmetic Mean	$(28.62, 28.93)$	$(21.60, 21.88)$

Table 5. *Statistical parameters of maximum and minimum temperature of the Vohemar station*

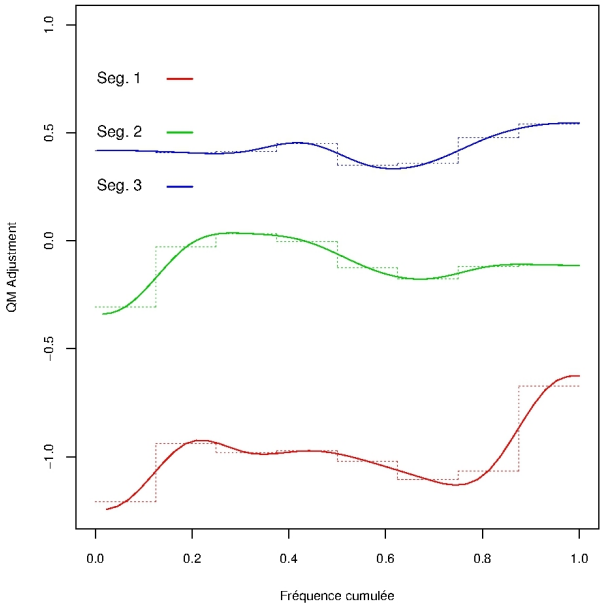


Figure 5. *The distribution of QM adjustments of maximum temperatures*

Antalaha Aero	Maximum temperature	Minimum temperature
Period	1991-2004	1991-2004
Length of Series	107	107
Missing values	61	61
Arithmetic average	28.63	21.10
Standard deviation	2.13	1.77
Variance	4.53	3.12
Variance Coefficiente	7.43%	8.37%
Coefficient of Kurtosis	-0.21	-0.15
Coefficient of Kurtosis	-1.35	-1.49
Maximum value	32.10 (1996)	23.70 (2002.08)
Minimal value	24.80 (1992.58)	18.10 (1992.50,1996.50)
1st Quartile (25%)	26.55	19.50
Median	28.90	21.30
3rd Quartile (75%)	30.50	22.80
Kolmogorov-Smirnov test	$D = 0.12(p = 0.11, O.K.)$	$D = 0.14(p = 0.03, Non)$
Linear Regression Model	$y = 28.47 + 0.00 \times x$	$y = 20.76 + 0.00 \times x$
Coefficient of T-test b1	$T = 0.59 < 1.98(95\%)$	$T = 1.43 < 1.98(95\%)$
Trend / 10 years	0.02	0.05
Determination Index (Correlation)	$0.00(0.06)$	$0.02(0.14)$
Variance (Residual + Estimate = Total)	$4.47 + 0.01 = 4.49$	$3.03 + 0.06 = 3.09$
Correlation Coefficient Series	$r1 = 0.82 < r1(Tg_{95\%}) = 0.15\,(Non)$	$r1 = 0.76 < r1(Tg_{95\%}) = 0.15\,(Non)$
Report by Von Neumann	$V = 0.36 > V(Tg_{95\%}) = 1.70\,(Non)$	$V = 0.48 > V(Tg_{95\%}) = 1.70\,(Non)$
Statistics of Rank Spearman	$rs = 0.07, t = 0.70 < Tkrit_{97.5\%} = 1.98\,(O.K.)$	$rs = 0.13, t = 1.36 < Tkrit_{97.5\%} = 1.98\,(O.K.)$
Statistics of Rang Mann-Kendall	$t = 0.03 < Tkrit_{95\%} = 0.13\,(O.K.)$	$t = 0.07 < Tkrit_{95\%} = 0.13\,(O.K.)$
Confidence Interval of Arithmetic Mean.	$(28.23, 29.04)$	$(20.76, 21.43)$

Table 6. *Statistical parameters of maximum and minimum temperature of Antalaha station*

Nosy Be Aero	Maximum temperature	Minimum temperature
Period	1950-2009	1950-2009
Length of Series	628	628
missing values	92	92
Arithmetic average	31.16	21.15
Standard deviation	1.12	1.92
Variance	1.24	3.69
Variance Coefficiente	3.58%	9.08%
Coefficient of Skew	-0.23	-0.48
Coefficient of Kurtosis	-0.45	-1.05
Maximum value	34.20 (2006.83)	24.60 (2008.08)
Minimal value	28.30 (1952.50,1956.50,1974.50)	16.70 (1968.50)
1st Quartile (25%)	30.40	19.40
Median	31.20	21.80
3rd Quartile (75%)	32	22.80
Kolmogorov-Smirnov test	$D = 0.06(p = 0.02, Non)$	$D = 0.15(p = 0, Non)$
Linear Regression Model	$y = 30.26 + 0.00 \times x$	$y = 20.95 + 0.00 \times x$
Coefficient of T-test b1	$T = 14.03 < 1.96(95\%)$	$T = 1.60 < 1.96(95\%)$
Trend / 10 years	0.03 (Non)	0.01
Determination Index (Correlation)	$0.24(0.49)$	$0.00(0.06)$
Variance (Residual + Estimate = Total)	$0.95 + 0.30 = 1.24$	$3.67 + 0.02 = 3.68$
Correlation Coefficient Series	$r1 = 0.66 < r1(Tg_{95\%}) = 0.06\,(Non)$	$r1 = 0.81 < r1(Tg_{95\%}) = 0.06\,(Non)$
Report by Von Neumann	$V = 0.68 > V(Tg_{95\%}) = 1.87\,(Non)$	$V = 0.38 > V(Tg_{95\%}) = 1.87\,(Non)$
Statistics of Rank Spearman	$rs = 0.49, t = 14.02 < Tkrit_{97.5\%} = 1.96\,(Non)$	$rs = 0.08, t = 1.88 < Tkrit_{97.5\%} = 1.96\,(O.K.)$
Statistics of Rang Mann-Kendall	$t = 0.31 < Tkrit_{95\%} = 0.05\,(Non)$	$t = 0.03 < Tkrit_{95\%} = 0.05\,(O.K.)$
Confidence Interval of Arithmetic Mean.	$(31.07, 31.25)$	$(21.00, 21.30)$

Table 7. *Statistical parameters of maximum and minimum temperature of Nossy Be station*

Sambava Aero	Maximum temperature	Minimum temperature
Period	1950-2009	1950-2009
Length of Series	566	566
missing values	154	159
Arithmetic average	28.75	20.51
Standard deviation	1.75	1.89
Variance	3.07	3.56
Variance Coefficiente	6.09%	9.19%
Coefficient of Skew	-0.05	-0.09
Coefficient of Kurtosis	-1.28	-1.29
Maximum value	32.30 (2007.08)	24.10 (1992)
Minimal value	24.80 (1984.50)	16.60 (1956.50,1959.67)
1st Quartile (25%)	27.10	18.77
Median	28.90	20.70
3rd Quartile (75%)	30.30	22.20
Kolmogorov-Smirnov test	$D = 0.10(p = 0.00, Non)$	$D = 0.10(p = 0.00, Non)$
Linear Regression Model	$y = 28.76 - 0.00 \times x$	$y = 19.94 + 0.00 \times x$
Coefficient of T-test b1	$T = -0.11 > -1.96(95\%)$	$T = 4.44 < 1.96(95\%)(O.K)$
Trend / 10 years	-0.00	$0.02\,(Non)$
Determination Index (Correlation)	$0.00(0.00)$	$0.03(0.18)$
Variance (Residual + Estimate = Total)	$3.06 + 0.00 = 3.06$	$3.43 + 0.12 = 3.55$
Correlation Coefficient Series	$r1 = 0.80 < r1(Tg_{95\%}) = 0.07\,(Non)$	$r1 = 0.83 < r1(Tg_{95\%}) = 0.07\,(Non)$
Report by Von Neumann	$V = 0.39 > V(Tg_{95\%}) = 1.87\,(Non)$	$V = 0.34 > V(Tg_{95\%}) = 1.86\,(Non)$
Statistics of Rank Spearman	$rs = -0.02, t = -0.48 < Tkrit_{97.5\%} = 1.96\,(O.K.)$	$rs = 0.22, t = 5.32 < Tkrit_{97.5\%} = 1.96\ (Non)$
Statistics of Rang Mann-Kendall	$t = -0.03 < Tkrit_{95\%} = 0.06\,(O.K.)$	$t = 0.14 < Tkrit_{95\%} = 0.06\,(Non)$
Confidence Interval of Arithmetic Mean.	$(28.60, 28.89)$	$(20.36, 20.67)$

Table 8. *Statistical parameters of maximum and minimum temperature of Sambava station*

Coupling Discontinuous Galerkin method and integral representation for solving Maxwell's system

Anis Mohamed* — Nabil Gmati** — Stephane Lanteri***

* Tunis El Manar University, National School of Engineering of Tunis
LAMSIN-ENIT, BP 37, LE BELVEDERE Tunis 1002
Tunisia
midani.anis@gmail.com
** Tunis El Manar University, National School of Engineering of Tunis
LAMSIN-ENIT, BP 37, LE BELVEDERE Tunis 1002
Tunisia
nabil.gmati62@gmail.com
*** INRIA, NACHOS project-team
2004 Route des Lucioles, B.P. 93 06902 Sophia Antipolis Cedex
France
Stephane.Lanteri@inria.fr

ABSTRACT. we present a mathematical and numerical study of the three-dimensional time-harmonic Maxwell equations solved by a discontinuous Galerkin method coupled with an integral representation. This study was completed by some numerical tests to justify the effectiveness of the proposed approach.

RÉSUMÉ. nous présentons une étude mathématique et numérique pour la résolution des équations de Maxwell tridimensionnelles en régime-harmonique, par une méthode de type Galerkin discontinu couplée à une représentation intégrale. Cette étude a été complétée par des tests numériques pour justifier l'efficacité de l'approche proposée.

KEYWORDS : Maxwell equations, time-harmonic, discontinuous Galerkin method, integral representation, fictitious domain

MOTS-CLÉS : Équations de Maxwell, régime-harmonique, méthode de Galerkin discontinu, représentation intégrale, domaine fictif

1. Introduction

The propagation of electromagnetic waves is a physical phenomenon that describes the analysis of an emitted wave, this phenomenon is described by mathematical equations. In this work, the electromagnetic wave propagation equation will result in Maxwell's equations.

What's interesting for Maxwell's equations is that the domain of validity extends to a wide variety of waves: radar, TV, radio, ... and even in radiation fields as varied: Ultra-violet, X-rays, infra-red, gamma, etc.

Various methods have been developed for numerical resolution of Maxwell's equations, however, it seems that no method is predominant if we take into account the BF–MF–HF domains we are interested in BF–MF domains. Our work in this paper is devoted to the resolution of three-dimensional time-harmonic Maxwell's equations by the discontinuous galerkin method coupled to an integral representation.

2. Maxwell's problem

We are interested in this paper to the solutions of the time-harmonic Maxwell's equations in the presence of an obstacle D, which are particular solutions and which shall check the following system:

$$\begin{cases} \nabla \times E + i\omega\mu\, H &= J, & \text{in} & \quad \mathbb{R}^n \backslash \overline{D}, \\ \nabla \times H - i\omega\varepsilon\, E &= 0, & \text{in} & \quad \mathbb{R}^n \backslash \overline{D}, \end{cases} \tag{1}$$

where E and H are respectively the electric and magnetic fields. The parameters ε is the relative dielectric permittivity, μ is the relative magnetic permeability and ω is the pulsation.

So the perfect conductor condition will be considered on boundary Γ_m define here:

$$\begin{cases} E \times n &= 0, \\ H \,.\, n &= 0. \end{cases}$$

This problem is posed on an initially infinite domain; the idea here is to limit our domain by a fictitious boundary we will note it Γ_a.

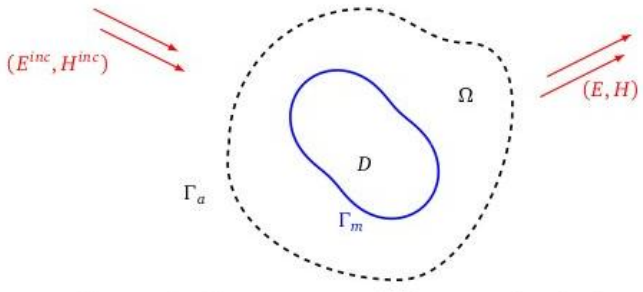


Figure 1. *Diffraction of an electromagnetic wave in the presence of an obstacle D where its boundary is noted Γ_m*

We consider on this boundary an exact condition in the form of an integral representation defined by:

$$n \times E \; + \; n \times (n \times H) \; = \; n \times \; \mathfrak{R}(E) \; + \; n \times (n \times \; \mathfrak{R}(H)) \; ,$$

where $\mathfrak{R}(E)$ and $\mathfrak{R}(H)$ are respectively the values of E and H on Γ_a expressed as a function of E and H in Γ_m defined using the integral representation by the Stratton-Schu formulas [3, 4] given by:

$$\mathfrak{R}(E) = \; \mathcal{L}\, g \; - \; \mathcal{K}\, f \quad \text{and} \quad \mathfrak{R}(H) = \; \mathcal{L}\, f \; + \; \mathcal{K}\, g \; ,$$

where $f = n \times E$, $g = -n \times H$ and for the fundamental solution of the Helmholtz problem (the Green function G):

$$\Big(\mathcal{G}\, u\Big)(x) \; = \; \int_{\Gamma} G(x,y)\, u(y)\, \mathrm{d}y \; , \; \mathcal{L}\, u \; = \; \frac{1}{i\, k}\, \nabla \times \nabla \times \mathcal{G}\, u \quad \text{and} \quad \mathcal{K}\, u \; = \; \nabla \times \mathcal{G}\, u$$

For simplicity we assume that $J = 0$. At this phase, we then come back to a problem:

$$\left\{ \begin{array}{lcll} \text{Find } E, H \in H(\nabla\times, \Omega) & , & \text{such as:} & \\ i\omega\varepsilon\, E - \nabla \times H & = & 0 & \text{in} \quad \Omega \\ i\omega\mu\, H + \nabla \times E & = & 0 & \text{in} \quad \Omega \\ n \times E & = & -n \times E^{inc} & \text{on} \quad \Gamma_m \\ n \times E - n \times (n \times H) & = & n \times \mathfrak{R}(E) - n \times (n \times \mathfrak{R}(H)) & \text{on} \quad \Gamma_a \end{array} \right. \tag{2}$$

where $H(\nabla\times, \Omega) = \{v \in L^2(\Omega)^3 : \nabla \times v \in L^2(\Omega)^3\}$ and:

$$E^{inc} = \begin{bmatrix} E_1^{inc} \\ E_2^{inc} \\ E_3^{inc} \end{bmatrix} \quad , \; E = \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix} \quad , \; H = \begin{bmatrix} H_1 \\ H_2 \\ H_3 \end{bmatrix} \quad \text{et} \quad n = \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix}$$

In the vector field W such that $W = \begin{bmatrix} E \\ H \end{bmatrix}$, the problem (2) will be written in this matricial form:

$$\left\{ \begin{array}{ll} i\omega Q W + G_x \partial_x W + G_y \partial_y W + G_z \partial_z W & = 0 \quad \text{on} \; \Omega \\ (M_{\Gamma_m} - G_n)\, (W + W^{inc}) & = 0 \quad \text{in} \; \Gamma_m \\ (M_{\Gamma_a} - G_n)\, (W - \mathfrak{R}(W)) & = 0 \quad \text{in} \; \Gamma_a. \end{array} \right. \tag{3}$$

Where $Q = \begin{bmatrix} \varepsilon\, I_3 & 0_{3\times3} \\ 0_{3\times3} & \mu\, I_3 \end{bmatrix}$

and $\mathfrak{R}(W) = \begin{bmatrix} \mathfrak{R}(E) \\ \mathfrak{R}(H) \end{bmatrix}$ such that: $\Big(\mathfrak{R}(W)\Big)(x) \; = \; \displaystyle\int_{\Gamma_m} K(x,y) W(y) \partial\sigma_y$ where $K : \mathbb{R}^6 \times \mathbb{R}^6 \to M_6(\mathbb{C})$ is a Green kernel.

In fact, denoting by (e_x, e_y, e_z) the canonical basis of $\mathbb{R}^3$, the matrices G_l for $k \in \{x, y, z\}$ are defined by:

$G_k = \begin{bmatrix} 0_{3\times3} & N_{e_k} \\ N_{e_k}^t & 0_{3\times3} \end{bmatrix}$ where for $l \in \{1, 2, 3\}$ a vector $v = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$, $N_v = \begin{bmatrix} 0 & v_3 & -v_2 \\ -v_3 & 0 & v_1 \\ v_2 & -v_1 & 0 \end{bmatrix}$

Furthermore, $G_n = G_x\, n_1 + G_y\, n_2 + G_z\, n_3$.

G_n^+ and G_n^- denote the positive and negative parts of G_n [1]. We also define $|G_n| \; =$

1. If $P\Lambda P^{-1}$ is the natural factorization of G_n then $G_n^{\pm} = P\Lambda^{\pm}P^{-1}$ where Λ^+ (resp. Λ^-) includes only positive eigenvalues (resp. negative).

$G_n^+ - G_n^-$. The matrices M_{Γ_m} et M_{Γ_a}, are then defined by:

$$M_{\Gamma_m} = \begin{bmatrix} 0_{3\times3} & N_n \\ -N_n^t & 0_{3\times3} \end{bmatrix} \qquad \text{and} \qquad M_{\Gamma_a} = |G_n|$$

3. Variational formulation of the problem and discretization

We decompose the domain Ω in tetrahedral elements, we denote by τ_h the set of elements K_i.
For all $K_i \in \tau_h$, we define the functional space

$$V_h = \{W \in [L^2(\Omega)]^6 \; ; \; W_{|_{K_i}} = W_i \in P_p(K)\}$$

By a development similar to that adopted by Ern and Guermond [1, 2] and adding the terms of the integral representation, the variational formulation of the problem (3) consists of:
$\forall \; V \in V_h \times V_h$, K_i an element of τ_h obtained:
Find $W_i = (E_i, H_i) \in V_h \times V_h$ such as:

$$\begin{aligned}
\int_{K_i} (i\omega Q W_i)^t \, \overline{V} dx \quad &- \quad \int_{K_i} W_i^t \, (\sum_{l\in\{x,y,z\}} G_l \partial_l \overline{V}) dx \\
&+ \quad \int_{F\in\Gamma_i^0} \left[(I_{FK_i} S_F \, [\![W_i]\!])^t \, \overline{V} + (I_{FK_i} G_{n_F} \, \{W_i\})^t \, \overline{V} \right] \partial\sigma \\
&+ \quad \int_{F\in\Gamma_i^a} (\frac{1}{2}(M_{F,K_i} + I_{FK_i} G_{n_F}) W_i)^t \overline{V} \partial\sigma \\
&- \quad \int_{F\in\Gamma_i^a} (\frac{1}{2}(M_{F,K_i} - I_{FK_i} G_{n_F}) \mathfrak{R}(W_i))^t \overline{V} \partial\sigma \\
&+ \quad \int_{F\in\Gamma_i^m} (\frac{1}{2}(M_{F,K_i} + I_{FK_i} G_{n_F}) W_i)^t \overline{V} \partial\sigma \\
&= \quad \int_{F\in\Gamma_i^m} (\frac{1}{2}(M_{F,K_i} - I_{FK_i} G_{n_F}) W_i^{inc})^t \overline{V} \partial\sigma
\end{aligned}$$

where: $\Gamma_i^0 = \bigcup_{K_j\in\tau_h} \overline{K_i} \cap \overline{K_j}$, $\Gamma_i^m = \bigcup_{K_i\in\tau_h} \overline{K_i} \cap \Gamma_m$ and $\Gamma_i^a = \bigcup_{K_i\in\tau_h} \overline{K_i} \cap \Gamma_a$.
I_{FK} represents the incidence matrix between facing surfaces and elements whose entries are given by:

$$I_{FK} = \begin{cases} 1 & \text{if } \; F \in K \; \text{ and orientations of } n_F \text{ and } n_K \text{ are match,} \\ -1 & \text{if } \; F \in K \; \text{ and orientations of } n_F \text{ and } n_K \text{ do not match,} \\ 0 & \text{if the face } \; F \; \text{ does not belong to the element } \; K. \end{cases}$$

where: n_F is the unitary normal associated to the oriented face F and n_K is the unitary normal associated to the cell K.
We also define respectively the jump and average of a vector V to $V_h \times V_h$ on the face F shared between two elements K and $\widetilde{K}$

$$[\![V]\!] = I_{FK} \, V_{|_K} + I_{F\widetilde{K}} \, V_{|_{\widetilde{K}}} \qquad \text{and} \qquad \{V\} = \frac{1}{2} \left(V_{|_K} + V_{|_{\widetilde{K}}} \right)$$

The matrices S_F and $M_{F,K}$ are defined following the choice of numerical fluxes:

3.1. Centered flux

In this case, $S_F = 0$ and $M_{F,K} = \begin{cases} I_{FK} \begin{bmatrix} 0_{3\times 3} & N_{n_F} \\ -N_{n_F}^t & 0_{3\times 3} \end{bmatrix} & \text{if } F \in \Gamma^m. \\ |G_{n_F}| \quad \text{if } F \in \Gamma^a. \end{cases}$

3.2. Upwind flux

In this case, $S_F = \begin{bmatrix} \alpha_F^E N_{n_F} N_{n_F}^t & 0_{3\times 3} \\ 0_{3\times 3} & \alpha_F^H N_{n_F}^t N_{n_F} \end{bmatrix}$ and

$$M_{F,K} = \begin{cases} \begin{bmatrix} \eta_F N_{n_F} N_{n_F}^t & I_{FK} N_{n_F} \\ -I_{FK} N_{n_F}^t & 0_{3\times 3} \end{bmatrix} & \text{if } F \in \Gamma^m. \\ |G_{n_F}| \quad \text{if } F \in \Gamma^a, \end{cases}$$

for a homogeneous medium, $\eta_F = \alpha_F^E = \alpha_F^H = \frac{1}{2}$

4. Linear system of the problem

We will treat the variational formulation term by term we can reduce our formulation in vectorial form

$$\left[D_i^1 - D_i^2 + D_i^{\Gamma_0} + \delta_{F_i^m}\, D_i^{\Gamma_m} + \delta_{F_i^a}\, D_i^{\Gamma_a}\right] W_i + \sum_{j\in V_i} E_{ij}\, W_j + \delta_{F_i^a} \sum_{j:K_j\cap\Gamma_m\neq\varnothing} C_{ij}\, W_j = \delta_{F_i^m}\, B_i^{inc}$$

where: $D_i^1 = i\omega \left(\Phi_i \otimes Q\right)$, $D_i^2 = \sum_{l=1}^{3} \left(\Phi_i^l \otimes G_l\right)$,

$D_i^{\Gamma_m} = \left(\Psi_{F_i^m} \otimes \left[\frac{1}{2}(M_{F,K_i} + I_{FK_i} G_{n_F})\right]\right)$,

$D_i^{\Gamma_a} = \left(\Psi_{F_i^a} \otimes \left[\frac{1}{2}(M_{F,K_i} + I_{FK_i} G_{n_F})\right]\right)$,

$D_i^{\Gamma_0} = \left(\Psi_i \otimes \left[I_{FK_i}(S_F\, I_{FK_i} + \frac{1}{2} G_{n_F})\right]\right)$,

$E_{ij} = \sum_{j\in V_i} \left(\Psi_{ij} \otimes \left[I_{FK_i}(S_F\, I_{FK_j} + \frac{1}{2} G_{n_F})\right]\right)$,

$C_{ij} = \frac{1}{2} \left(\Psi^{\Gamma_i^a} \otimes I_6\right) \widetilde{K}_{ij} \left(\Psi^{\Gamma_j^m} \otimes I_6\right)$,

$B_i^{inc} = Z_i W_i^{inc} = \left(\Psi_{\Gamma_i^m} \otimes \left[\frac{1}{2}(M_{F,K_i} - I_{FK_i} G_{n_F})\right]\right) W_i^{inc}$,

$F_{ij} = \overline{K_i} \cap \overline{K_j}$, $F_i^m = \overline{K_i} \cap \Gamma_m$, $F_i^a = \overline{K_i} \cap \Gamma_a$,

V_i : the set of indices of neighboring elements of K_i ,

$$\delta_{F_i^a} = \begin{cases} 1 & \text{if} \quad \Gamma_a \cap K_i = F_i^a \\ 0 & \text{if} \quad \Gamma_a \cap K_i = \varnothing \end{cases} \quad \text{and } \delta_{F_i^m} = \begin{cases} 1 & \text{if} \quad \Gamma_m \cap K_i = F_i^m \\ 0 & \text{if} \quad \Gamma_m \cap K_i = \varnothing \end{cases}$$

we can reduce our problem as a linear system:

$$(A - C)\, X = B$$

– A is the square matrix of size:

$$N = 6 \times \underbrace{\text{Number of degrees of freedom}}_{d_i} \times \underbrace{\text{Number of cells}}_{N_c}$$

this matrix is a sparse matrix defined by block size $(6\ d_i\ \times\ 6\ d_i)$ such as:

- For $i = 1, \ldots, N_c$:

$$A(i,i) = D_i^1 - D_i^2 + D_i^{\Gamma^0} \times \delta_{ij} + D_i^{\Gamma^m} \times \delta_{\Gamma_i^m} + D_i^{\Gamma^a} \times \delta_{\Gamma_i^a}$$

- For $i, j = 1, \ldots, N_c$:

$$A(j,i) = E_{ij} \times \delta_{ij}$$

with:

$$\delta_{ij} = \begin{cases} 0 & \text{if} \quad K_i \cap K_j = \varnothing \\ 1 & \text{else} \end{cases}$$

– C is a square matrix of the same size as A, defined by block size $6\ d_i\ \times\ 6\ d_j$ such as:

- For $i, j = 1, \ldots, N_c$:

$$C(i,j) = -C_{ij} \times \delta_{\Gamma_i^a} \times \delta_{\Gamma_j^m}$$

where:

$$\delta_{\Gamma_j^m} = \begin{cases} 0 & \text{if} \quad K_j \cap \Gamma_a = \varnothing \\ 1 & \text{else} \end{cases}$$

– X is the vector of size N, Where its components are the unknowns of our problem.

– B is the vector of size N such as: $B(i) = B_i^{inc} \times \delta_{F_i^m}$

5. Numerical results

Following the mathematical study of the resolution of the Maxwell equations in unbounded domain by a method of type coupled with an integral representation (DG+IR), we present a sample of the numerical results.

We will give some numerical results by making the comparison between the approximate solution and the exact solution.

Mesh	#M1	#M2	#M3
Distance between Γ_m and Γ_a	0.2	0.4	0.6
h_{max}	0.1	0.1	0.1
Number of elements	204222	476454	830879
Relative error (DG)	0.467×10^{-1}	0.288×10^{-1}	0.286×10^{-1}
Relative error (DG+IR)	0.843×10^{-2}	0.883×10^{-2}	0.909×10^{-2}

Table 1. *Variation of external radius, k=5*

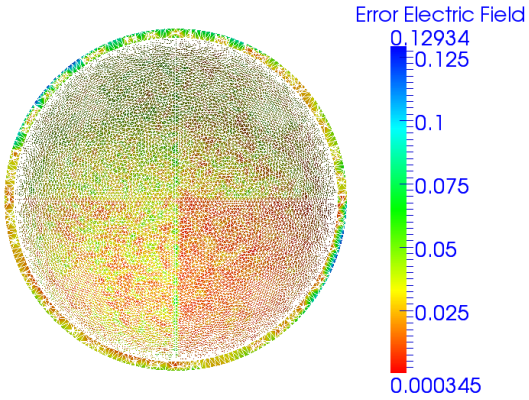


Figure 2. *Meshing of the volume between a first sphere of radius $R = 1$ and a second sphere of radius $R = 1.06$. A mesh size $h = 0.07$.*

5.1. Performance of methods with centered flux & upwind flux

The comparison results between the two methods DG+IR and DG are illustrated in the form of the relative error between the exact solution and the approximate solution either using a centered flux (see also figure (3)) or an upwind flux (see also figure (4)).

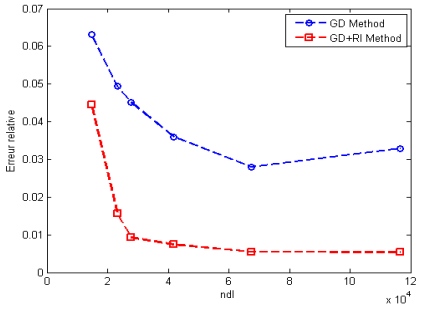


Figure 3. *Electric Field Error according to degree of freedom: Centered flux*

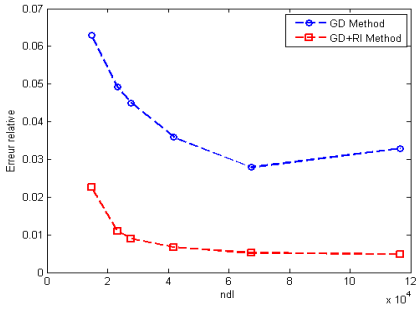


Figure 4. *Electric Field Error according to degree of freedom: Upwind flux*

A good improvement of the convergence is observed by using the DG method coupled to an integral representation using either a centered flux or an upwind flux.

5.2. Error depending on the size of the domain of study

We are interested in the case where the discretization step h and the waves number $k = 10$ are fixed and by varying the distance delimited between the boundary of the obstacle Γ_m and the artificial boundary Γ_a by keeping a choice of wavelength equal to $20h$.

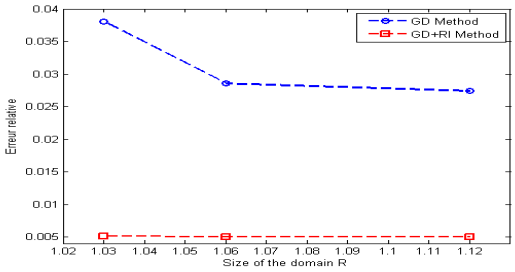


Figure 5. *Error according the size of the domain R.*

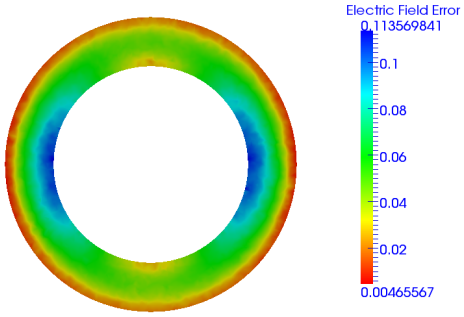


Figure 6. *Electric Field Error by the DG method.*

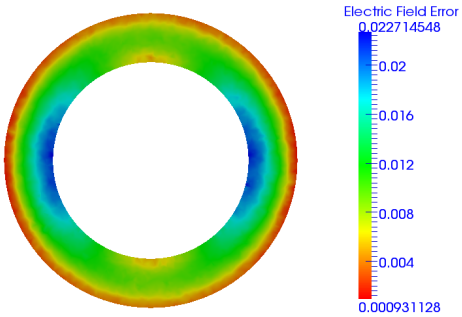


Figure 7. *Electric Field Error by the DG+IR method.*

It is clear that the results obtained by the DG+IR method are better, which shows that the coupling method is the most efficient. They show an show an improvement in accuracy, especially when the fictitious border is close to the boundary of the obstacle.

6. References

[1] A. ERN, J.-L. GUERMOND, "Discontinuous Galerkin methods for Friedrichs systems I. General theory", *SIAM J. Numer. Anal.*, vol. 44, num. 2, 2006.

[2] A. ERN, J.-L. GUERMOND, "Discontinuous Galerkin methods for Friedrichs systems II. Second-order elliptic PDE's", *SIAM J. Numer. Anal.*, vol. 44, num. 6, 2006.

[3] D. COLTON, R. KRESS, "Inverse Acoustic and Electromagnetic Scattering Theory", *Springer Verlag*, 1997.

[4] B. -MIREBEAU, J. BOURGUIGNON, "Preconditioning domain decomposition methods for electromagnetic scattering problems involving a deep cavity", *Thesis*, 2011.

[5] N. GMATI, S. LANTERI, A. MOHAMED, "Discontinuous Galerkin method coupled with an integral representation for solving the three-dimensional time-harmonic Maxwell equations", *Applied Acoustics*, vol. 108, 2016.

Split-convexity method for image restoration

Hamdi Houichet* — Anis Theljani** — Badreddine Rjaibi* — Maher Moakher*

* Université de Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis,
Laboratoire de Modélisation Mathématique et Numérique dans les Sciences de l'Ingénieur
B.P. 37, 1002 Tunis, Tunisia.
hamdi.houichet@enit.utm.tn (H. Houichet), badreddine.rjaibi@lamsin.rnu.tn (B. Rjaibi),
maher.moakher@enit.utm.tn (M. Moakher).
** Liverpool Centre for Mathematics in Healthcare, Department of Mathematical Sciences, University of Liverpool, Liverpool, UK.
A.Theljani@liverpool.ac.uk

ABSTRACT. Image processing such as image denoising, edge detection and image segmentation, etc, plays an important role in various fields. The objective of image denoising is to reconstruct an original image from a noisy one. In this talk, we propose a nonlinear equation based on the $p(\cdot)$-biharmonic operator to denoise the images. First, we consider a topological gradient approach in order to detect important objectes of the image. Then, we chose the variable exponent $p(\cdot)$ adaptively based on the map furnished by the topological gradient. Finally, we consider the split convexity method in order to linearize the proposed equation. We present some numerical examples to show the performance of the proposed method.

RÉSUMÉ. Le traitement d'images tel que le débruitage, la détection de contour et la segmentation, etc., jouent un rôle important dans divers domaines. L'objectif du débruitage d'image est de reconstruire une image originale à partir d'une image bruitée. Dans cet exposé, nous proposons un modèle non linéaire basé sur l'opérateur $p(\cdot)$-biharmonic pour restaurer les images. Premièrement, nous considérons une approche de gradient topologique afin de détecter les importants objets de l'image. Ensuite, nous avons choisi la variable $p(\cdot)$ en fonction du gradient topologique. Finalement, nous considérons la méthode de la décomposition convexe afin de linéariser le modèle proposé. Nous présentons quelques exemples numériques pour montrer la performance de la méthode proposée.

KEYWORDS : variational method, image denoising, biharmonic operator, topological gradient, speckle noise, nonlinear PDE.

MOTS-CLÉS : Méthode variationnelle, débruitage d 'image, opérateur biharmonique, gradient topologique, bruit de type speckle, EDP non linéaire.

1. Introduction

In this work, we are interested in the restoration of images highly corrupted with multiplicative noise. The objective of image denoising is to reconstruct an image $u : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$ from an observed one $f : \Omega \to \mathbb{R}$ which is degraded and contaminated by noise. The degradation model that we consider is the following:

$$f = u + \eta\sqrt{u}, \tag{1}$$

where $\eta : \Omega \to \mathbb{R}$ is a positive function and that follows the Rayleigh-distribution. Model (1) represents the degradation of an image corrupted by speckle noise, usually present in medical ultrasound imaging [5, 6]. We consider the following partial deferential equation for denoising the ultrasound image:

$$\begin{cases} \Delta^2_{p(x)} u + \alpha \frac{u^2 - f^2}{u^2} = 0, & \text{in } \Omega, \\ \partial_n \Delta u = \partial_n u = 0, & \text{on } \partial\Omega, \end{cases} \tag{2}$$

where $\Delta^2_{p(x)} u := \Delta(|\Delta u|^{p(x)-2}\Delta u)$ is the $p(\cdot)$-biharmonic operator, $p : \Omega \to]1, 2]$ is measurable function called exponent and α is a positive parameter. For more details about the exponent functions, we refer the reader to [3, 7].

The variable exponent $1 < p(x) \leq 2$ is chosen so that to slow diffusion near edges in order to highlight them, and to enhance diffusion in smooth regions. A classical idea of choosing the values of the exponent p is to make an adaptive procedure as follows: first, we consider the topological gradient method with $p(x) = 2$ to identify the edges in order to preserve them. Second, we perform a local selection of the exponent $1 < p(x) \leq 2$ with the help of the map furnished by the topological gradient.

2. Well-posedness

Let Ω be a bounded and Lipschitz open subset in $\mathbb{R}^2$. In the following theorem, we establish the well-posedness of equation (2).

Theorem 2.1. *Let $f \in X = \{u \in W^{2,p(x)}(\Omega)$ such that $\frac{\partial u}{\partial n} = 0$ on $\partial\Omega\}$ with $\inf_\Omega f > 0$. Then, equation* (2) *admits a unique solution u in X satisfying the maximum principal*

$$\inf_\Omega f \leq u \leq \sup_\Omega f.$$

Proof. First, we note that (2) is the Euler-Lagrange equation of the following minimization problem

$$\min_{\{u \in X, u > 0\}} \left\{ E_{p(x)}(u) := \int_\Omega |\Delta u|^{p(x)} \, \mathrm{d}x + \alpha \int_\Omega \frac{(f-u)^2}{u} \, \mathrm{d}x \right\}. \tag{3}$$

By the classical compactness, semi-continuity and convexity arguments of the energy $E_{p(x)}(\cdot)$, it is easy to verify that (3) has a unique minimizer, which is equivalently the unique weak solution of (2). □

3. Edges-detection and preservation

The big challenge in image restoration and segmentation problems is how to accurately detect features such as arteries, filaments, internal organ, etc. To meet with this challenge, we employ here the topological gradient method which was widely used in edge detection [1, 5]. We start by inserting a small crack $\sigma_\varepsilon := \{x_0 + \varepsilon\sigma(n)\}$ in the domain Ω, where $x_0 \in \Omega$, $\varepsilon > 0$, $\sigma(n)$ is a straight crack, and n is a unit vector normal to the crack and we minimize the cost function

$$j(\Omega_\varepsilon) := J(u_\varepsilon) = \int_{\Omega\setminus\sigma_\varepsilon} |\Delta u_\varepsilon|^2 \,\mathrm{d}x,$$

where u_ε is the unique solution of the following equation defined on the *perturbed domain* $\Omega_\varepsilon \overset{\text{def}}{=} \Omega \setminus \overline{\sigma}_\varepsilon$:

$$\begin{cases} \Delta^2 u_\varepsilon + \alpha \frac{u_\varepsilon^2 - f^2}{u_\varepsilon^2} = 0, & \text{in } \Omega_\varepsilon, \\ \frac{\partial \Delta u_\varepsilon}{\partial n} = \frac{\partial u_\varepsilon}{\partial n} = 0, & \text{on } \partial\Omega, \\ \frac{\partial \Delta u_\varepsilon}{\partial n} = \Delta u_\varepsilon = 0, & \text{on } \partial\sigma_\varepsilon. \end{cases} \tag{4}$$

After that, we measure the impact of such a modification of the domain on this cost function by computing the following asymptotic expansion as ε goes to zero

$$J(u_\varepsilon) - J(u_0) = \rho^2 G(x_0, n) + o(\rho^2),$$

where $G(x_0, n)$ is the topological gradient given by [1, 5]:

$$G(x_0, n) = -\pi \Delta u_0(x_0) \cdot (n, n) \Delta v_0(x_0)(n, n), \tag{5}$$

and v is the solution of the adjoint problem

$$\begin{cases} \Delta^2 v + \alpha \frac{2f^2}{u^3} v = -\Delta^2 u, & \text{in } \Omega, \\ \frac{\partial \Delta v}{\partial n} = \frac{\partial v}{\partial n} = 0, & \text{on } \partial\Omega. \end{cases} \tag{6}$$

4. Numerical computation

4.1. Split convexity method

However, for such a choice of $p(\cdot)$, equation (2) is strongly nonlinear. For that, we introduce an artificial time variable t and for any fixed number T, we transform our problem to the following time-dependent one:

$$\begin{cases} u_t = -\nabla E_{p(\cdot)}(u), & \text{in } \Omega \times (0,T], \\ u(.,t=0) = f, & \text{in } \Omega, \end{cases} \tag{7}$$

where $\nabla E_{p(\cdot)}(u)$ denotes the Gateaux derivative of $E_{p(\cdot)}(\cdot)$ about u.

After that, we consider *the split convexity* method (see [2, 4]) to solve problem (7). The basic idea of this method is to split the functional $E_{p(\cdot)}$ into a convex part treated implicitly, and a concave one treated explicitly. In our case, we split the energy $E_{p(\cdot)}$ as follows:

$$E_{p(\cdot)} = E_{1,2} - E_{2,p},$$

where

$$\begin{cases} E_{1,2} = \dfrac{c_1}{2}\displaystyle\int_\Omega |\Delta u|^2 \,\mathrm{d}x + \dfrac{c_2}{2}\int_\Omega |u|^2 \,\mathrm{d}x, \\ E_{2,p} = -\displaystyle\int_\Omega \frac{1}{p(x)}|\Delta u|^{p(x)} \,\mathrm{d}x - \alpha\int_\Omega \frac{(f-u)^2}{u} \,\mathrm{d}x + \frac{c_1}{2}\int_\Omega |\Delta u|^2 \,\mathrm{d}x + \frac{c_2}{2}\int_\Omega |u|^2 \,\mathrm{d}x, \end{cases}$$

and c_1 and c_2 are two positive constants. Let τ be the time-step, and write $t_k = k\tau$, $u^k(x) = u(x,t_k)$, with $k = 1, 2, \ldots, [T/\tau] - 1$, then, the resulting discrete time stepping scheme for an initial condition u^0 is is given by

$$\frac{u_{k+1} - u_k}{\delta t} + c_1\Delta\Delta u_{k+1} + c_3 u_{k+1} = -\Delta^2_{p(x)} u_k + c_1\Delta\Delta u_k + c_3 u^k - \alpha\frac{u_k^2 - f^2}{u_k^2}. \tag{8}$$

We use Neumann boundary conditions on $\partial\Omega$:

$$\frac{\partial u_k}{\partial n} = \frac{\partial \Delta u_k}{\partial n} = 0.$$

4.2. Algorithm

The steps of the restoration algorithm are the following:

Algorithm 1 MAIN ALGORITHM

Given f and α.

1) For $p(\cdot) \equiv 2$, compute u and v which solve equations (4) and (6), respectively.

2) Compute the topological gradient $G(x_0, n)$ for each point $x_0 \in \Omega$.

3) Update $q(\cdot)$ as function on $G(x_0, n)$ to obtain a new exponent and solve (8).

In order to update to exponent $p(\cdot)$, we use the following formula

$$p_a(x) = 1 + \exp(-\mu|G(x, n)|), \ \forall x \in \Omega,$$

where $\mu > 0$ is a constant.

4.3. Results

We present here some numerical results in order to show the efficiency of our method. In Figure 1(c), we observe that the blood vessels are well detected by the topological gradient method. The main difference between the noisy image (Figure 1(b)) and restored one (Figure 1(b)) is compared quantitatively by using the SNR and SSIM indicators. The segmented image is presented in Figure 1(d). We remark that all the regions of the image are well identified.

This remark holds true for the synthetic image presented in Figure 2.

5. References

[1] A. Drogoul, *Numerical analysis of the topological gradient method for fourth order models and applications to the detection of fine structures in imaging*, SIAM Journal on Imaging Sciences, vol. 7, pp. 2700–2731, 2016.

[2] D. J. Eyre, *Unconditionally gradient stable time marching the Cahn-Hilliard equation*, MRS Proceedings, vol. 529, Cambridge Univ. Press (1998), p. 39.

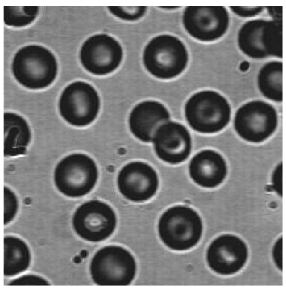
[3] X. Fan and D. Zhao, *On the spaces $L^{p(x)}(\Omega)$ and $W^{m,p(x)}(\Omega)$*, J. Math. Anal. Appl., vol. 263, pp. 424–446, 2001.

[4] K. Glasner and S. Orizaga, *Improving the accuracy of convexity splitting methods for gradient flow equations*, Journal of Computational Physics, vol. 315, pp. 52–64, 2016.

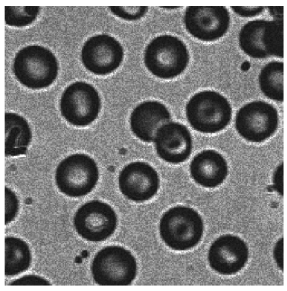
[5] H. Houichet, A. Theljani, B. Rjaibi and M. Moakher, *A nonstandard higher-order variational model for speckle noise removal and thin-structure detection*, submitted to International Journal of Computer Mathematics, p. 31, 2018.

[6] K. Krissian, R. Kikinis, C.-F. Westin, and K. Vosburgh, *Speckle-constrained filtering of ultrasound images*, in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 2, pp. 547–552, 2005.

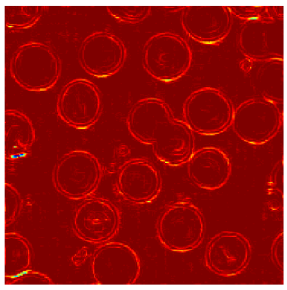
[7] A. Zang and Y. Fu, *Interpolation inequalities for derivatives in variable exponent Lebesgue-Sobolev spaces*, Nonlinear Anal., vol. 69, pp. 3629–3636, 2008.



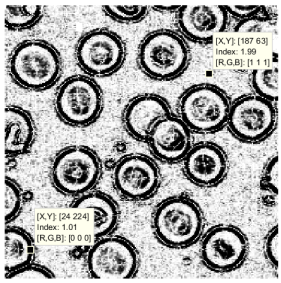
(a) Original image.



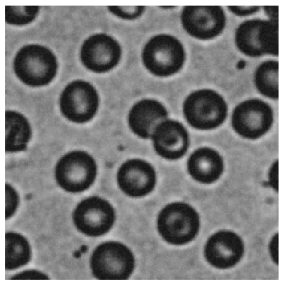
(b) Noisy image (SSIM =0.43, SNR =12.5).



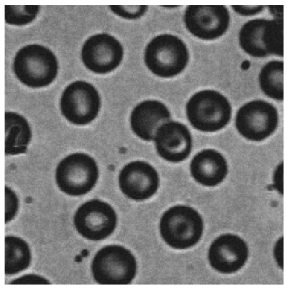
(c)



(d)

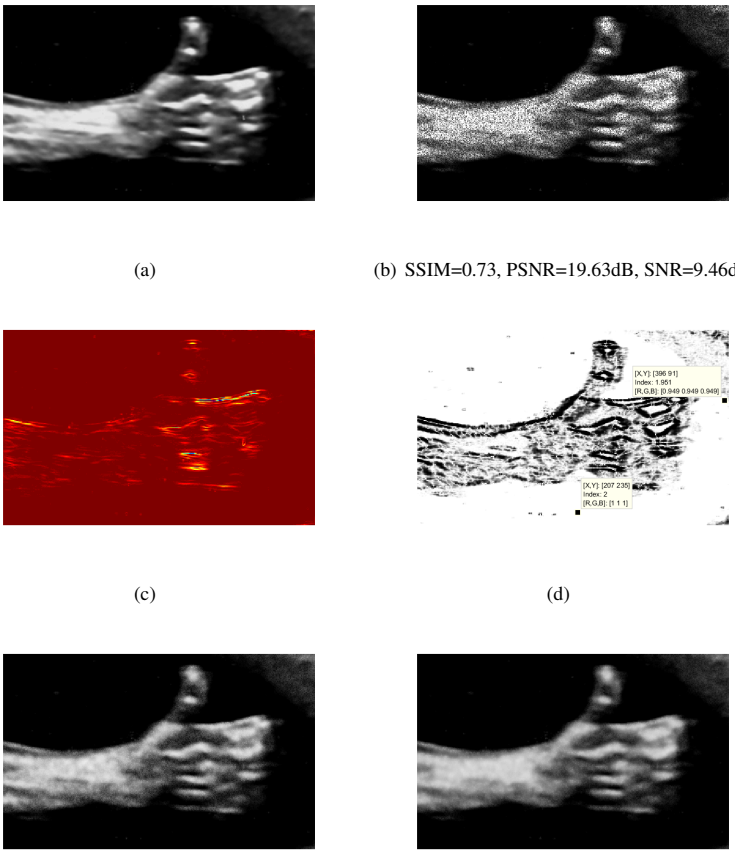


(e) biharmonic model (SSIM=0.76, SNR=18.95dB)



(f) $p(\cdot)$-biharmonic model (SSIM =0.81, SNR =18.93).

Figure 1. *From left to right and top to bottom: (a) Original image, (b) Noisy image, (c) Topological gradient , (d) The variable exponent* $p(\cdot)$ *(e) biharmonic model and (f)* $p(\cdot)$*-biharmonic model.*



(a) (b) SSIM=0.73, PSNR=19.63dB, SNR=9.46dB

(c) (d)

(e) SSIM=0.93, PSNR=27.77dB, SNR=17.47dB (f) SSIM=0.94, PSNR=30.6dB, SNR=20.31dB

Figure 2. *From left to right and top to bottom: (a) Original image, (b) Noisy image, (c) Topological gradient , (d) The variable exponent $p(\cdot)$ (e) biharmonic model and (f) $p(\cdot)$-biharmonic model.*

Discontinuous Galerkin Method (DGM): from classical to isogeometric

hyperbolic problems

Asma Gdhami* — Regis Duvigneau** — Maher Moakher*

* Laboratoire de Modélisation Mathématique et Numérique dans les Sciences de l'Ingénieur
ENIT
Tunis
Tunisie

** INRIA
Université Côte dAzur
Sophia-Antipolis
FRANCE

RÉSUMÉ. L'Analyse isogéométrique (AIG) est une stratégie moderne de résolution numérique des équations différentielles, proposée à l'origine par Tom Hughes, Austin Cottrell et Yuri Bazilevs en 2005. Cette technique de dicretisation est une généralisation de lanalyse par éléments finis classique (AEF), conçue pour intégrer la conception assistée par ordinateur (CAO) et AEF, afin de combler l'écart entre la description géométrique et l'analyse des problèmes d'ingénierie [1]. Le but de ce travail est d'examiner et d'évaluer la méthode de Galerkin discontinue (GD) classique et la méthode de GD dans le contexte isogéométrique (IG) pour résoudre le problème d'advection. Ces deux méthodes sont basées sur le choix d'une base lagrangienne locale et d'une base de Bernstein respectivement.

ABSTRACT. Isogeometric analysis (IGA) is a generalization of classical finite element analysis (FEA) with the main aim of closing the gap between the geometrical description and the analysis of engineering problems. The basic IGA concept, based on the isoparametric paradigm, consisted of using basis functions commonly found in CAD geometries, such as B-spline, to represent both the geometry and the physical fields in the solution of problems governed by partial differential equations (PDEs) [1]. The purpose of this work is to examine and evaluate classical discontinuous Galerkin (CDG)method and discontinuous Galerkin method in the isogeometric context (IGDGM) for solving time dependent, advection problem. These two methods are based on the choice of a local Lagrangian basis and Bernstein basis respectively.

MOTS-CLÉS : Galerkin Discontinu, analyse isogéométrique, flus de Lax-Friedrichs, extraction de Bézier.

KEYWORDS : Discontinuous Galerkin, isogeometric analysis, Lax-Friedrichs flux, Bézier extraction.

1. Introduction & background

The CDG method was originally introduced in 1971 by Reed and Hill [5], for the numerical solution of the nuclear transport PDE problem. Subsequently the method has found far greater use in broad application in large-scale data intensive science and engineering problems. In contrast to the stabilized continuous Galerkin FEM, DG method produce stable discretizations without the need for stabilization parameters. However, this method combine the best properties of the finite volume (FV) method and continuous Galerkin FEM. In the fact, FV method can only use lower degree polynomials, and continuous FEM require higher regularity due to the continuity requirements, therefore, the idea of this method is to decompose the original problem into a set of subproblems that are connected using an appropriate transmission condition (known as the numerical flux). Though DG methods have gained increasing traction in large-scale application modeling and analysis, a shortcoming in the DG methodology is the inability to fully recover complex underlying geometries in the meshing domain. To overcome this problem, we combine IGA with the DG method to get IGDG method. As mentioned before, IGA is a computational technique that improves on and generalizes the classical FE method, the main benefit of thismethod is the exact representation of the geometry in the language of computer aided design (CAD) tools. This simplifies the meshing as the computational mesh is implicitly created by the engineer using the CAD tool. The IGDG method combines the best properties of the FV method and IGA, in fact FV method can only use lower degree polynomials, and IGFE method require to use functions from CAD like Bernstein (B-spline, NURBS) to determine the field where the PDE takes place and to numerically solve it. Therefore, The IGDG method is the DG method formulated on element that exactly preserve the geometries generated by CAD tools. An important property of B-spline in the context of IGA is the ability to perform Bézier extraction. Bézier extraction provides the capability of recovring a local Bernstein-Bézier representation of the geometry from the global B-spline CAD. In this work we will discusses specific details of implementation of IGDG method for the advection problem.

2. Bernstein basis

Definition (Univariate Bernstein).
The Bernstein polynomials of degree p are defined explicitly over the interval $[0, 1]$ by :

$$B_p^k(\zeta) = C_p^k \zeta^k (1-\zeta)^{p-k} \qquad \forall \quad k = 0, ..., p$$

Definition (Multivariate Bernstein).
In order to define Bernstein in higher dimensions, we make use of the tensor product.
Let $p = (p_1, p_2, ..., p_d)$ be a vector in N^d. The d-dimensional Bernstein polynomials are

defined by a tensor product of d univariate Bernstein polynomials with possibly different degrees $p_1, p_2, ..., p_d$ and multi-indices $k_1, k_2, ..., k_d$. Therefore, $\forall \zeta = (\zeta_1, \zeta_2, ..., \zeta_d) \in [0,1]^d$ we get :

$$B_p^k(\zeta) = B_{p_1}^{k_1}(\zeta_1) \otimes B_{p_2}^{k_2}(\zeta_2) \otimes ... \otimes B_{p_d}^{k_d}(\zeta_d)$$

where, the multi-indices $k = (k_1, k_2, ..., k_d)$.

3. B-spline functions

Univariate B-spline functions are defined in parametric space using a so-called vector denoted Ξ, in unit size (1D) is a set of m non-decreasing coordinates : $\Xi = \{\xi_1, \xi_2, ..., \xi_m\}$.
The univariate B-spline function $\mathcal{N}_{i,p}$ of degree p is defined according to the Coxde Boor recursion formula [2] :
for $p = 0$:

$$\mathcal{N}_{i,0}(\xi) = \begin{cases} 1 & if \quad \xi_i \leq \xi < \xi_{i+1} \qquad \forall i = 1, ..., m-1 \\ 0 & otherwise \end{cases} \qquad [1]$$

for $p \geq 1$:

$$\mathcal{N}_{i,p}(\xi) = \Big(\frac{\xi - \xi_i}{\xi_{i+p} - \xi_i}\Big)\mathcal{N}_{i,p-1}(\xi) + \Big(\frac{\xi_{i+p+1} - \xi}{\xi_{i+p+1} - \xi_{i+1}}\Big)\mathcal{N}_{i+1,p-1}(\xi) \qquad [2]$$

In order to define **multivariate B-splines functions in higher dimensions**, we make use of the tensor product.
Let $p = (p_1, p_2, ..., p_d)$ be a vector in N^d and let for all $j = 1, ..., d$, Ξ_j is a $1D$ knot vector defined by :

$$\Xi_j = \{\xi_1^j, \xi_2^j, ..., \xi_{n_1+p_1+1}^j\}$$

Furthermore, we denote the i_j univariate B-spline of degree p_j defined on the knot vector Ξ_j by $\mathcal{N}_{i_j,p_j}(\xi^j)$. Then, with the multi-indices $i = (i_1, i_2, ..., i_d)$, $p = (p_1, p_2, ..., p_d)$ and $n = (n_1, n_2, ..., n_d)$ the d-dimensional tensor product B-spline is defined by :

$$\mathcal{N}_{i,p}(\xi) = \mathcal{N}_{i_1,p_1}(\xi^1) \otimes \mathcal{N}_{i_2,p_2}(\xi^2) \otimes ... \otimes \mathcal{N}_{i_d,p_d}(\xi^d)$$

3.1. B-spline curves

Given n basis functions $\mathcal{N}_{i,p}$, $i = 1, ..., n$ and corresponding control points $P_i \in R$, $i = 1, ..., n$
thus a piecewise-polynomial B-spline curve is given as :

$$\mathcal{C}_p(\xi) = \sum_{i=1}^{n} \mathcal{N}_{i,p}(\xi) P_i$$

3.2. Extracting Bézier curves from B-splines

To decompose a set of B-spline basis functions to its Bézier elements, called Bézier decomposition, a straightforward approach consists in using the knot-insertion procedure p times, for each of the existing interior knots $(\xi_{p+2}, ..., \xi_n)$. Theoretically, the interior knots should have multiplicity of $(p+1)$ to form truly separated Bézier elements. By doing so, the multiplicity of p is sufficient to represent the Bernstein polynomials, which in this context are also referred to as Bézier basis functions. It is important to point out that the Bézier patch is a particular case of B-spline patch, for which the number n of functions (and control points) is equal to $p+1$.

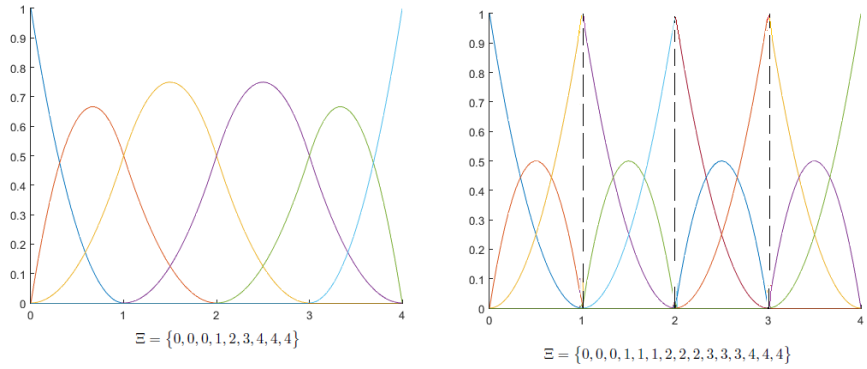


Figure 1. *Bézier decomposition (right) from a quadratic B-spline basis (lef) by knot insertion.*

4. Classical discontinuous Galerkin method

DG is a class of FEM using completely discontinuous basis functions. In contrast to the stabilized continuous Galerkin FEM, DG method produce stable discretizations without the need for stabilization parameters, due to their flexibility in local approximation they offer, together with their good stability properties [3] [4].
In the following, we describe the discretization of the advection problem by the classical DG method :

$$\begin{cases} \partial_t u(X,t) + \nabla.\left(\overrightarrow{c} u(X,t)\right) & = 0 & \forall (X,t) \in \Omega \times [0,T] \\ u(X,0) & = u_0(X) & \forall X \in \Omega \end{cases} \quad [3]$$

where $u(X,t)$ is a scalar quantity transported by a continuous velocity field $\overrightarrow{c}$.
In the DG method, the domain Ω is subdivided into a union of finite number N_{el} of cells $\{D_k\}_{k=1}^{N_{el}}$, such that :

$$\Omega = \bigcup_{k=1}^{N_{el}} D_k \qquad with \quad D_k \bigcap D_l = \emptyset \quad \forall 1 \leq k \neq l \leq N_{el}$$

Thus, we denote by $\mathcal{T}$ a subdivision of Ω into N_{el} elements D_k.

$$\mathcal{T} = \Big\{ D_k, \quad 1 \leq k \leq N_{el} \Big\}$$

So on each cell D_k, the discrete unknown u_h^k is represented as a linear combination of well chosen basis functions of the space of polynomials of degree p. Then, the finite-dimensional subspace $\mathcal{V}_h^p$ is defined as :

$$\mathcal{V}_h^p = \Big\{ \mathbf{v} \in L^2(\Omega) \quad | \, \mathbf{v}_{|_{D_k}} \in P_p(D_k) \quad \forall 1 \leq k \leq N_{el}, \quad D_k \in \mathcal{T} \Big\}$$

where $P_p(D_k)$ represents the space of polynomials of degree up to p defined on the element D_k. By applying Greens formula and introducing the numerical flux $f^\star$ (in the present work, we use the local Lax- Friedrichs recipe), the weak formulation can be written as :
For each element $D_k \in \mathcal{T}$:

$$\int_{D_k} \frac{\partial u_h^k(X,t)}{\partial t} \mathbf{v}_h(X) dX = \int_{D_k} u_h^k(X,t) \vec{c}.\nabla \mathbf{v}_h(X) dX - \int_{\Gamma^k} u_h^k(X,t) \mathbf{v}_h(X) \vec{c}.\vec{n}^k d\Gamma^k \quad \forall t \in [0,T] \quad \forall \mathbf{v}_h \in \mathcal{V}_h^p \quad [4]$$

We denote $\overrightarrow{n}^k$ the outer unit normal to Γ^k of the element D_k.
Therefore, the local problem takes the form of a linear system, which can be written in the following matrix form :

$$M^k \partial_t u^k = \mathbf{R}^k(u^k) + f^\star(u^k) \qquad \forall t \in [0,T] \quad k = 2, ..., N_{el} - 1 \qquad [5]$$

Therefore, in the present work, a $RK2$ and $RK4$ are used for time integration. Because we are focusing on DG schemes, we discuss the limits for the C_{cfl} number when the DG method is used in conjunction with the RK time integration approach. An extra condition on the size of the timestep must also be satisfied, a Courant Friedrichs-Lewy (CFL) condition :

$$| \, c \, | \, \frac{\Delta t}{h_x} \leq \frac{1}{2p+1}$$

where $| \, c \, |$ is the largest wave speed, h_x is the smallest element width, Δt is the length of the time step and p is the degree of the approximating polynomial.

The L^2 error of the numerical approximations are depicted in Fig. (2) which indicate that the rates of convergence are of the type $O(h_x^{p+1})$.

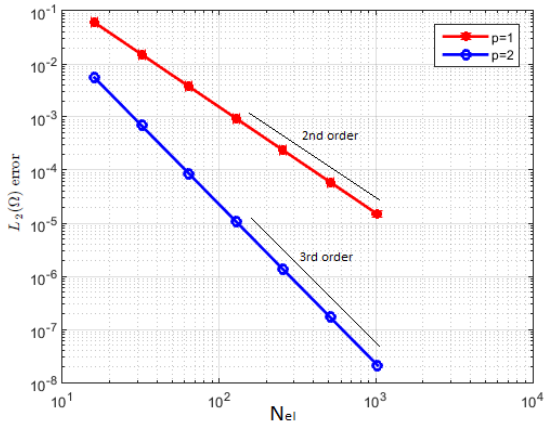


Figure 2. *$1D$ advection problem - L^2-errors from DGFE method in conjunction with RK method for a sinusoidal initial condition and Lax-Friedrichs flux.*

5. Isogeometric discontinuous Galerkin method

In this section, we present a method that combines isogeometric analysis (IGA) with the discontinuous Galerkin (DG) method for solving hyperbolic equations. The basis functions are continuous within each patch, and discontinuous only on patch boundaries. We also highlight that IGA space is local to patches rather than elements, in comparison with FEA. Therefore, the DG application in IGA is a patch to patch relation instead of an element to element. This fact is important to remember, since every time we mention about partitions in the domain, we are referring to patches that consist of elements. In order to apply the IGA methodology, the physical domain Ω is subdivided into patches Ω^e,

$$\mathcal{S}(\Omega) := \left\{\Omega^e\right\}_{e=1}^{N_{el}}$$

such that :

$$\overline{\Omega} = \bigcup_{e=1}^{N_{el}} \overline{\Omega}^e \quad with \quad \Omega^e \bigcap \Omega^l = \emptyset \quad \forall \quad 1 \leq e \neq l \leq N_{el}$$

Then, we define the test functions in the physical domain Ω^e such as :

$$\Phi^{p,q}(x,y)_{|\Omega^e} = \Big(\Phi^{p,q}(x,y)\Big)^e = \Big(B^{p,q}(x,y)\Big)^e = \Big(B^{p,q}\big(T(\xi,\eta)\big)\Big)^e = \big(B^{p,q}(\xi,\eta)\big)^e$$
$$= \Big(B^p(\xi)\Big)^e \otimes \Big(B^q(\eta)\Big)^e = \Big(\widetilde{\Phi}^p(\xi)\Big)^e \otimes \Big(\widetilde{\Phi}^q(\eta)\Big)^e$$

where, T is the transformation of the parametric domain $\widetilde{\Omega}$ to the physical domain Ω :

$$T : \widetilde{\Omega} \longmapsto \Omega, \qquad (\xi,\eta) \longmapsto \big(x(\xi,\eta), y(\xi,\eta)\big)$$

Applying a IGDG method, the solution u is approximated by $u_h \in V^p$, we can postulate the following approximation to the solution :

$$u_h^e(x,y) = \sum_{i=1}^{p+1}\sum_{j=1}^{p+1} \Big(B_i^p(\xi)\Big)^e \Big(B_j^p(\eta)\Big)^e u_{ij}^e$$

where $u_{ij}^e : [0,T] \longmapsto R^2, \quad \forall \quad 1 \leq i,j \leq p+1$ are local unknown functions.
Therefore, the local problem takes the form of a linear system of size $(p+1)^2 \times (p+1)^2$, which can be written in the following matrix form :

$$\mathbf{M}^e \partial_t u^e = \mathbf{R}^e u^e + F^e \qquad \forall t \in [0,T] \quad \forall 1 \leq e \leq N_{el} \qquad [6]$$

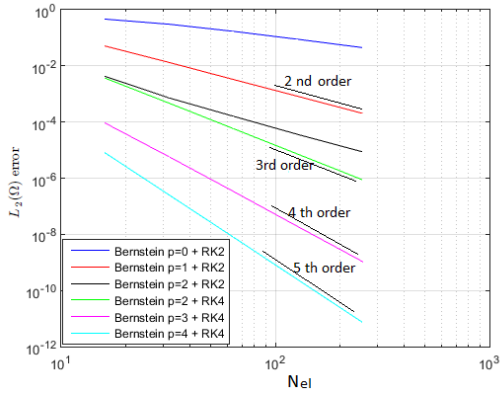


Figure 3. *$1D$ advection problem - L^2-errors for a sinusoidal initial condition, $RK2$ and $RK4$.*

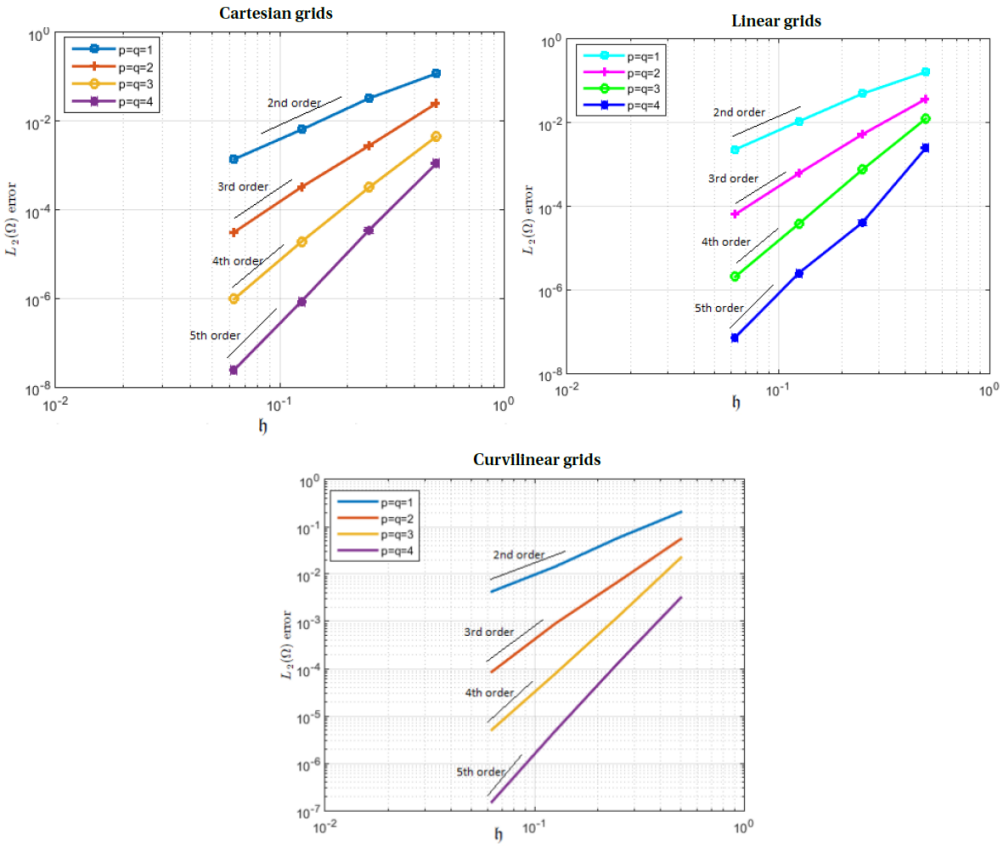


Figure 4. *$2D$ advection problem - L^2 errors for the IGDG method in conjunction with $RK4$ for different grids.*

An optimal convergence rate is found, the method being of order $p+1$ with respect to L^2-norm.

6. Conclusion

As mentioned before, the major reason for using DG methods lies with their ability to provide stable numerical methods for first order PDEs problems, for which classical FEM is well known to perform poorly. However, for geometric partitioning of the computational domain, the DG method uses standard disjoint finite element meshes, each element determines a single subproblem, the solution is calculated separately for each element of the computational mesh. The solution for the whole computational domain is achieved by summing over all the elements of the mesh. In this work a new family of discontinuous Galerkin methods which combines the IGA with the DG method, called IGDG method has been developed for the advection problem, our method takes advantage of both IGA and the DG method. In the fact, DG ideology is adopted at patch level, i.e., we employ the traditional IGA within each patch, and employ the DG method across the patch interfaces to glue the multiple patches. Obviously, due to IGA (NURBS), all conic sections can be represented exactly, thus eliminating the geometrical errors at the beginning.

7. Bibliographie

[1] *Hughes, Thomas JR and Cottrell, John A and Bazilevs, Yuri*. Isogeometric analysis : CAD, finite elements, NURBS, exact geometry and mesh refinement. Computer Methods in Applied Mechanics and Engineering.

[2] *De Boor, Carl*. On calculating with B-splines. Journal of Approximation Theory.

[3] *Shu, Chi-Wang*. Discontinuous Galerkin methods : general approach and stability. Numerical Solutions of Partial Differential Equations.

[4] *Xu, Qinwu and Hesthaven, Jan S*. Discontinuous Galerkin method for fractional convection-diffusion equations. SIAM Journal on Numerical Analysis.

[5] *Reed, William H and Hill, TR*. Triangular mesh methods for the neutron transport equation. Los Alamos Scientific Lab., N. Mex.(USA).

Dynamic resource allocations in virtual networks through a knapsack problem's dynamic programming solution

Vianney Kengne Tchendji*, Kerol Roussin Donteu Djoumessi*, Yannick Florian Yankam*

*Department of Mathematics and Computer Science
Faculty of Science
University of Dschang
PO Box 67, Dschang-Cameroon
vianneykengne@yahoo.fr, djoumessikerol@gmail.com, yyankam@yahoo.fr

RÉSUMÉ. La multitude des services à forte valeur ajoutée offert par Internet et améliorés considérablement avec l'intégration de la virtualisation réseau et de la technologie des réseaux définis par logiciels (Software Defined Networking), suscite de plus en plus l'attention des utilisateurs finaux et des grands acteurs des réseaux informatiques (Google, Amazon, Yahoo, Cisco, ...); ainsi, pour faire face à cette forte demande, les fournisseurs de ressources réseau (bande passante, espace de stockage, débit, ...) doivent mettre en place les bons modèles permettant de bien prendre en main les besoins des utilisateurs tout en maximisant les profits engrangés ou le nombre de requetes satisfaites dans les réseaux virtuels. Dans cette optique, nous montrons que le problème d'allocation des ressources aux utilisateurs en fonction de leurs requetes, se ramène à un problème de sac à dos et peut par conséquent être résolu de façon efficiente en exploitant les meilleures solutions de programmation dynamique pour le problème de sac à dos. Notre contribution considère l'allocation dynamique des ressources comme une application de plusieurs instances du problème de sac à dos sur des requetes à valeurs variables.

ABSTRACT. The high-value Internet services that have been significantly enhanced with the integration of network virtualization and Software Defined Networking (SDN) technology are increasingly attracting the attention of end-users and major computer network companies (Google, Amazon, Yahoo, Cisco, ...). In order to cope with this high demand, network resource providers (bandwidth, storage space, throughput, etc.) must implement the right models to understand and hold the users'needs while maximizing profits reaped or the number of satisfied requests into the virtual networks. From this perspective, we show that the problem of resource allocation to users based on their queries is a knapsack problem and can therefore be solved efficiently by using the best dynamic programming solutions for the knapsack problem. Our contribution takes the dynamic resources allocation as a multiple knapsack's problem instances on variable value queries.

MOTS-CLÉS : Réseau virtuel, allocation des ressources, sac à dos, programmation dynamique, fournisseur de services, fournisseur d'infrastructures

KEYWORDS : Virtual network, ressource allocation, knapsack, dynamic programming, service provider, infrastructure provider

1. Introduction

The limits of the Internet (security, architectural rigidity due to IP protocol, ...) like its resistance to the adoption of new services (such as VOD, telephony over IP, etc) generally known as the phenomenon of Internet ossification [2, 3], led to rethink its architecture. This is how network virtualization was proposed, the idea being the maximum exploitation of physical resources through their sharing and reusability in order to meet the dynamic needs of users ; the integration of the Software Defined Networking (SDN)[1] allowed to better face this resources allocation challenge (known as virtual network embedding problem[9]) through a central equipment called controller, which defines the management policies of the network. This resource allocation is a subproblem of a most global one, commonly known as the Virtual NetWork Embedding (VNE), which is NP-hard to solve[4] because of the number of constraints involved.

Nowadays, since the network virtualization involves the Internet operators to be divided into infrastructure providers (InP) who hold the physical resources and the service providers (SP) who exploit these resources to offer services, both parts must setup appropriate techniques to match their resources allocation with the varied requests of end-users[4]. Thus, techniques such as auctions [6] or game theory can be used to allocate these resources, although they do not always make it possible to decide in all cases.

Our contribution in this paper is the proposition of a 0-1 knapsack-based model for resources allocation in a virtual network environment integrating the SDN architecture. We exploit the dynamic programming solutions of the 0-1 knapsack problem to build an efficient allocation solution within the limits of the available substrate network resources. The goal is to find the best solution that satisfies the majority sides in competition. We also model a dynamic resource allocation as a multiple resource allocation instances with various requests at different times.

The rest of this paper is organized as follow : In section 2, we present network virtualization and SDN paradigms. Section 3 presents a formulation of the resource allocation problem, showing the equivalence with the 0-1 knapsack one, followed by the problem resolution through a dynamic programming solution related to the knapsack problem. A conclusion ends this paper.

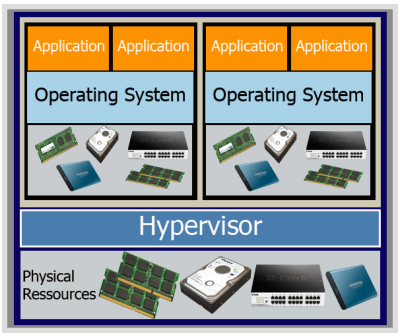
2. Network virtualization and SDN paradigms

Our work environment is made up of several virtual networks under the supervision of a network controller.

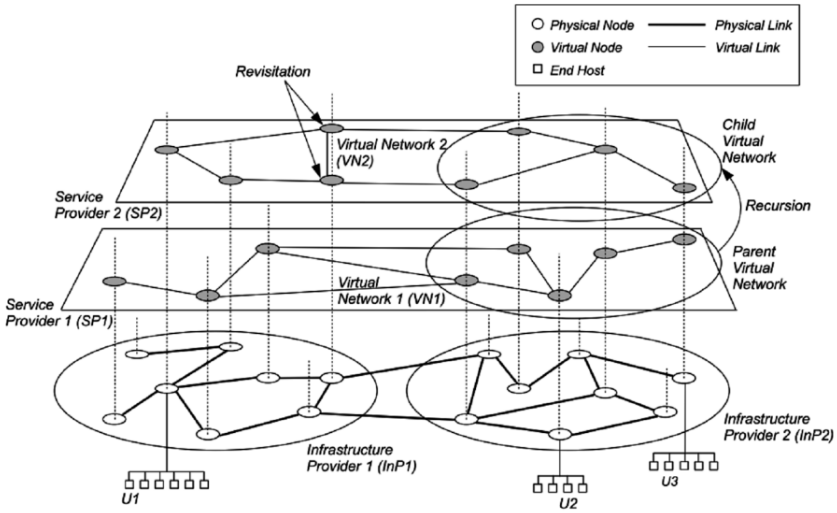
2.1. Virtual networks

A virtual network is a set of virtual devices interconnected by virtual links through a physical infrastructure [3]. In each virtual network, we find components created from a physical component by a special software called hypervisor : these are virtual machines [8] (see figure 1a). Thus, the resources used within a virtual network are provided by the substract network (see figure 1b). Basic physical network resources are provided by an Infrastructure Provider (InP) (see figure 1b). This Infrastructure Provider hires resources from service providers (SP) which creates virtual networks to exploit them. There are three levels of resource allocation : virtual network, SP and InP ; all of these levels are under the supervision of the controller which can initiate cooperation requests with other

InPs when needed. Without this controller, it would not be easy to manage resources with a large virtual network instances.



(a) Virtual machines.



(b) A network virtualization environment.

Figure 1 – Virtualization principles.

2.2. The Software Defined Networking solution

Software Defined Networking (SDN) is a new network architecture paradigm where the control plane is completely decoupled from the data plane for each network equipment [10]. The control plane is a part of network which permits to calculate the network topology or to exchange routing information, while data plane or forwarding plane is a part of network where the packets are commutated. A network controller who have the control plane, defines the network management policies (routing, bandwith allocation, topology discovery,...) and assign it to the equipments. This decoupling allows to deploy a

monitoring plane on standard servers with flexible computing capabilities [11], compared to conventional switches. Thus it opens the opportunity to design an efficient centralized control plane. In addition, the creation of a standardized API (Application Programming Interface) between the control plane and the data plane allows developing network services. The control plane is capable of injecting states in the network elements.

3. The resource allocation problem

3.1. Problem description

Intuitively, resource allocation is a problem of finding the best way to satisfy the most important parts of possible queries from a given set, taking into consideration several constraints involved[5]. It can also consist in satisfying a less important range of requests submitted with the same constraints. There are several problem formulations for virtual network provision [5, 7]. However, these different formulations focus on the allocation of virtual links and bandwidth [7] in a restricted virtual network, while it shall be more general. Another work [12] proposes in the context of the Internet of Things, a power allocation knapsack-based model which approaches the optimal solution, whereas ours allows to reach it using the dynamic programming solution for our resource allocation problem. In this work we look at this allocation problem as a sharing problem, that is, a problem from which we have resources to share among multiple users. The SDN controller ensure the monitoring and the provision of that resources to the end-users ; this controller can also initiate and manage some cooperation between Infrastructure Providers (IP) to get the resources matching the users'constraints. It is therefore an optimization or decision problem that takes as input :

– a set of n applicants. In our context we associate it to the term of user ;

– limited common resource (s) ;

– a common language for expressing preferences and preferences of n users on the resource (s) ;

– a set of constraints on the possible resources to be allocated ;

– an optimization or decision criterion.

As output, we have a resource allocation model, matching the constraints and optimize the criterion. Note that shared resources can be continuous (split), indivisible, discrete or mixed, though in this paper, we consider divisible and shareable resources. This means that a supplier can divide the resources in its basket before sharing them. In this light, resource allocations can be defined and characterized in the following ways :

Definition 1 : Let be a population $P = p_1, p_2, ..., p_n$ of n requests and a set of m resources $R = r_1, r_2, ...$ owned by a resource provider. A resource allocation between these n applicants is a list of n baskets containing the resources $R_i \subseteq R$ obtained by each applicant, matching the following properties : $\cup_{i \in m} R_i = R$ and $\cap_{i \in m} R_i = \phi$

We define the physical infrastructure provider network as an undirected graph $G = (N, L)$ where N is a set of nodes and L is a set of links. Similarly, the virtual network of a service provider is defined as a graph $G^{'} = (N^{'}, L^{'})$ in which $N^{'}$ and $L^{'}$ are the nodes and virtual links built on the substract network of an InP. Since each resource is associated with a constraint, at each node $n \in N$ we also associate a constraint $C^N(n)$ and with each link $l \in L$ a constraint $C^L(l)$. These constraints can represent at the level

of nodes, constraints on the portion of resources available for packets process and delay constraints at the link level.

At the request of a user (see figure 2), the SP submits a request composed of a set of resources that it wants to get from the InP_k. This request consists of a matrix in which the SP specifies its needs.

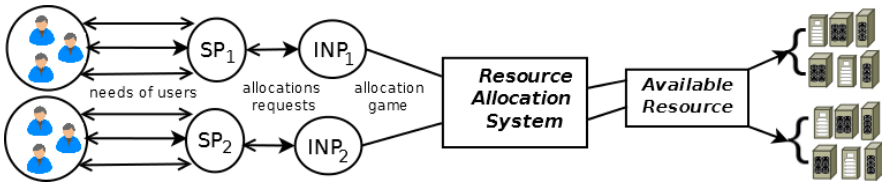


Figure 2 – Ressource allocation process.

This matrix defines the SP's needs (resource and quantity) to satisfy the end users. The physical InP ensures that requested resource quantities do not exceed the total capacity available at the physical network level. In all cases, for a set of requests to satisfy according to given criteria, a set $D = d_1, d_2, ..., d_n$ of n allocation requests to be satisfied, a quantity of available resources $W \in N$ at time t, a quantity $p_i \in N \setminus \{0\}$ of the resource i wanted through the application $d_i \in D$ and criteria $v_i \in N \setminus \{0\}$ to optimize when selecting grant requests to satisfy, the problem can be summarized as :

$$min \sum_{i=1}^{n} x_i p_i \tag{1}$$

or

$$max \sum_{i=1}^{n} x_i p_i \tag{2}$$

under the constraint :

$$\sum_{i=1}^{n} x_i p_i \leq W \tag{3}$$

where W is the total of available resources.

3.2. Correspondence between knapsack problem and that of resources allocation

The knapsack problem consists of determining among a set of objects, a selection with a maximum total value and not exceeding the total permissible weight in the knapsack. This principle is similar to the resource allocation ones, which consists in finding the resource price combination that maximizes the supplier's profits within the limits of available resources for a set of expressed demands. That is to say for each resource allocation problem, there is a knapsack formulation that matches.

Formally, for a set of n demands in resource allocation, we consider a set S of n objects with weight $p_i > 0$ and values $v_i > 0$. We have to find binary variables $x_1, x_2, ..., x_n \in 0, 1$ such as : $\sum_{i=1}^{n} x_i.p_i \leq W$, and $\sum_{i=1}^{n} x_i.v_i$ is maximum. For a variable x_i, value 1 means the element will be put in the knapsack (ie the resource demand i will be supplied)

and 0 means that it will not be selected.
Generally, some constraints are added to avoid singular cases :

– $\sum_{i=1}^{n} p_i > W$: We cannot take all the objects (The SP cannot supply all the needs at the same time) ; that is because in virtual networks, a spare resource must be always available in the substract network for the network recovery.

– $p_i \leq W, \forall i \in 1, ..., n$: no object weight could exceed the knapsack capacity (each resource demand is less than the total capacity of the knapsack ;

– $v_i > 0, \forall i \in 1, ..., n$: each object has a value and brings a gain (the profit collected by the supplier for the allocated resources) ;

– $p_i > 0, \forall i \in 1, ..., n$: any object has a weight (In ressource allocation, there is not null request).
So, to sort out an allocation resource problem, we can use some solutions of the knapsack problem like the dynamic programming solution.

3.3. Solving the resource allocation problem using a dynamic programming solution of the 0-1 knapsack's problem

The dynamic programming resolution method aims at obtaining the optimal solution to a problem by combining optimal solutions with similar, smaller and overlapping sub-problems. Using it involves a recurrent formulation of the problem that will be used to find the optimal solutions. We proceed as follow :

Decomposition of the problem into sub-problems : Let be $M(k, w), 0 \leq k \leq n$ and $0 \leq w \leq W$ the maximum cost that can be obtained with objects $1, ..., k$ of S, and a maximum load knapsack W (We assume that the p_i and w are integers). If we can compute all the entries of this array, then the array entry $M(n, W)$ will contain the maximum computing time of files that can fit into the storage, that is, the solution to our problem. The Cost could be the number of requests or the profit collected.

The recursive equation : Now, we recursively define the value of an optimal solution in terms of solutions to sub-problems. we have two cases :

– **we don't select the object** k : in this case, $M(k, w)$ is the maximum benefit by selecting among the $k-1$ first objects with the limit w ($M(k-1, w)$) ;

– **we select the object** k : $M(k, w)$ is the value of the object k plus the maximum benefit by selecting among the $k-1$ first objects with the limit $w - p_k$.

The recursive equation is then :

$$M(k, w) = \begin{cases} 0 & \text{if } i = 0 \\ M(k-1, w) & \text{if } p_i > w \\ max\{M(k-1, w), v_k + M(k-1, w-p_k)\} & \text{else} \end{cases} \quad (4)$$

This recursive equation result in the dynamic programming algorithm 1 with a space complexity $O(nW)$. We choose this algorithm to perform a bottom-up computation (see figure 3), looking for the optimal solution. This bottom-up computation means that the resource evaluation values will increase gradually during computations. The horizontal red arrows show that calculations are made from left to right ; the vertical red arrow shows that calculations are also done vertically taking into consideration dependency relationships.

Application to resource allocation :

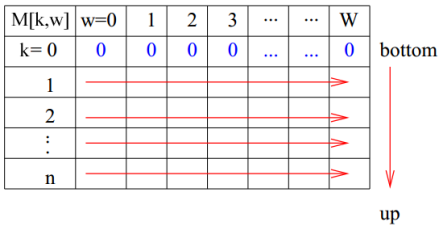


Figure 3 – Bottom-Up Computation principle.

Let us consider a total available resources $W = 11$ in the network. This resource could be the bandwidth, the storage space or throughput. We also consider a set of k applicants with values v_k as the number of requests sent, and weight p_k as the resource quantity corresponding, as given in table 1. Let us assume that all the requests are about the same resource type and they arrive at the same time.

k	weight(p_k)	cost(v_i)
1	1	1
2	2	6
3	5	18
4	6	22
5	7	28

Table 1 – Request sets to an InP for 5 simultaneous arrivals.

Looking for the optimal solution (the maximum requests satisfied by the InP which have resources) with the bottom-up computation, we obtain table 2. M is the different amounts of available resources. Each n-uplet $\{a_{i1}, a_{i2}, ...a_{in}\}$ represents the fact that the element a_{in} have dependencies with the previous elements $a_{i1}, a_{i2}, ...a_{in-1}$; this means that according to the recursive equation 4, the resource computation for a_{in} is linked to those of $a_{i1}, a_{i2}, ...$ and a_{in-1}. For example, to obtain the cost for $M[4, 11]$ which is also written {1,2,3,4}, the computations made are :
$M[4, 11] = max\{M[4-1, 11], v_4+M[4-1, 11-p_4]\} = max\{M[3, 11], 22+M[3, 11-6]\} = max\{25, 22 + 18\} = max\{25, 40\} = 40$

M	0	1	2	3	4	5	6	7	8	9	10	11
$\emptyset$	0	0	0	0	0	0	0	0	0	0	0	0
{1}	0	1	1	1	1	1	1	1	1	1	1	1
{1,2}	0	1	6	7	7	7	7	7	7	7	7	7
{1,2,3}	0	1	6	7	7	**18**	19	24	25	25	25	25
{1,2,3,4}	0	1	6	7	7	18	22	24	28	29	29	**40**
{1,2,3,4,5}	0	1	6	7	7	18	22	28	29	34	35	**40**

Table 2 – Bottom-up costs evaluation.

Table 2 shows that the maximum request numbers could be up to 40 UoC (Unit of Cost) with this example. Then, the optimal solution is {4,3} based on algoritm 1 and the applicants number 3 and 4 would be satisfied by the InP firstly ; the provided resources will be used during a time before they are allowed to other applicant. Within this period of time, other applicant requests are saved in a waiting mode. When the previously allocated

resources are totally or partially released, other applicant requests could be satisfied. For each allocation game, the dynamic programming solution is used with various data at different times. This allocation process is presented in figure 4.

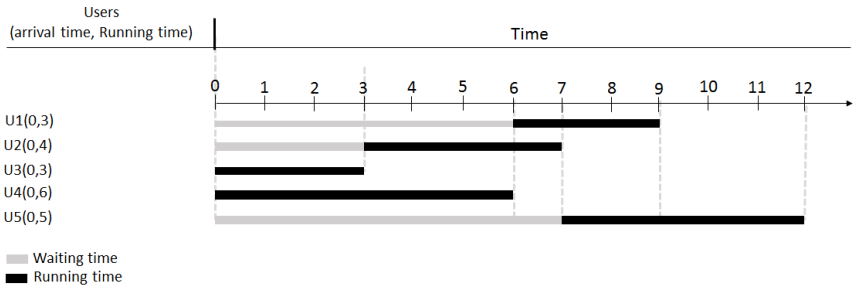


Figure 4 – Gannt chart for a set of five applicants for resources.

Depending on the objectives targeted by the InP (maximizing the number of requests fulfilled, maximizing the economic benefit derived from the allocation of resources), the previous example can be adapted.

4. Conclusion

In this paper, we have presented a knapsack-based dynamic resource allocation model that allows Infrastructure Providers (InP) in a network virtualization environment to select the most suitable users'requests meeting the aims of this InP. Our aim was to provide an efficient decision mechanism to face challenging difficulties encountered by the InP with the multiple requests of end-users or Service Providers. We propose a solution based on a knapsack dynamic programing solution to choose the most suitable users to satisfy. We managed dynamic allocations as multiple simple resource allocation instances occuring at different times.

In an upcoming future, we intend to work on a decision mechanism taking into consideration important constraints as the fidelity of the user to an InP. It would not be suitable that a new customer, even providing a good profit to an InP, is chosen in replacement of an older and regular customer.

5. Bibliographie

[1] Jain, Raj, Paul, Sudipta, « Network virtualization and software defined networking for cloud computing : a survey », *Mobile Networks and Applications*, Vol. 51, N° 11, p. 24-31, 2013.

[2] Niebert, Norbert, El Khayat, , Baucke, Stephan, Keller, Ralf, Rembarz, René, Sachs, Joachim, « Network virtualization : A viable path towards the future internet », *Wireless Personal Communications*, Vol. 45, N° 4, p. 511–520, 2008.

[3] N.M. Mosharaf Kabir Chowdhury, RRaouf Boutaba, « A survey of network virtualization », *Elsevier, IEEE*, Vol. 54, p. 862-876, 2010.

[4] Haider, Aun and Potter, Richard and Nakao, Akihiro, « Challenges in resource allocation in network virtualization », *20th ITC Specialist Seminar*, Vol. 18, N° 2009, 2009.

[5] Mohamed Said Seddiki, « Allocation dynamique des ressources et gestion de la qualité de service dans la virtualisation des réseaux », *PhD thesis, Université de Lorraine*, 2015.

[6] Amraoui, Asma and Benmammar, Badr and Krief, Francine and Bendimerad, Fethi Tarik, « Négociations à base d'Enchères dans les Réseaux Radio Cognitive », *Nouvelles Technologies de la répartition-Ingénierie des protocoles NOTERE/CFIP 2012*, 2012.

[7] Zhu, Yong and Ammar, Mostafa H, « Algorithms for Assigning Substrate Network Resources to Virtual Network Components », *INFOCOM*, Vol. 1200, N° 2006, p. 1–12, 2006.

[8] Popek, G. J., Goldberg, R. P, « Formal requirements for virtualizable third generation architectures », *Communications of the ACM*, Vol. 17, July, 1974.

[9] Fischer, Andreas and Botero, Juan Felipe and Beck, Michael Till and De Meer, Hermann and Hesselbach, Xavier, « Virtual network embedding : A survey », *IEEE Communications Surveys & Tutorials*, Vol. 15, N° 4, p. 1888–1906, 2013.

[10] Kreutz, Diego and Ramos, Fernando MV and Verissimo, Paulo Esteves and Rothenberg, Christian Esteve and Azodolmolky, Siamak and Uhlig, Steve, « Software-defined networking : A comprehensive survey », *Proceedings of the IEEE*, Vol. 103, N° 1, p. 14–76, 2015.

[11] Kim, Hyojoon and Feamster, Nick, « Improving network management with software defined networking », *IEEE Communications Magazine*, Vol. 51, N° 2, p. 114–119, 2013.

[12] Morimoto, Naoyuki, « Power allocation optimization as the multiple knapsack problem with assignment restrictions », *Network of the Future (NOF), 2017 8th International Conference on th*, p. 40–45, 2017.

A. Knapsack dynamic programing algorithm

Algorithm 1 provide the optimal solution on a set of objects for the knapsack problem, and also indicates which subset gives this optimal solution. From line 1 to 15, we compute the maximum requests to satisfy. From line 16 to 21, the algorithm select the applicants to provide with resources.

B. A practical example of resource allocation with succeeding request arrivals of 8 applicants to the InP

In this example, we suppose that the applicant requests reach the InP at different times. So, those requests are satisfied successively. When new requests occur from another applicant, preceding allocated resources can be divided to provide the other ones.

let us assume a total available resources $W = 10$ in the InP network. We also consider a set of k applicants with values v_k as in the previous example, as given in table 3. Let us assume that all the requests are concerned with the same resource type and they arrive successively according to time.

We suppose that requests from the applicants number 1, 2 and 3 come first. The computation of the maximum satisfied requests will be 70 UoC (see table 4). This means that the optimal solution is {3,4}. In case of competition, applicants 3 and 4 would be selected before the others.

When other applicant requests will reach the InP, another computations will be made to choose the most suitable user to provide with resources. Table 5 illustrates the computations done for the requests coming at the time 6, and result in a maximum of 110 requests that could be satisfied by the InP. The applicant numbers 1 and 2 correspond respectively to numbers 3 and 4 in table 3.

Algorithm 1: knapsack

Data: p,v,n,M
Result: A maximum benefit on objects p

Let $M[0..n, 0..W]$ be a new table ;
Let $x[1..n]$ be a new table ;
begin
  for $w = 1$ **to** W **do**
    $M[0, w]$=0 ;
  for $k = 1$ **to** n **do**
    $M[k, 0]$=0 ;
  for $k = 1$ **to** n **do**
    for $w = 1$ **to** W **do**
      if $p[k] > w$ **then**
        $M[k, w] = M[k - 1, w]$;
      else if $M[k - 1, w] > v[k] + M[k - 1, w - p[k]]$ **then**
        $M[k, w] = M[k - 1, w]$;
      else
        $M[k, w] = v[k] + M[k - 1, w - p[k]]$;
  w=W ;
  for $k = n$ **to** 1 **do**
    if $M[k, w] == M[k - 1, w]$ **then**
      $x[k] = 0$;
    else $x[k] = 1$; $w = w - p[k]$;
  return x ;

k	weight(p_k)	cost(v_i)	Arrival time
1	5	10	0
2	4	40	
3	6	30	
4	5	50	6
5	4	60	
6	3	80	13
7	5	20	16
8	7	30	

Table 3 – Request sets to an InP for 8 applicants.

M	0	1	2	3	4	5	6	7	8	9	10
$\emptyset$	0	0	0	0	0	0	0	0	0	0	0
{1}	0	0	0	0	0	10	10	10	10	10	10
{2}	0	0	0	0	**40**	40	40	40	40	50	50
{3}	0	0	0	0	40	40	40	40	40	50	**70**

Table 4 – Bottom-up costs evaluation with applicants coming at the time 0.

M	0	1	2	3	4	5	6	7	8
$\emptyset$	0	0	0	0	0	0	0	0	0
{1}	0	0	0	50	50	50	50	50	50
{2}	0	0	60	60	60	110	110	110	**110**

Table 5 – Bottom-up costs evaluation with applicants coming at the time 6.

with regards to what is stated above, the results of table 6 are obtained for applicants number 7 and 8 coming at the time 16.

M	0	1	2	3	4	5	6	7	8	9
$\emptyset$	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	20	20	20	20	20
2	0	0	0	0	0	0	0	30	30	**30**

Table 6 – Bottom-up costs evaluation with the applicants coming at the time 16.

Gant chart of the figure 5 presents the resource allocation order of all different applicants, mapping with their requests. It considers that the running time of each applicant is proportional to its weight p_k.

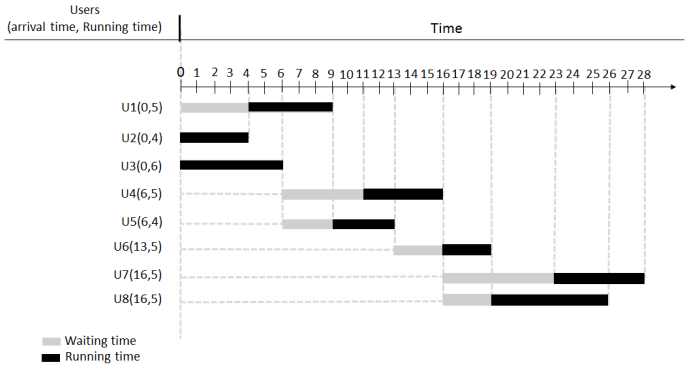


Figure 5 – Gannt chart for 8 sequential arrivals.

C. An enhanced example of resource allocation with 24 applicants and 150 UoC of resources to the InP

In this example, we enhance the resource allocation scenario presented in appendix B.

let us assume a total available resources $W = 150$ in the InP network. We also consider a set of $k = 24$ applicants with values v_k as given in table 7. The column A.t. (t) is the arrival time represented as t. Let us assume that all the requests are concerned with the same resource type and they arrive successively according to time t. Such configuration provide a maximum of 420 satisfied requests with the following provision scheme for the users at $t = 0$: users' requests 2, 3 and 5 will be satisfied firstly, then users 4 and 1. In the same way, at time $t = 20$, users' requests 11 and 13 will be satisfied before 12, resulting a maximum requests number of 808. At $t = 30$, the maximum satisfied requests is 543 and the resource allocation process will consider the users 18 and 19 before user 20. These maximum satisfied requests are computed using the algorithm 1. In each period of the allocation process, this maximum request number can be increased with the running requests of the preceding period. The Gannt chart is provided by the figure 6.

k	weight(p_k)	cost(v_i)	A.t. (t)	k	p_k	v_i	A.t. (t)	k	p_k	v_i	A.t. (t)
1	103	200		9	62	120	18	17	90	210	27
2	30	101		10	45	138		18	16	187	
3	54	174	0	11	35	350		19	107	356	30
4	101	250		12	92	670	20	20	88	231	
5	46	145		13	110	750		21	42	199	33
6	22	80	13	14	63	680	21	22	61	225	
7	6	20	16	15	102	110	23	23	115	165	38
8	14	30		16	87	220	25	24	84	194	

Table 7 – Request sets to an InP for 24 applicants.

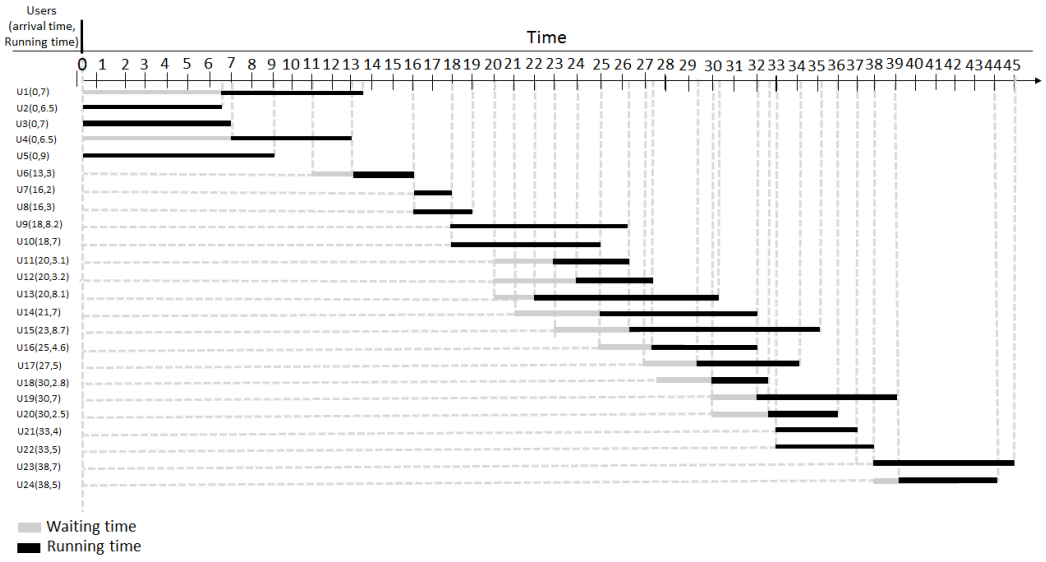


Figure 6 – Gannt chart for 24 sequential arrivals.

Secure and Energy-Efficient Geocasting Protocol for Hierarchical Wireless Sensor Networks

Vianney Kengne Tchendji*, Blaise Paho Nana*, A. Yvan Guifo Fodjo*

*Department of Mathematics and Computer Science
Faculty of Science
University of Dschang
PO Box 67, Dschang-Cameroon
vianneykengne@yahoo.fr, blaisepaho@gmail.com, yvanguifo@gmail.com

ABSTRACT. Wireless sensor networks are now used in many applications (medical, agricultural, military, fire detection, the Internet of Things, etc.) needing a high security level and better energy saving. In this paper, we present a geocasting protocol, secured by the use of elliptic curves as a generator of secret keys. For this purpose, we begin by presenting a technique of grouping sensors in clusters or zones, then in cliques. This clustering creates a virtual architecture that facilitates the routing, the management of the network, and reduces the energy expenditure. We then propose for this architecture, our secure geocasting protocol, energy efficient and with low memory needs, which makes good use of the built architecture. The geocasting uses the skills of the sink to quickly and more efficiently reach geocast regions. The security aspect of our protocol, based on elliptic curves, offers a good level of security that can also limit damages in case of eventual attacks, facilitates the distribution and keys refreshment (a recurrent problem) without ever having to pass these keys in the network, and remains effective against several types of classic attacks such as passive or active listening, identity theft, black hole, data replication, etc .; but remains vulnerable to the problem of frequency jamming.

RÉSUMÉ. Les réseaux de capteurs sans fil sont de nos jours utilisés dans plusieurs domaines (médicale, agricole, militaire, surveillance, Internet des objets, etc.) à fortes exigences en matière de sécurité et d'économie d'énergie. Dans ce document, nous présentons un protocole de géocasting, sécurisé grâce à l'utilisation des courbes elliptiques comme générateur de clefs secrètes. Pour ce faire, nous commençons par présenter une technique de clustérisation de la région d'intérêt en zones ou clusters, puis en cliques. Cette clustérisation met en place une architecture virtuelle qui facilite le routage, permet une meilleure gestion du réseau, et réduit la consommation énergétique. Nous proposons ensuite pour cette architecture, notre protocole de géocasting, économe en énergie et en espace mémoire, qui utilise à bon escient l'architecture sous-jacente. Le géocasting utilise les atouts du sink pour atteindre plus rapidement et efficacement les régions de géocast. L'aspect sécuritaire de notre protocole, basé sur les courbes elliptiques, offre un bon niveau de sécurité qui permet de limiter les dégâts en cas d'éventuelles attaques, facilite la distribution et le rafraîchissement des clefs (un problème récurrent) sans jamais avoir à faire transiter ces clefs dans le réseau. Il reste efficace face à plusieurs types d'attaques classique tels que l'écoute passive ou active, l'usurpation d'identité, le trou noir, la réplication de données, etc.; mais reste vulnérable au problème de brouillage de fréquence.

KEYWORDS : Wireless sensor networks, geocasting, hierarchical clustering, virtual architecture, clique, cluster, elliptic curve, security.

MOTS-CLÉS : Réseaux de capteurs sans fil, géocasting, clustérisation hiérarchique, architecture virtuelle, clique, cluster, courbes elliptiques, sécurité.

1. Introduction

Several progress have been achieved this recent years in the fields of microelectronics, micro technology, and wireless communication technologies. Theses advances enable the production of sensors of small sizes, at low cost and at the forefront of the technology. Yet, they are still subject to several constraints: limited energy, low computing power, small memory storage, etc. Massively deployed in a given area (usually for monitoring reasons) which is most of the times hostile to humans, they are able to organize and self-configure into a wireless network called **W**ireless **S**ensor **N**etwork (WSN). This type of network requires hundreds or even thousands of units that are mass-produced in an environment where testing is a luxury. Each unit is usually equipped with a single use battery, irreplaceable and non-rechargeable for various reasons (cost of batteries replacement, hostility of the area to be monitored, etc.).

Many WSN-based applications make use of geocasting to send messages to sets of receivers in a well-defined geographic area [5, 7, 14]. This technique is of a particular interest especially for civil security. During a sinister or a natural disaster for example, police forces and firefighters may need a geocasting mechanism to join any other actor in an area of the disaster. This technique is also used in agriculture when watering a specific area, or even in military applications, where information must often be provided to all soldiers located in a given area. In addition, it can provide a commercial use, especially to allow anyone passing near a store to immediately receive advertising information.

Promoted for their ease of deployment, WSNs faces many challenges, some of the most important are related to the energy consumption, security and reliability of information circulating in such a network. We are presently seeing WNS-based applications flourishing and moving towards the Internet of Things (IoT) [5, 15]. Thus, for military or medical applications, the need to provide a reliable security solution seems important or even compulsory [3]. Even if the context has evolved, from the machine not connected to wired and wireless networks, the goal of security has always been the same overall, namely to provide basic security services such as authentication, control and security access, confidentiality, integrity, availability, etc. However, because of the characteristics of WSNs (lack of preset infrastructure, dynamic topology, large number of sensors, limited physical security, modes and deployment areas offering multiple possibilities of attacks, etc.) coupled to the inherent constraints of sensor nodes, securing sensor networks is nowadays the source of many scientific and technical researches [3, 4, 11, 12]. Earlier research has shown that the security solutions offered for wireless networks (mobile and ad hoc), especially those based on the use of public cryptography key are very heavy for WSNs [8]. It is therefore important that the security solutions implemented should be the least expensive in terms of resource consumption and aim to reduce delays, number of communications, and the bandwidth occupancy. In other words, these solutions should provide maximum security while preserving the lifetime of the sensors.

In this paper, we are interested in the geocasting and the security problems in WSN. The geocasting is energy-efficient, fast and with less flooding thanks to the use of the Wadaa et al.'s virtual architecture protocol [1] and the Sun et al.'s secure clique formation protocol [12]. We also seek to define inexpensive energy mechanisms and solutions for wireless sensor networks that take into account the relative weaknesses of defense of an autonomous network. For this purpose, we apply energy-saving symmetric cryptography solutions based on elliptic curve-based asymmetric cryptography mechanisms to secure the geocasting protocol that we propose.

The rest of this paper is organized as follows: section 2 presents the construction of the virtual architecture; section 3 presents our geocasting protocol; we secure the whole protocol in section 4; and we conclude after the simulation results of section 5.

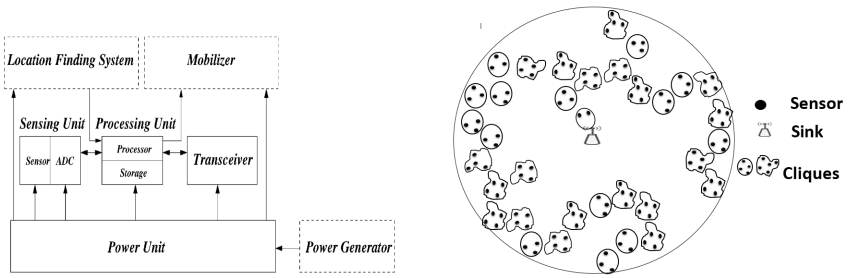
2. Setting up the virtual architecture of the network

2.1. Hypotheses

The work done in this paper is subject to the following assumptions: each sensor of the network is static, has an unique *ID*, a GPS (or ASP [9]) and is loaded with an elliptic curve that will help secure data in transit; it is also able to estimate its residual energy; its internal clock is synchronized with that of the base station (sink) so that it can wake up at defined time intervals to collect informations and remain idle the rest of the time; all the deployed sensors are connected and are aware of the number "c" of coronas and the number "s" of angular sections of the architecture; Adding or removing a sensor is allowed but it is a rare event. The base station has the possibility to broadcast messages in the network either at different radii coverage or at different angles and constitutes the only reliable and incorruptible entity of the network; Time is slotted and each message sent by a sensor is received by the sensors in its vicinity within a slot.

2.2. Sensor's model

The sensor is the basic element of any WSN. It must have a sensing unit, a radio module to receive or transmit data, a storage and computing module, all powered by a small battery. This equipment can be completed according to the requirements of the application. Figure 1a shows an example of a sensor with other optional equipment (framed with dotted lines).



(a) Generic model of a wireless sensor.

(b) Deployment of sensors and formation of disjointed cliques.

Figure 1: Deployment of sensors and cliques' formation.

2.3. Formation of cliques

After a massive deployment around the base station also called the sink (figure 1b), it is time for sensors to self-configure, collect environmental data and route them to the sink. To do this, we start by grouping the sensors into small disjoined groups using Sun et

al. protocol [12]. This technique presents a network partitioning protocol using the cluster first (CF) approach. This approach requires the fact that each node accepts membership in a group before the leader's election. This is how the network is partitioned into cliques in which each sensor is at a single hop from any other sensor of the same clique. This protocol has the following properties: it is essentially distributed and each node calculates its clique's membership by sending messages to its direct neighbors; when the partitioning algorithm terminates, the participating nodes that do not follow the specifications of the protocol (sending superfluous or conflicting messages) are systematically identified and removed from each clique; at the end of the partitioning algorithm, the network consists of disjoint cliques and each node has a clear view of the member nodes of its membership clique. After the cliques' formation as on figure 1b, the formation of the zones follows.

2.4. Formation of zones

Consider here that the sink is a special node capable of performing omnidirectional transmissions with certain radii for the formation of concentric discs (or coronas) and directional transmissions at certain angles for the formation of angular sections. Once deployed in the supervised zone, the sensors, each having a unique identifier, are grouped in clusters or zones (as described in [1]) according to the angular sectors and coronas (figure 2). In this way, the intersection of a corona c and an angular sector s constitutes the zone (c, s). In addition, since the network is parse, it will be important to identify empty zones or clusters (section 2.5). This will allow the sink node to have an overview of the areas actually covered by the sensors. Note that when a clique is shared between several zones, after the CH1 (clusterhead of a clique) election, the sensors of this clique will behave as part of the zone where their CH1 is located. Now, let's build the architecture and proceed to the CH1 and CH2 (clusterhead of a zone) elections.

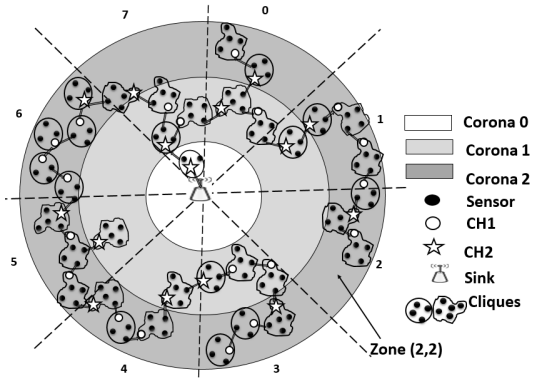


Figure 2: Formation of zones.

2.5. Hierarchical structure, detection of empty zones and election of clusterheads

The empty zones discovery process is similar to the one presented in one of our previous work [13] with some differences and inspired by [10]. In this process, there's a part executed by the sink, and a part performed by the sensors.

Base Station Algorithm: Before running the algorithm, the sink node periodically broadcasts an alert message across the entire monitored area, specifying to the sensors the date the discovery algorithm will begin. All the sensors are awake and the sink node initiates the detection of the empty clusters by spraying in the first corona a $Detect(-1,-1)$ message containing a variable $NumberOfHops$ initialized to 0. This variable allows each sensor to evaluate its distance (in terms of the number of hops) from the base station; It will also serve as a selection criteria for the different clique (CH1) and cluster (CH2) leaders. Initially, each sensor initializes its $NumberOfHops$ variable to $+\infty$. The network being connected, it is sure that at least one sensor of the first corona will receive the message $Detect$. During the algorithm, the sink listens to answers coming from the area of interest. At each message reception, the sink node maintains two routing tables: h and $relay$. The cells of table h initialized to 0 contains a bit 1 in the cell (i,j) if the sink has received a message from the cluster (i,j) or a 0 if not; and the relay table contains in its cell (i,j) the coordinates of the relay cluster (or relay zone) of the zone (i,j), i.e. the closest zone to the sink in the vicinity of cluster (i,j) through which its data passes to reach the sink. This allows the base station to have an overview of the areas covered in the network.

Algorithm of a sensor: When the $Detect$ message arrives especially in a cluster of the first corona and in general in any other cluster, the sensors execute approximately the same algorithm. This algorithm is inspired by the one proposed in [13]. When a sensor receives such a message, it checks that it comes from a neighboring cluster or a neighboring clique. In the case it is the first to receive this message (i.e. it has not yet received a $DetectTimer$ message), he broadcasts a message $DetectTimer$ containing the election's beginning date of the leader of its clique, and retransmits the $Detect$ message to allow the discovery of other zones once it has incremented the $NumberOfHops$ variable.

Election of leaders: Within a clique, the election of CH1s takes place in two steps requiring only two message transmissions (Head1 and Head2). Each sensor having a residual energy greater than the threshold energy Es calculates and starts the countdown of a timer that lasts $(1 - \frac{1}{NumberOfHops} + \frac{e^{-ID}}{\Lambda})$ slots. When this timer expires, it broadcasts a Head1 message to propose itself as leader of the clique if it has not yet received one. The sender and receivers of the previous Head1 message calculate and arm a second timer (which lasts $(\frac{1}{Er} + \frac{e^{-ID}}{\Lambda})$ slots) at the end of which, only the CH1 will be able to broadcast the message Head2 preventing at the same time any sensor that receives this message to diffuse another. These timers are calculated such that the priority should be given to the sensor having the most energy, the one closest to the base station, and the *ID* is used to decide in case of concurrent access. Note that after the $Detect$ message is broadcast by a sensor, it specifies the election start date within its clique. And when the election begins within its clique, it is after a slot that it will start in the clique which just received the message $Detect$. After more than 2 slots, if within a clique there is no longer message transmission, then the election of CH1 is over and we can start the election of the CH2 (leader of a given zone).

When the election of CH1s ends within a zone, there is an immediate broadcast of a Head3 message from the elected CH1. At each reception of a Head3 message, the elected CH1s memorize the parameters of the transmitter and make updates to keep each time the best candidate for the CH2's post i.e. the one closest to the sink. Note that with the CH1 already in place, the communication between two direct neighbors needs at most 2 slots. A CH1 knows that the election is over if after more than 2 slots it no longer receives a Head3 message. This mechanism allows the various elected CH1 and CH2 to know the number

of CH1 in their area. After the election of the CH2 of the zone, the latter broadcasts a message Head4; The neighboring CH1s will receive the Head4 message in one slot and will immediately rebroadcast the Head4 message. Any other CH1 that receives the message Head4 keeps the sending clique as its relay clique i.e. the one through which its messages will pass to reach the sink.

Hierarchical structure and packet routing: We have thus constructed an hierarchical structure with several levels. The bottom of this structure consists of ordinary sensors grouped into cliques. Each clique is driven by a leader of the first level: the CH1. The CH1s are managed by super clusterheads (CH2) which also report to the sink. In a given clique, when data are collected at the level of the ordinary sensors, these data are transmitted to their CH1 which will then be responsible for retransmitting them to their relay clique. This message can be received either directly by the CH1 of the relay clique at the best of the cases, or by a member of this clique in which case it transmits it to its leader (the CH1 of the relay clique). When the CH1 of the relay clique is also the CH2 of the zone, the message is transmitted to the relay zone either to a CH1 of a clique of the relay zone (at best), or to an ordinary sensor of a clique of the relay zone. This is how messages are routed from cliques to zones and zones after zone until they reach the sink.

3. Scenario of geocasting

Here we implement a procedure that allows any sensor in the network to communicate with sensors of other areas. Note that sensors have no way of keeping a global view of the network because of their limited means. On the other hand, the communication between two zones, though neighboring or geographically close zones, is not always obvious because of the empty zones bypassing during the discovery. We assume that sensors know only their direct neighboring cliques (good use of memory) and that a sensor of clique A has a data D that it wants to send to the sensors of a region B defined by a set of GPS coordinates. For this purpose, the sensor begins by transmitting this data to its leader. The CH1 in procession of this data will be in several situations that we describe below:

The CH1 can directly reach the sensors of the B region: since the CH1 has a restricted view of its neighbors, it can know if it can reach the CH1 in charge of the region B or at least one sensor of this region. In this case, it broadcasts the message toward this region. Any sensor in this region that receives the message informs its leader, which will also inform all sensors in the given region.

The CH1 cannot directly reach the sensors of region B: In this case, the CH1 processes the geocast message like any other normal data message collected i.e. this message must go from cliques to cliques and from zone to zone until it reaches the sink as described in the above method. If along the road, a CH1 can reach a sensor of the region B, this brings us back to the first case and the preceding algorithm is executed. Otherwise, the message will necessarily reach the base station. This leads us to the third case.

The message reaches the base station: In this case, the sink will use its directional transmitting antenna to directly send the information in the region B. For this, it performs directional transmissions to each CH2 in the region B by modulating the transmission power so that the beam reaches the target region and also by modulating the transmission angle to not divulge the message to sensors that are not concerned. This is illustrated by figure 3. The CH2s will send the message to the CH1s in the geocast region which will also transmit it to the concerned sensors.

Because the transmission is not secure, sensors on the passage of the beam can listen to

the message which is not intended for them and try to send a useless message back to the sink or to a neighboring clique. To avoid this, any sensor that receives the message from the base station verifies that it is in the geocast region to know if this message is also intended for it.

In case of multi-geocasting i.e. for a message intended for several regions, this protocol is easily adaptable by replacing for example the geocast region with a list of regions.

Up to here, our protocol has no security mechanism. The purpose of the next section is to secure all the exchanged messages of this protocol from the clustering stage.

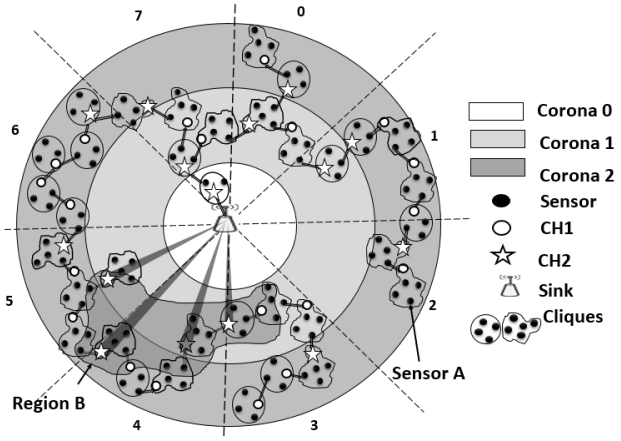


Figure 3: Multizones geocasting.

4. Securing the protocol

4.1. Elliptic Curve Integrated Encryption Scheme (ECIES)

The Elgamal protocol is rarely used directly with elliptic curves. Before encrypting a message, it must first be converted to a point on the elliptical curve used. There are different techniques that exist, but the conversion requires more calculation. In several researches, elliptic curves are commonly used to establish a shared key between the 2 parts of a conversation, after which, a symmetric cryptography algorithm is used to secure the communication between them [4, 8].

The ECIES protocol is indeed a standardized version of Elgamal. Suppose that Alice wants to send a message M to Bob in a secure way, they must first have all the following information: Key Derivation Function (KDF: A key derivation function that allows to generate several keys from a secret reference value); Message Authentication Code (MAC: Code transmitted with the data in order to ensure the integrity of the data); a symmetric encryption algorithm (SYM); an elliptic curve $E(Fp)$ used with the generator point G with $Ord_p(G) = n$; the public key of Bob $K_B = k_B.G$ where $k_B \in [1, n-1]$ is his private key. In this protocol, the critical value is k with which Bob can compute $Z = k.K_B$, and generate the pair $(k1, k2)$ that is used to decrypt and authenticate the message. Due to the difficulty in solving the discrete logarithm problem, Alice can send $R = k.G$ without any problem.

4.2. Security integration

To return to our goal of securing our protocol, here we describe the steps of our secure geocast protocol. We combine symmetric cryptography with the techniques offered by elliptic curves to generate secret keys. Before the deployment, the BS calculates the initial parameters (ID, KDF_{initBS}, MAC_{initBS}, SYM_{initBS}, $E(F_p)$ and $K_{initBS} = k_{initBS}.G$ where $k_{initBS} \in [1, n-1]$) that will be used to execute the symmetrical cryptography algorithm, and charge them in each sensor. The BS also puts the cryptographic material so that once deployed, the nodes can communicate securely and build the network topology. For this purpose, the sink builds a secrete key *k* that will be the same for all the sensors, an elliptic curve *E* which will allow the sensors to know the point of the curve used to secure communications and an initial point *P* (belonging to the first cyclic group *Fp*).

Before the election of the CH2 in a zone (period of confidence), the messages exchanged are secured by the initial parameters loaded in each sensor before the deployment. When the CH2s are all elected, the sink can easily change the keys used by the CH2s (keys of zones) at any time to make it more difficult to an attacker to penetrate the network. Likewise, the technique can be at will repeated by each CH2 to refresh the keys of its CH1s, and by each CH1 to refresh those of its common sensors (keys of cliques). The keys refreshment mechanism is done here in a secure way and with little effort. For this purpose, after a time arbitrarily chosen by the sink, the latter randomly generates a number *x* belonging to *Fp*, it encrypts it with the current key and broadcasts it in the entire network to the attention of the CH2s. When the CH2s receive and decipher the message with their current key, they understand that they must change the current key by jumping *x* points from the current position *P* of the elliptic curve. With this new position, each CH2 easily calculates the new encryption parameters. We are sure that through this strategy, a malicious sensor that has also received information from the sink, according to which the sensors have to change their point on the curve, will not be able to interfere with the operation of the network because they do not have any elements allowing it to determine the points of the elliptic curve.

With this security mechanism, we note that the keys are refreshed without ever having to circulate in the network. The phase of formation of the cliques and zones, election of CH1 and CH2 is entirely secured by the key initially generated and loaded in each sensor. All further exchanges are secure thanks to the random and localized key refreshment performed by the sink, the CH2s and the CH1s. Each time a CH1 changes the key used in its clique (by generating a number *x*), it also notifies the cliques linked to it clique, so that they can recalculate the encryption and decryption key corresponding to each neighboring cluster. Note that two neighboring cliques or two neighboring areas do not necessarily use the same key which would help to circumscribe the damages in case of possible attacks.

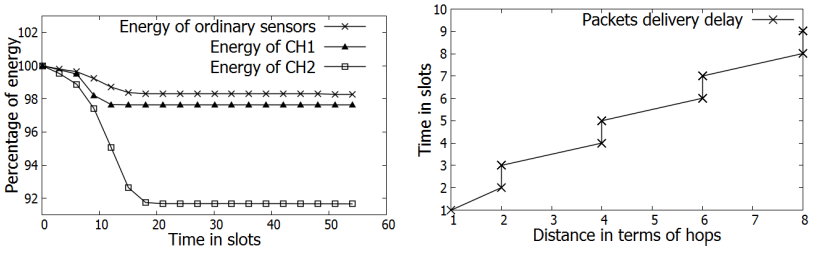
To pass information from a sensor to the sink, the sensor encrypts the message with the key of the clique and sends it to its CH1. The CH1 deciphers the message and encrypts it with the key of the relay clique. The message received by any sensor of the relay clique is then transmitted to the CH1 of the clique which can be at the same time the CH2 of the zone. This is how the collected data are securely routed during date collection and geocasting. When the sink wants to transmit a secure information in a zone, it encrypts it with the key currently used by the CH2 of the zone before transmitting it with its directional antenna. This allows to not divulge the carried messages. In the next section, we perform simulations to study the behaviors of this protocol.

5. Simulation and analysis of the results

Simulations were performed on a laptop (Intel Core i3 CPU M350 @ 2.27GHz $\times$ 4, 8GB RAM, Ubuntu 16.04) with the J-Sim simulator[6]; We deployed 500 sensors each equipped with a battery of 100 joules, a radius of transmissions of 20m. Each transmission requires 35.28×10^{-3} joule and each reception costs 31.32×10^{-3} joule [2]. A slot lasts $78\mu s$. The virtual architecture covers a radius of 400m and has 8×8 zones. The simulation includes the partitioning into cliques, creating the virtual architecture, routing the data to the sink and sending data to the geocast regions. Curves are made with gnuplot 5.0.

Energy consumption: The figure 4a shows the energy consumption for both ordinary sensors and clusterheads of level 1 and 2. This allows us to observe that the energy consumption is more accentuated respectively at the CH2 level, followed by the level of CH1 and finally at the level of ordinary sensors. This is explained by the fact that the CH2 are responsible for the management of the entire zones.

Messages delivery delay: The axis of distance of figure 4b denotes the distance separating the sender sensor from the geocast region. The figure shows that the delivery delay is quite proportional to this distance. Due to the lack of an aggregation technique, the geocast message could be duplicated in some cliques on the way to the geocast region. Nevertheless, these late messages reaches the destination not more than one slot after the expected time. That is why this curve has stairs' shape.



(a) Energy consumption. (b) Packets delivery delay.

Figure 4: Simulations results.

6. Conclusion

In this paper, we have set up a hierarchical clustering protocol to build a layered virtual architecture. We have defined mechanisms to achieve geocasting for one region and for several regions while remaining energy efficient. Subsequently we secured this protocol using cryptographic methods based on elliptic curves. Elliptic curves are used here to generate secret key. Keys are randomly refreshed after a random time without having to circulate in the network. The tests conducted gives acceptable results showing that the proposed solution works and is valid. In addition, the security integrated in our contribution avoids many types of attacks and is equipped with a control mechanism at different levels, which makes it a robust solution. In our next work, we intend to increase a lit-

tle more the challenges by addressing quality of service, fault tolerance mechanisms and including mobile sensors.

7. References

[1] A. Wadaa, S. Olariu, L. Wilson, M. Eltoweissy, K. Jones, "Training a wireless sensor network", *Mobile Networks and Applications*, Vol. 10, Num. 1-2, p. 151–168, 2005.

[2] 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.

[3] Das, Ashok Kumar, "A secure and effective biometric-based user authentication scheme for wireless sensor networks using smart card and fuzzy extractor", *International Journal of Communication Systems, Wiley Online Library*, Vol. 30, Num. 1, 2017.

[4] Dou, Yunqi, Weng, Jiang, Ma, Chuangui, Wei, Fushan, "Secure and efficient ECC speeding up algorithms for wireless sensor networks", *Soft Computing, Springer*, Vol. 21, Num. 19, p. 5665–5673, 2017.

[5] Khan Imran, Belqasmi Fatna, Glitho Roch, Crespi Noel, Morrow Monique, Polakos Paul, "Wireless sensor network virtualization: A survey", *IEEE Communications Surveys & Tutorials, IEEE*, Vol. 18, Num. 1, p. 553–576, 2016.

[6] A discrete event network simulator, J-Sim : *https://sites.google.com/site/jsimofficial/*, 2016.

[7] Panta, Rajesh Krishna, Auzins, Joshua Marc, Fernandez, Maria F, Hall, Robert J, "Geocast protocol for wireless sensor network", *Google Patents, US Patent 9,210,589* , dec 2015.

[8] Saqib, Najmus, Iqbal, Ummer, "Security in wireless sensor networks using ECC, *Advances in Computer Applications (ICACA), IEEE International Conference on, IEEE*, p. 270–274, 2016.

[9] S. Faye, C. Chaudet, I. Demeure, "A Distributed Algorithm for Adaptive Traffic Lights Control", *15th International IEEE Annual Conference on Intelligent Transportation Systems, Anchorage, USA*,September, 2012.

[10] 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, Num. 12, December, 2012.

[11] Sébastien Faye, Jean-Frédéric Myoupo, "An Ultra Hierarchical Clustering-Based Secure Aggregation Protocol for Wireless Sensor Networks", *AISS: Advances in Information Sciences and Service Sciences*, Vol. 3, Num. 9, p. 309 – 319, 2011.

[12] Sun Kun, Peng Pai, Ning Peng, Wang Cliff, "Secure distributed cluster formation in wireless sensor networks", *Computer Security Applications Conference. ACSAC'06. 22nd Annual, IEEE*, p. 131–140, 2006.

[13] Vianney Kengne Tchendji, Blaise Paho nana, "Management of Low-density Sensor-Actuator Network in a Virtual Architecture", *ARIMA Journal*, Vol. 27, p. 75–100, 2018.

[14] Wang, Neng-Chung, Wong, Shih-Hsun, "Agrid-Based Geocasting Protocol for wireless sensor networks", *Machine Learning and Cybernetics (ICMLC), 2016 International Conference on, IEEE*, Vol. 2, p. 530–534, 2016.

[15] Yassen Muneer Bani, Aljawaerneh Shadi, Abdulraziq Reema, "Secure low energy adaptive clustering hierarchal based on internet of things for wireless sensor network (WSN): Survey", *Engineering & MIS (ICEMIS), International Conference on, IEEE*, p. 1–9, 2016.

A Survey on e-voting protocols based on secret sharing techniques

Wafa Neji* — Kaouther Blibech** — Narjes Ben Rajeb ***

* Higher Institute of Technological Studies of Beja, The General Directorate of Technological Studies, Tunisia
wafa.neji@gmail.com

** Higher Institute of the Environment, Town planning and Building Technologies, University of Carthage, Tunisia
kaouther.blibech@gmail.com

*** National Institute of Applied Sciences and Technology, University of Carthage, Tunisia
narjes.benrajeb@gmail.com

ABSTRACT. Secret sharing techniques allow sharing a secret between a group of participants such that each of them holds one part of it. The secret can be reconstructed only when a subset of valid shares are combined together. These techniques are used in e-voting protocols since they allow to distribute the trust between several authorities and thus, achieve a greater degree of security. In this paper, we propose a classification of existing e-voting protocols based on secret sharing techniques and analyze their main advantages and drawbacks. We also identify security properties that could be ensured when using secret sharing techniques.

RÉSUMÉ. Le partage de secrets permet de partager un secret entre un ensemble de participants, chacun d'entre eux disposant d'une part du secret. Le secret ne peut être reconstruit que quand un sous-ensemble de parts valides sont réunies. Ces techniques sont utilisées dans les protocoles de vote électronique afin de distribuer la confiance entre plusieurs autorités électorales et atteindre ainsi un degré de sécurité plus important. Dans ce papier, nous proposons une classification des protocoles de vote électronique préexistants basés sur les techniques de partage de secret et nous analysons leurs avantages et inconvénients. Nous déterminons aussi les exigences de sécurité qui pourraient être satisfaites grâce à l'utilisation des différentes techniques de partage de secret.

KEYWORDS : Electronic voting, security, secret sharing, distributed key generation protocol

MOTS-CLÉS : Vote électronique, sécurité, partage de secret, protocole de génération de clé distribuée

1. Introduction

Electronic voting protocols are based on different approaches and cryptographic mechanisms that allow them to guarantee the validity of the voting process. Secret sharing techniques are one of the commonly used approaches because they avoid that a single electoral authority has the power to decrypt individual ballots, to access partial result, or to compute exclusively the final result. These behaviors compromise the security requirements of the voting process [1, 2]. Indeed, during an election, e-voting protocols must satisfy security properties of voting process. Foremost, all votes must be kept secret (privacy), and no traceability between the voter and his vote can be established (anonymity). Moreover, anyone can check the validity of the final voting result and voters must be able to ensure that their votes have been taken into account (verifiability) while preventing them from proving for any party how they voted (receipt-freeness). In addition, voters must also be able to vote correctly even if they are under a threat of an adversary (Incoercibility). Furthermore, the protocol must be robust against a coalition of a partial number of dishonest authorities (robustness).The complexity of the protocol is an important element which must be taken into consideration. Indeed, an efficient e-voting protocol has to be scalable according to time, communication and computation costs needed to include a larger number of voters (scalability).

Note that the use of cryptographic mechanisms in e-voting protocol could contribute to ensure these security requirements. Several comparative studies of these mechanisms have been proposed in the literature [1, 2, 3, 4]. Most of these works only provide an overview of these approaches and a basic understanding of e-voting protocols. Security requirements are not studied according to the used cryptographic techniques.

In this paper, we propose a particular analysis of e-voting protocols based on the used Secret Sharing Techniques (SST). In addition, we provide the security requirements that could be ensured thanks to the use of SST techniques. First, we introduce the notion of SST. Second, we present a classification of e-voting protocols according to the used SST. After that, based on this classification, we analyze the main advantages and drawbacks of e-voting protocols based on SST and we identify the most important security requirements that could be ensured through that.

2. Secret sharing techniques

The secret sharing is a cryptographic mechanism that allows to divide a secret data s, chosen initially by a trusted party named the dealer, in several parts $s_1, ..., s_n$. These shares are distributed among n participants such that only the coalition of a subset t of them allows the reconstruction of the original secret s. This mechanism is called the (t, n)

threshold secret sharing scheme. The first secret sharing schemes appeared in 1970 and were proposed simultaneously by Shamir [5] and Blakley [6]. The main problem of these schemes is that there is no guarantee of an adequate reconstruction of the secret if the dealer cheats by generating and distributing invalid shares or if one of the participants cheats by restoring invalid shares. Verifiable Secret Sharing (VSS) schemes [7] partially solve this problem because they allow participants to verify the validity of the shares received from the dealer. Unfortunately, dishonest participant can still restore invalid shares and skew the reconstruction of the secret. The solution of this problem is provided by Publicly Verifiable Secret Sharing (PVSS) schemes that allow not only the participants but also any external party to verify the validity of the distributed and/or the restituted shares. In the literature, several PVSS schemes has been proposed [8, 9, 10, 11, 12, 13]. These schemes can be used as a building block to design secure e-voting protocols.

3. Classification of e-voting protocols based on SST

In e-voting protocols, when a single authority supports the execution of the whole voting process, it is not possible to guarantee security requirements of the protocol. For example, if a private secret key used to decrypt ballots is owned by a single authority, this latter can decrypt voters' ballots and know the vote of each voter. To avoid this kind of situation, the secret key should not be held by a single authority and must be shared among several authorities. This process distributes the trust to achieve a greater degree of security, and reduces the risk of the presence of any dishonest authorities. During the voting process, secret sharing can be used in three different ways, as follows:

1) Class 1 : The secret is a private key shared between authorities. This private key is used to decrypt all ballots.

2) Class 2 :The secret is a ballot. Each voter uses a secret sharing scheme to share its ballot between authorities.

3) Class 3 :The secret is a decryption key of a single ballot. Each voter uses a secret sharing scheme to share the decryption key of its ballot between authorities.

Based on these three approaches, we propose in what follows a classification of e-voting protocols using SST.

3.1. Class 1: Authorities' shared key

For this first class, SST are used before the beginning of the voting process. During the initialization phase of the election, a trusted party generates and shares a private key between authorities. The public key associated with this private key is used by voters to encrypt their ballots. The secret key is used by authorities during the tallying phase to compute the final voting result. The use of this technique appeared in [14] and has been

improved in [15, 16, 17, 18, 19, 20, 21, 22, 23, 24]. In general case, the generation of the secret key is carried out using secret sharing schemes [17, 18, 19, 20]. In the literature, a multitude of e-voting protocols of Class 1 use Shamir's secret sharing scheme as a building block. The major disadvantage of this is to involve a single trusted dealer who initially holds the secret key. In fact, this latter can decrypt individual ballots and compute partial result of final tallying. This compromises the security of the voting process. Another disadvantage is the lack of verification protocols for distributed and/or restituted shares. Thus, it is more appropriate to use VSS or PVSS schemes that include verification protocols. This ensures the validity of the shares distributed by the dealer, and restituted by the authorities. To avoid the need for a single trusted distributor, several e-voting protocols use distributed key generation protocols (DKG) which involve multiple parties to jointly generate and share the secret key. A multitude of protocols defined in the literature [14, 15, 16, 21] are based on the DKG protocol of Pedersen [25]. However, this DKG protocol has some drawbacks. It has been proven in [26] that this protocol cannot ensure a uniform distribution of the generated keys. Thus, the use of a secure DKG protocol to define e-voting protocols belonging to Class 1 is essential. Several authors proposes protocols of the Class 1 [14, 15, 18, 16, 19, 20, 21] which are based on secure DKG protocols. These protocols propose new versions of threshold cryptosystems used to encrypt ballots. The use of threshold cryptosystems is useful: the authorities cooperate to perform a multiple decryption of the final result without decrypting the ballots one by one. In addition, the secret key of the authorities is never reconstructed and is used implicitly in the tallying phase.

3.2. Class 2: Shared ballot

For this class, the SST are used during the voting process. Each voter acts as a dealer and shares its ballot between authorities using a secret sharing scheme. Each authority receives a different part of each ballot. To compute the final voting result, authorities use the homomorphic property [28] of the secret sharing scheme and multiply all the shares received from voters. In the literature, a multitude of e-voting protocols use this technique. E-voting protocols of Class 2 appeared first in [27] and were later improved in [28, 30, 31, 32, 33]. However, these protocols have some drawbacks. On the one hand, the majority of them, like those proposed in [30, 31, 33], struggle to prove the validity of the voting value contained in the shared ballot. On the other hand, some of these protocols [30, 31, 33] use non-verifiable secret sharing schemes as a building block. Thus, these schemes don't provide means to check the validity of ballots' shares distributed by voters and restituted by authorities. Voters can send invalid shares of their ballots and authorities can falsify the voting result by giving shares that do not actually come from the voters. For this purpose, protocols of Class 2 based on non-verifiable secret sharing schemes cannot ensure the verifiability property. The solution to this problem can be provided using VSS or PVSS schemes, which add verification protocols to check the validity of

ballots' shares. This is the case, for example, of the voting protocol proposed by Cramer et al. in [28] which is based on the VSS scheme of Pedersen [29]. Unfortunately, the use of VSS or PVSS schemes is not possible in all cases. This is due to the use of verification protocols that compromise the confidentiality of the vote. This is the case, for example, of Feldman's VSS scheme, in which the content of the vote can be revealed from the public commitments of the Shamir polynomial coefficients using a simple exhaustive search (since each vote belongs to a predefined set of values). The use of Pedersen's VSS scheme [29] is more appropriate in this case since the public commitments used to verify the shares do not provide any information on the value of the vote. For this purpose, the most appropriate secret-sharing schemes for protocols of Class 2 are VSS or PVSS schemes which preserve the confidentiality of the vote and which have an homomorphic property facilitating the computing of the final voting result.

3.3. Class 3: Ballot's shared key

Electronic voting protocols in this class use a similar technique to the one used in the protocols of the second class. However, instead of sharing his/her ballot, each voter will share a secret key between authorities. This secret key is used by the voter to encrypt the ballot. Then, the coalition of authorities is needed to reconstruct the secret key of each voter and to decrypt ballots. The first e-voting protocol that uses this technique was proposed by Schoenmakers in 1999. It's also is the first e-voting protocol based on a PVSS scheme. In 2002, Kiayias and Yung [34] took inspiration from Schoenmakers' protocol to propose a protocol allowing voters to participate in the tallying phase to compute the final voting result. However, this protocol is based on (m, m) secret sharing scheme (where m is the number of voters) and requires the presence of all voters to compute the final result. In 2014, Zou et al. also propose in [35] an e-voting protocol based on (m, m) secret sharing scheme. The major disadvantage of this approach is that if only one of the shares is lost, the voting result will be permanently inaccessible. Moreover, the time and complexity of communication defined in [34, 35] depend on the number of voters. These protocols can be applied only to elections with a small number of voters.

4. Analysis of classes' security requirements

In this section, we focus our analysis on the security requirement that could be ensured by voting protocols thanks to the use of SST.

4.1. Privacy

The use of SST helps to satisfy this security requirement. For protocols belonging in each class, violating the privacy of a ballot implies that an adversary can compute the

secret key shared between authorities (Class 1), can reconstruct the ballot from the shares sent to authorities (Class 2), or can compute the secret key related to the ballot from shares sent to authorities (Class 3). In all these cases, compromising the vote's privacy implies getting a secret data shared between authorities with a secret sharing scheme. Therefore, privacy of ballots depends on the security of the secret sharing scheme which must satisfy the secret property. Note that a secret sharing scheme satisfies the secret property for a secret data s if a dishonest party cannot get s, or any information related to s. Thus, if the secret property is satisfied, a dishonest party cannot decrypt ballots and cannot get any information related to the votes' values. This helps to ensure the privacy of the vote.

Remark 1. *The use of a secret sharing scheme that verifies the secret property contributes to ensure the privacy of the vote.*

For this purpose, e-voting protocols based on SST satisfy the privacy if the used secret sharing scheme verifies the secret property. Note that secret sharing schemes use several cryptographic primitives to ensure secrecy. Most of these primitives are based on NP-hard problems.

4.2. Anonymity

E-voting protocols belonging to Classes 1, 2 and 3 satisfy voter's anonymity by assuming the honesty of a subset of authorities who will not cooperate to decrypt individual ballots. For this purpose, this assumes that authorities will not cooperate to explicitly reconstruct the authorities'secret key (Class 1), or will not cooperate to reconstitute individual ballots from received shares (Class 2), or will not explicitly reconstitute the secret key related to each ballot (Class 3).

Remark 2. *Voter's anonymity could be satisfied only assuming the honesty of at least t of the authorities who will not cooperate to decrypt ballots one by one.*

4.3. Receipt-Freeness

E-voting protocols based on STT in Class 2 and Class 3 fail to satisfy receipt-Freeness. This is due to the random values chosen by the voter to share his ballot (Class 2), or to share the secret key related to his ballot (C lass3). Thus, a voter can construct a receipt which can prove the content of his vote by revealing the random value that he used during the dealing phase. In general, in a voting process, when the voter chooses a random value to encrypt his vote, the voter can easily use it to construct a receipt of his vote [15, 16].

Remark 3. *Receipt-Freeness is not satisfied by e-voting protocols based on SST used in Class 2 and Class 3.*

In e-voting protocol of Class 1, the voter does not execute the secret sharing process to encrypt his ballot. He only uses the authorities'secret key. If he has to choose in

addition a random value it is possible to re-encrypt the value of encrypted ballot. The re-encryption can be performed using re-encryption mix-net [36], permutations carried out by authorities or by using a secure hardware device [16]. This process prevents the voter from keeping chosen random values. The voter is then unable to prove to an adversary the content of his vote

Remark 4. *Receipt-Freeness can be satisfied by e-voting protocols based on SST used in Class1 and combined with mix-net or re-encryption technique.*

Note that the combination of these techniques has some disadvantages. The use of re-encryption technique implies adding new proofs of verification to prove the validity of the re-encryption. In the case of using of re-encryption mix-nets, this leads to an important communication and computational complexity. New validity proofs must be added in each stage.

4.4. Incoercibility

In e-voting protocols based only on SST, it is not possible to satisfy the incoercibility. To avoid this, it is possible to resort to the use of anonymous credentials in combination with SST. Indeed, for the protocols of Class 1 and Class 3, to vote, each voter must submit a credential with his encrypted ballot to validate it. For protocols of Class 2, during the dealing phase, each voter must submit several credentials with the ballot shares to validate them. In any case, if an adversary forces the voter to vote in a certain way, the voter may submit an invalid credential. Recall that neither the voter nor the attacker can prove or verify the validity nor the invalidity of the submitted credential.

Remark 5. *The combination of SST with anonymous credentials contribute to ensure incoercibility.*

4.5. Robustness

The robustness implies that the voting protocol can tolerate the presence of a number of dishonest authorities. E-voting protocols based on SST can ensure this requirement. These protocols assume the presence of a minimal number t of honest authorities to share the authorities' secret key (Class 1), the ballots (Class 2) or the secret keys related to the ballots (Class 3). Inadequate behavior of $t-1$ coalition of authorities can be tolerated. No coalition of dishonest voters can disrupt the election.

Remark 6. *The use of (t, n) threshold secret sharing schemes in e-voting protocol contributes to satisfy the robustness property.*

4.6. Verifiability

In PVSS schemes, a public validity proof is added to allow any party to verify the validity of the distributed and restored shares. The application of a PVSS scheme to an e-voting protocol ensures verifiability. In fact, any participant can ensure that the ballots have been counted correctly by verifying the validity of ballots' shares (Class 2) or decryption keys' shares (Class 3) distributed by voters and given back by authorities. The decryption that leads to the final result from valid ballots is also verifiable. Note also that the use of a PVSS scheme or a DKG protocol based on PVSS scheme for protocols of Class 1 also allows any participant to verify the validity of the distributed decryption performed by authorities.

Remark 7. *The use of SST based on PVSS schema could contribute to ensure verifiability.*

4.7. Scalability

When examining the performance of the voting process, we notice that the work done by the voter in protocols of Class 1 seems requiring less computation operations than Class 2 and Class 3 [14]. An interesting property of protocols of Class 1 is to see whether that complexity and communication time are independent of the number of voters and authorities. Indeed, in protocols of Class 1, a voter will simply send a single encrypted ballot accompanied by a single proof that proves the validity of his vote. Nevertheless, in protocols of Classes 2 and 3, the voter must send several encrypted shares according to the number of authorities and must prove the validity of each share.

Remark 8. *The property of scalability could be provided by e-voting protocols of Class 1.*

For this purpose, protocols belonging to Classes 2 and 3 seem to be more appropriate for small elections, on account of the complexity of computational operations made during the voting and tallying phases and the manifold proofs generated by voters. The protocols belonging to Class 1 can be applied for large-scale elections.

4.8. Summary of the analysis

Table 1 provides a summary of the security requirements that could be satisfied by each class. From the Table 1, we deduce that the use of the secret sharing technique related to Class 1 is the most appropriate for the design and the definition of e-voting protocols. Combined with other cryptographic approaches (anonymous credentials, re-encryption, mix-nets, etc.), this technique helps to satisfy the security requirements of the voting process. Note that the most appropriate way to exploit this technique is using DKG protocols based on PVSS schemes. Indeed, on the one hand, this avoid having recourse

to a trusted party who initially holds the secret key, and on the other hand contributes to ensure the property of verifiability.

Table 1 – Classification: Analysis of security requirements

	Privacy	Anonymity	Robust-ness	Receipt-freeness	Incoercibil-ity	Verifiability	Scalability
Class 1	Com	C	✓	CCA	CCA	PVSS / DKG based on PVSS	✓
Class 2	Com	C	✓	X	CCA	PVSS	X
Class 3	Com	C	✓	X	CCA	PVSS	X

Com : computational privacy, **C** : conditionally satisfied, **CCA** combination with other cryptographic approaches, ✓ : satisfied, **X** : not satisfied

5. Conclusion

In this paper, we have studied the use of SST in e-voting protocols and analyze their main advantages and drawbacks. We have also proposed a classification of e-voting protocols based on the used SST. This classification led us to identify security properties that could be satisfied for each class. Depending on the targeted security requirements, our analysis may help in the selection of building blocks and cryptographic mechanisms that could be used in order to define secure electronic voting protocols.

It should be noted that the use of specific cryptographic approaches does not necessarily imply that e-voting protocols satisfy the required properties of e-voting process. For each protocol, it should be necessary to verify that the combination of all cryptographic building blocks contributes to ensure security requirements. Indeed, it would be interesting to formally prove the security of e-voting protocols. Thus, as future research, we intend to construct formal proofs in order to prove the satisfaction of security requirements related to e-voting protocols.

6. References

[1] MURSI, M. F. , ASSASSA, G. M. , ABDELHAFEZ, A. , SAMRA, K. M. A., " On the development of electronic voting: a survey ", *International Journal of Computer Applications*, Vol. 61, no 16, 2013.

[2] QADAH, G. Z. , TAHA, R., " Electronic voting systems: Requirements, design, and implementation ", *Computer Standards and Interfaces*, Vol. 29, no 3, 376-386, 2007.

[3] LAMBRINOUDAKIS, C. , GRITZALIS, D. , TSOUMAS, V. , KARYDA, M. , IKONOMOPOULOS, S., " Secure electronic voting: The current landscape", *In Secure electronic voting, Springer, Boston*, 101-122, 2003.

[4] SMITH, W. D., "Cryptography meets voting", *Technical report*, Vol. 10, 80, 2005.

[5] SHAMIR, A, "How to share a secret", *Communications of the ACM*, Vol. 22, no 11, 612-613, 1979.

[6] BLAKLEY, G. R., "Safeguarding cryptographic keys", *In Proceedings of the national computer conference*, Vol. 48, 313-317, 1979.

[7] FELDMAN, P., " A practical scheme for non-interactive verifiable secret sharing", *28th Annual Symposium on. IEEE*, 427-438, 1987.

[8] BEHNAD, A. , EGHLIDOS, T., "A new, publicly verifiable, secret sharing scheme", *Sci. Iran.* , 246-251, 2008.

[9] SCHOENMAKERS, B., "A simple publicly verifiable secret sharing scheme and its application to electronic voting", *In Annual International Cryptology Conference, Springer, Berlin, Heidelberg*, 148-164, 1999.

[10] FUJISAKI, F. , OKAMOTO, T., "A practical and provably secure scheme for publicly verifiable secret sharing and its applications", *In: Proceedings of the annual international conference on Theory and application of cryptographic techniques, EUROCRYPT'98, Springer-Verlag, Berlin, Heidelberg*, 32-46, 1998.

[11] Behnad08,Schoenmakers99,Fujisaki,Heidarvand09 HEIDARVAND, S. , VILLAR, J. L., "Selected Areas in Cryptography", *chapter Public Verifiability from Pairings in Secret Sharing Schemes, Springer-Verlag, Berlin, Heidelberg*, 294-308, 2009.

[12] JHANWAR, M. P., "A Practical (Non-interactive) Publicly Verifiable Secret Sharing Scheme", *In: ISPEC'11*, 273-287, 2011.

[13] SHIL, A. , BLIBECH, K. , ROBBANA, R. , NEJI, W., " A New PVSS Scheme with a Simple Encryption Function", *arXiv preprint arXiv:1307.8209*, 2013.

[14] CRAMER, R. , GENNARO, R. , SCHOENMAKERS, B. , "A secure and optimally efficient multi-authority election scheme", *Transactions on Emerging Telecommunications Technologies*, 8(5), 481-490, 1997.

[15] HIRT, M. , SAKO, K. , "Efficient receipt-free voting based on homomorphic encryption", *In International Conference on the Theory and Applications of Cryptographic Techniques* , 539-556, 2000.

[16] LEE, B. , KIM, K., "Receipt-free electronic voting scheme with a tamper-resistant randomizer", *In International Conference on Information Security and Cryptology*, 389-406, 2002.

[17] DAMGARD, I., "A generalisation, a simplification and some applications of Paillier's probabilistic public-key system", *Proceedings of the 4th International Workshop on Practice and Theory in Public Key Cryptosystems*, 119-136, 2001.

[18] FOUQUE, P. A. , STERN, J., "One round threshold discrete-log key generation without private channels", *Public Key Cryptography*, 300-316, 2001.

[19] ACQUISTI, A. , "Receipt-Free Homomorphic Elections and Write-in Ballots", *IACR Cryptology ePrint Archive*, p 105, 2004.

[20] PORKODI, C. , ARUMUGANATHAN, R. , VIDYA, K., "Multi-authority Electronic Voting Scheme Based on Elliptic Curves", *IJ Network Security*, Vol 12(2), 2011.

[21] PHILIP, A. A. , SIMON, S. A. , OLUREMI, A., " A receipt-free multi-authority e-voting system", *International Journal of Computer Applications*, Vol 30, no 6, 15-23, 2011.

[22] CHONDROS, N. , ZHANG, B. , ZACHARIAS, T. , DIAMANTOPOULOS, P. , MANEAS, S. , PATSONAKIS, C. , ROUSSOPOULOS, M., "D-DEMOS: A distributed, end-to-end verifiable, internet voting system", *In Distributed Computing Systems (ICDCS), 2016 IEEE 36th International Conference*, 711-720, 2016

[23] CULNANE, C. , RYAN, P. Y. , SCHNEIDER, S. , TEAGUE, V. , "vVote: a verifiable voting system", *ACM Transactions on Information and System Security (TISSEC)*, Vol 18, no 1, 2015.

[24] CHAIDOS, P. , CORTIER, V. , FUCHSBAUER, G. , GALINDO, D. , "Beleniosrf: A non-interactive receipt-free electronic voting scheme", *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 1614–1625, 2016.

[25] PEDERSEN, T. P., "A threshold cryptosystem without a trusted party", *In Workshop on the Theory and Application of Cryptographic Techniques*, 522-526, 1991.

[26] GENNARO, R. , JARECKI, S. , KRAWCZYK, H. , "Secure distributed key generation for discrete-log based cryptosystems", *Journal of Cryptology*, 20(1), p. 51-83, 2007.

[27] BENALOH, J. D. C. , " Verifiable secret-ballot elections", 1987.

[28] CRAMER, R. , FRANKLIN, M. , SCHOENMAKERS, B. , YUNG, M. , "Multi-authority secret-ballot elections with linear work", *In International Conference on the Theory and Applications of Cryptographic Techniques*, 72-83, 1996.

[29] PEDERSEN, T. P., "Non-interactive and information-theoretic secure verifiable secret sharing", *In Annual International Cryptology Conference*, 129-140, 1991.

[30] IFTENE, S., "General secret sharing based on the chinese remainder theorem with applications in e-voting", *Electronic Notes in Theoretical Computer Science*, 186, 67-84, 2007.

[31] SPIRIDONICĂ, A. M. , PISLARU, M., "THE ASSURANCE OF SECURITY OF ELECTRONIC VOTING THROUGH THE USE OF SECRETS SHARING SCHEMES AND BENALOH ELECTRONIC VOTING SCHEME", 2010.

[32] OTSUKA, A. , IMAI, H. , "Unconditionally secure electronic voting", *In Towards Trustworthy Elections*, 107-123, 2010.

[33] NAIR, D. G. , BINU, V. P. , KUMAR, G. S. , "An improved e-voting scheme using secret sharing based secure multi-party computation", *arXiv preprint arXiv:1502.07469*, 2015.

[34] KIAYIAS, A. , YUNG, M., "Self-tallying elections and perfect ballot secrecy", *International Workshop on Public Key Cryptography*, 141-18, 2002.

[35] ZOU, X. , LI, H. , SUI, Y. , PENG, W. , LI, F., "Assurable, transparent, and mutual restraining e-voting involving multiple conflicting parties", *INFOCOM, 2014 Proceedings IEEE*, 136-144, 2014.

[36] NEFF, C. A., "A verifiable secret shuffle and its application to e-voting", *Proceedings of the 8th ACM conference on Computer and Communications Security*, 116-125, 2001.

Appendix - Some existing e-voting proyocols : Analysis of security requirements

The Table 2, Table 3 and Table 4, give respectively a summary of the security requirements satisfied by some e-voting protocols belonging to Class 1, Class2 and Class3.

Table 2 – Class 1: Analysis of security requirements

Protocol	Privacy	Anonymity	Robustness	Receipt-freeness	Incoercibility	Verifiability	Scalability
Cramer et al. (1997)	Com	C	✓	X	X	✓	✓
Damgard et Jurik (2000)	Com	C	✓	X	X	✓	✓
Hirt et Sako (2000)	Com	C	✓	✓	X	✓	X
Fouque et al. (2001)	Com	C	✓	✓	X	✓	✓
Lee et Kim (2002)	Com	C	✓	C	X	✓	✓
Acquisti (2004)	Com	C	✓	AP	AP	X	X
Civitas/JCJ (2008)	Com	C	✓	✓	✓	✓	X
Porkodi et al. (2011)	Com	C	✓	X	X	✓	✓
Philip et al. (2011)	Com	C	✓	✓	X	✓	X
Chondros et al. (2015)	Com	C	✓	C	X	✓	✓
BeleniosRF (2016)	Com	C	✓	✓	✓	✓	✓

Com : computational privacy, **C** : conditionally satisfied, **AP** : attack proved, ✓ : satisfied, **X** : not satisfied

Table 3 – Class 2: Analysis of security requirements

Protocol	Privacy	Anonymity	Robustness	Receipt-freeness	Incoercibility	Verifiability	Scalability
Cramer et al. (1996)	Com	C	✓	X	X	✓	X
Iftene (2007)	Com	C	✓	X	X	X	X
Spiridoncia et al. (2010)	Com	C	✓	X	X	X	X
Otsku et Imai (2010)	Com	C	✓	X	X	✓	X
Mukhopadhyay (2014)	Com	C	X	X	X	X	X
Nair et al. (2015)	Com	C	✓	X	X	X	X

Com : computational privacy, **C** : conditionally satisfied ,✓ : satisfied, **X** : not satisfied

Table 4 – Class 3: Analysis of security requirements

Protocol	Privacy	Anonymity	Robustness	Receipt-freeness	Incoercibility	Verifiability	Scalability
Schoenmakers (1999)	Com	C	✓	X	X	✓	X
Kiayias et Yung (2002)	Com	C	X	X	X	✓	X
Zou et al. (2014)	Com	C	X	X	X	✓	X

Com : computational privacy, **C** : conditionally satisfied, ✓ : satisfied, **X** : not satisfied

Survey of Internet of Things Applications in Smart Agriculture: A typical architecture

Salim Chikhi* — Badreddine Miles**

* MISC Laboratory
University Abdelhamid Mehri Constantine 2
salim.chikhi@univ-constantine2.dz

** MISC Laboratory
University Brother Mentouri Constantine 1
badr-eddine.miles@umc.edu.dz

ABSTRACT. The world will need to produce 70% more food in 2050 than it did in 2006 in order to feed the growing population of the Earth, according to the UN Food and Agriculture Organization. To meet this demand, farmers and agricultural companies are turning to the Internet of Things for analytics and greater production capabilities. The IoT is set to push the future of farming to the next level thanks to agricultural drones and sensors. In this article, we present a survey of the various technological advances used in the Internet of Things (IOT) for smart agriculture, and we end by raising the positive contribution of the use of IoT in the development of agriculture by projects and applications that have already been realized. At the same time, we are attracting the attention of African researchers to look into the possibilities of developing IoT-based systems to meet the challenges of agriculture on our continent by the proposition of a typical architecture.

RÉSUMÉ. Selon l'Organisation des Nations Unies pour l'alimentation et l'agriculture, le monde devra produire 70% de plus de nourriture en 2050 qu'en 2006 pour nourrir la population croissante de la planète. Pour répondre à cette demande, les agriculteurs et les entreprises agricoles se tournent vers l'Internet des objets pour des analyses et des capacités de production plus importantes. L'IoT devrait faire passer l'avenir de l'agriculture à un niveau supérieur grâce aux drones et aux capteurs agricoles. Dans cet article, nous présentons une étude des différentes avancées technologiques utilisées dans l'Internet des objets (IOT) pour l'agriculture intelligente, et nous concluons en soulignant la contribution positive de l'utilisation de l'IoT dans le développement de l'agriculture par des projets et des applications déjà réalisés. Dans le même temps, nous attirons l'attention des chercheurs africains sur les possibilités de développer des systèmes basés sur l'IoT pour relever les défis de l'agriculture sur notre continent par la proposition d'une architecture typique.

KEYWORDS : Internet of things, Applications of IoT, Smart agriculture, Sensors, LPWAN.

MOTS-CLÉS : Internet des objets, Applications de l'IoT, Agriculture intelligente, Capteurs, LPWAN.

1. Introduction

The applications of IoT are numerous and can improve the daily life of citizens. IoT has a good chance of succeeding in developing countries because of existence of means to overcome challenges for effective deployment of IoT solutions. The issues across Africa can be very different from other continents [1, 2].

IoT can be used, among others, to:

– help deliver clean water to thousands of people,

– better protect endangered species,

– diagnose and follow patients remotely ,

– make roads and streets safer for citizens,

– better inform farmers and increase crop production.

The application of IoT in Africa faces the barriers of the required investment and the weak existing infrastructure. However, Literature shows that many countries have started experimenting Iot based applications, such as:

– intelligent traffic lights in Nairobi are helping to ease traffic congestion,

– new smart city in the suburb of Algiers,

– smart meters to reduce the load and avoid power outages in south africa,

– unmanned aerial vehicles (UAVs) are used to protect some african national parks.

In the agricultural domain, Iot can play a crucial role in Africa because there are one billion farmers across the continent who contribute significantly to the economy of their countries.

Through this paper, we try to review some important works done in developed countries in the agricultural field to draw some lessons that encourage African researchers to invest in this area for the good of our continent. We give some recommendations for the application of IoT in the different agricultural applications.

The rest of the paper is organized as follows: Section 2 summarizes some well known IoT definitions and the nature of data involved in most corresponding systems. Section 3 is on the communication aspects which are the heart of any IoT system. In section 4 we present a brief overview of some interesting IoT solutions for intelligent agriculture. Section 5 is reserved for the proposition of a basic architecture to be used in any system dedicated to making farms intelligent. In section 6, we conclude the present study.

2. IoT Definition

According to the International Telecommunication Union, the Internet of Things (IoT) is a " global infrastructure for the information society, which provides advanced services by interconnecting objects (physical or virtual) with the technologies of the Internet. Ex-

isting and evolving interoperable information and communication.
According to [3] defines "the internet of things as a network of networks that allows, via standardized and unified electronic identification systems, and mobile wireless devices, to identify directly and unambiguously digital entities and objects physical and thus to be able to recover, store, transfer and process, without discontinuity between the physical and virtual worlds, the data related thereto".
According to [4] defines "IoT is a dynamic global network infrastructure with self- configuring capabilities based on standard and interoperable communication protocols where physical and virtual Things have identities, physical attributes, and virtual personalities and use intelligent interfaces, and are seamlessly integrated into the information network".
The semantic meaning of the "Internet of Things" is presented as "a globally network of uniquely addressable interconnected objects based on standard communication protocols"
According to Gartner, 25 billion devices will be connected to the Internet by 2020 and these connections will make it easier to use data to analyze, plan, manage and make intelligent decisions independently. In this context, IoT can be used in several sectors, such as transportation, smart city, intelligent home automation, intelligent health, life support, logistics, automation, industry, and agriculture.
Although the data generated daily in agriculture continue to rise. BI Intelligence, Busi-

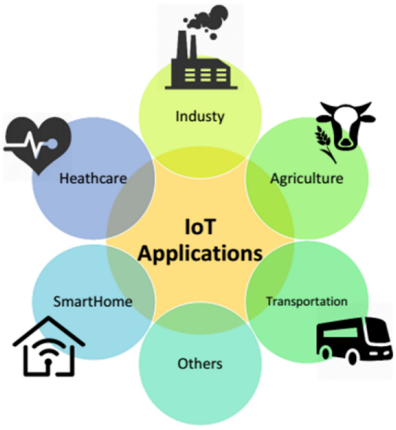


Figure 1. *different areas of IoT applications*

ness Insider's premium search service [5], predicts that IoT device installations in the agricultural world will grow from 30 million in 2015 to 75 million in 2020, for a compound annual growth rate of 20

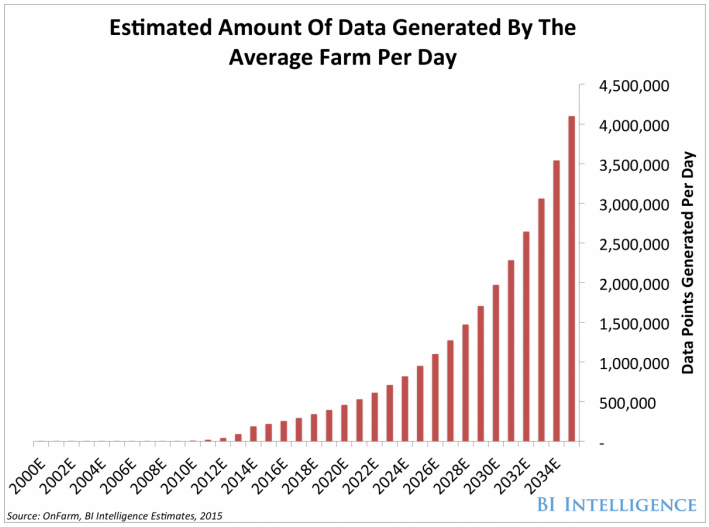


Figure 2. *Estimated Amount of data generated by the average farm per day [5]*

Precision agriculture is sometimes referred to as "smart farming", involving different types of technologies such as data analysis, sensing technologies, communication systems, and hardware and software systems. These technologies are needed to implement the computing system that gathers, analyzes, and then presents the data to initiate an appropriate response to the information received.

For farmers and producers, a wide variety of information regarding soil and crop behavior, animal behavior, condition of machinery, storage tank status from isolated sites is presented for farmers can make decisions and improve production.

It is true that for smart agriculture we need sensors that measure the temperature, the humidity, the climate, and the acidity of the soil. and also the sensors placed in the fields to allow farmers to obtain detailed maps of the topography and resources of the area. This data must be sent to the server for storage and analysis. Therefore, communication is very important in precision farming. There is a lot of recent technology for communication in IoT, for example Zigbee, z-wave, Bluetooth Low Power, Wifi, and LPWAN communication networks (SigFox, LoRa, NB-IoT). These proves very useful and practical in intelligent agriculture, because communication is used for a great distance up to 40 Km and inexpensive in energy, a battery can emit some messages a day for 10 years.

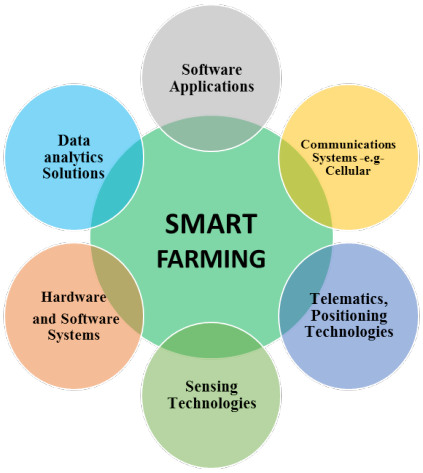


Figure 3. *The different types of technologies involved in smart farming*

3. LPWAN

Recently, a new generation of low power wide area network (LPWAN) has emerged to bridge the gap between wireless and mobile network technologies. This LPWAN network is still little known, but behind it hides more high-profile technologies such as LoRaWan, SigFox, and NB-IoT. These LPWAN networks have several features that make them particularly attractive for devices and applications that require low mobility and low levels of data transfer. This makes it possible to send and receive messages of very small size over very long periods of time. range (from 5km to 40km), with a major advantage to issue messages very inexpensive and very low energy (it is possible with a single battery, to issue a few messages per day for 10 years). To see the benefits of these LPWAN networks, Figure 4 shows a comparison of this network with two other commonly used communicating technologies, GSM (3G, 4G, and 5G) and ZigBee.

3.1. SigFox

The Sigfox technology was developed in 2010 by the start-up Sigfox (in Toulouse, France), both company and network operator LPWAN. Sigfox operates and markets its own IoT solution in 31 countries around the world, including two in Africa (South Africa and Tunisia), and is still being rolled out worldwide through partnership with various network operators [6].

Sigfox uses unlicensed ISM bands, for example, 868 MHz in Europe, 915 MHz in North America and 433 MHz in Asia. Using the ultra-narrow band, Sigfox uses bandwidth effectively and experiences very low noise levels, resulting in very low power consumption, high receiver sensitivity and economical antenna design at the cost of a maximum speed

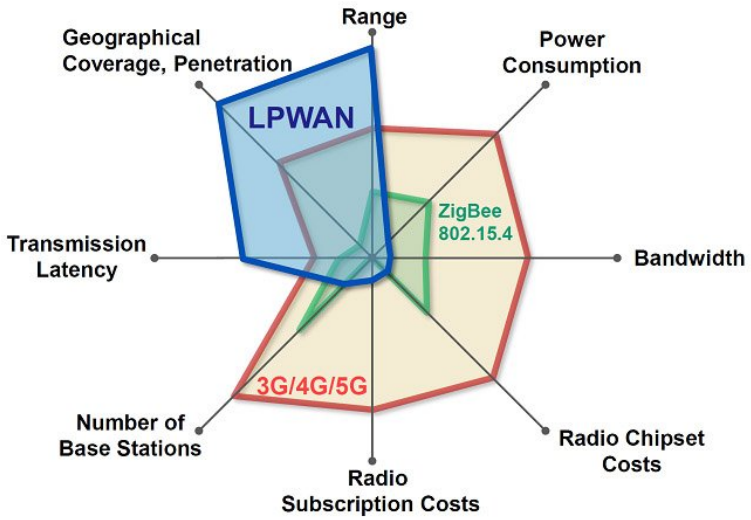


Figure 4. *comparative communicating technology*

of 100 bps.

The number of messages on the uplink is limited to 140 messages per day. The maximum payload length for each uplink message is 12 bytes.

3.2. LoRa

LoRa technology, developed by Semtech, is a physical layer technology that modulates signals in the sub-GHZ ISM band. It is the most widely used technology for LPWAN in the sub-GHz unlicensed band [7].Due to the use of unauthorized tapes; the LoRa network is open to customers who do not have the authorization of radio frequency regulators. As a result, the LoRa network is easy to deploy for a distance of more than several kilometers and serves customers with minimal investment and maintenance costs.

LoRa technology has been tested in 56 countries and in different areas, for example, traffic tracking, intelligent health car [8]. LoRa uses unlicensed ISM bands, namely 868 MHz in Europe, 433 MHz in Asia and 915 MHz in North America.

LoRa technology provides two-way communication through spectrum modulation. The maximum payload length for each message is 243 bytes. The communication protocol based on LoRa technology was standardized by LoRA-Alliance in 2015, this protocol is called LoRaWAN. To improve the success rate for receiving messages sent by the end device, the end device sends the message to all stations near that device. The resulting duplicate receptions from this operation are filtered in the backend system (network server) which also has the intelligence to check security, send acknowledgments to the end de-

vice, and send the message to the corresponding application server.
In addition, LoRaWAN provides various classes of terminal devices to meet the different requirements of a wide range of IoT applications. Class A, Class B and Class C bidirectional terminal devices. Class A uses less power, while Class C uses the maximum amount of energy.

3.3. NB-IoT

NB-IoT is an LPWAN technology based on narrow-band radio technology and standardized by the 3rd Generation Partnership Project (3GPP). Its specifications were published in version 13 of the 3GPP in June 2016.
IoT NB is still in test in Europe. In December 2016, Vodafone and Huawei integrated NB-IoT into the Spanish Vodafone network and sent the first NB-IoT compliant message to a water meter. Huawei is currently expanding partnerships to deploy this technology around the world. In May 2017, the Ministry of Industry and Information Technology of China announced its decision to accelerate the commercial use of Io-NB for utilities and smart city applications.
NB-IoT can coexist with GSM (Global System for Mobile Communications) and LTE (Long Term Evolution) under licensed frequency bands (eg MHz, 800 MHz and 900 MHz).
The NB-IoT communication protocol is based on the LTE protocol. In fact, NB-IoT reduces the functionality of the LTE protocol to a minimum and improves it as required for IoT applications.
The improvement of NB-IoT continues with version 15 of 3GPP. Under the current 3GPP plan, the Io-NB will be extended to include location methods, multicast services (eg, terminal and message software update for a variety of devices), mobility and data storage. Other technical details applications of NB-IoT technology.

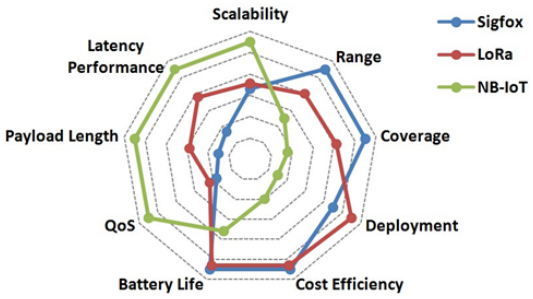


Figure 5. *Respective advantages of Sigfox, LoRa, and NB- IoT in terms of IoT factors [7]*

4. Smart agriculture applications projects

For farmers and producers, the Internet of Things has opened up extremely productive ways to grow crops and raise livestock, thanks to the use of cheap sensors. Intelligent agricultural applications are gaining ground with the promise of ubiquitous visibility on soil and crop health, machinery used, storage conditions, animal behavior and energy consumption. In this section, we discuss some IoT projects developed in intelligent agriculture.

4.1. CROPX SOIL MONITORING SYSTEM

Cropx produces hardware and software systems that measure moisture, temperature, and electrical conductivity in the soil (figure 6). The sensor reads measurements of humidity and temperature in different areas of the farmer's field, and sends this data over the internet to the server. The Cropx system analyzes and customizes irrigation plans for different areas of the field, this for better agricultural productivity and savings in water and energy.

Figure 6. *CROPX Soil Monitoring System*

4.2. MONITORING THE TEMPUTECH WIRELESS SENSOR

TempuTech offers an IoT solution for silo safety and monitoring and grain elevators. TempuTech has implemented an IoT system that allows farmers to understand their silo data with the wireless sensors installed at the silo level (figure 7). The platform enables manufacturers to establish benchmark performance standards and set alert and alarm conditions related to temperature, vibration, humidity and other conditions.

Figure 7. *TempuTech*

4.3. INTELLIGENT CLAAS EQUIPMENT

CLAAS is one of the world's leading manufacturers of agricultural engineering equipment. Their equipment can be controlled automatically. Their system provides advice to the farmer that minimizes grain loss and improves the flow of crops (figure 8). The system collects and makes good use of data through field mapping, fertilization planning, nutrient balance, and scheduling and planning programs.

Figure 8. *Intelligent CLAAS Equipment*

4.4. PRECISIONHAWK DRONE DATA PLATFORM

PrecisionHawk has created an autonomous UAV that collects high quality data through a series of sensors used for surveying, mapping and imaging farmland. The drone makes observations and a follow-up flight (figure 9). Using artificial intelligence, the drone can change course depending on the wind speed or the air pressure taken by sensors. During the flight, the drone collects visual, thermal and multispectral images. In this section,

Figure 9. *PRECISIONHAWK Drone Data Platform*

we have presented some projects in smart agriculture, these applications provide higher agricultural productivity and savings of water and energy and money to farmers. Other projects exist, such as Mb North America's connected cows, precision planting's corn maze, and Teamdev's Libelium network for tobacco crop quality.

5. Proposed architecture

In agriculture, the long life of the sensor battery is of great importance. Temperature, humidity and alkalinity sensors could significantly reduce water consumption and improve efficiency. The devices update the detected data only a few times a day because environmental conditions do not change dramatically. Thus, LPWAN technologies are ideal for smart agriculture applications. In Figure 10 we propose a typical IoT architecture for agricultural applications based on LPWAN networks. They represent the communication system that is the heart of the architecture. This system can be realized by three different technologies: Sigfox, LoRa and NB-IoT. We suggest the use of SigFox or LoRa because they use free frequency bands while NB-IoT requires LTE cellular coverage, which is not the case for most farms. The gateway receives data from different farm sensors installed in remote locations of the antennas, then, it transmits them to the server through an Internet access network. It stores them, analyzes them and makes them available to applications.

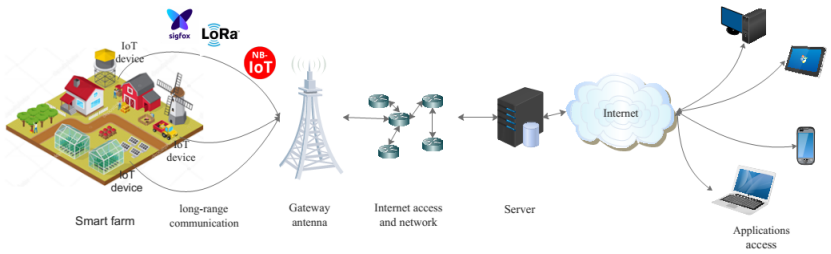


Figure 10. *Proposed architecture for smart farm*

6. Conclusion

Although vital, African agriculture is in difficulty, the sub-Sahara is made up of 95% of arable land which depend on the rain. This means that agricultural productivity is often low, making food insecurity a permanent danger. It is in this context that this paper is trying to sound the alarm to encourage researchers and governments to opt for a massive use of IoT through the design of appropriate architectures to help African agriculture in Africa. to get up. This paper provides basic architecture and communication system recommendations for efficient designs of intelligent farming systems based on connected components. We present particularly, a study of the different IoT technologies used in intelligent agriculture. This study shows the importance of LPWAN networks in the field of intelligent agriculture given the long range of communication (up to 40 KM) and low

energy consumption (one year of autonomy).
Finally, note that the success of IoT in Africa depends on close collaboration between companies, telecom providers, device suppliers and developers.

7. References

1 D. Kariuki, "The Internet of Things: Making Smart Farms in Africa," 15 01 2016.

2 S. Writer, "How is Internet of Things Shaping Up new African Businesses?," 2017.

3 P.-J. Benghozi, S. Bureau and F. Massit-Folléa, The Internet of Things , What Challenges for Europe?, Paris: La Maison des sciences de l'homme, 2009.

4 Kranenburg and R. van, "The Internet of Things: A Critique of Ambient Technology and the All-Seeing Network of RFID," vol. 2, 2008.

5 "The Internet of Everything," 2016. [Online]. Available: http://bewiser.eu/admin/resources/internetofeverything2016-2.pdf.

6 "Sigfox world coverage," [Online]. Available: www.sigfox.com/en/coverage/.

7 K. Mekki, E. Bajic, F. Chaxel and F. Meyer, "A comparative study of LPWAN technologies for large-scale IoT deployment," 2018.

8 B. JP, M. T and S. Seller O, "IoT: The era of LPWAN is starting now," in 42nd European Solid-State Circuits Conference, Lausanne, Switzerland, 2016.

9 [Online]. Available: https://www.cropx.com. [Accessed 23 03 2018].

10 R. S. Sinha, Y. Wei and S.-H. Hwang, "A survey on LPWA technology: LoRa and NB-IoT," ICT Express, 2017 .

11 2018. [Online]. Available: http://www.temputech.com/.

12 [Online]. Available: http://www.claas.co.uk/.

13 2018. [Online]. Available: http://www.precisionhawk.com/.

Images as sequence of points of an elliptic curve

Cidjeu Djeuthie Diderot*, Tieudjo Daniel**

Department of Mathematics and Computer Science
The University of Ngaoundere
PO Box 455 Ngaoundere
Cameroon
*cidjeu@gmail.com, **tieudjo@yahoo.com

RÉSUMÉ. Plusieurs transformations sont effectuées sur les images (Transformée en Cosinus Discrète, Transformée en Ondelettes Discrete, etc.) pour faciliter leur traitement et garantir leur sécurité. Cependant, ces transformations ne donnent pas toujours meilleure satisfaction lorsqu'on applique les algorithmes cryptographiques (crypto-compression par exemple) sur ces images. La cryptographie basée sur les courbes elliptiques offre de nos jours des performances remarquables. En vue d'appliquer la cryptographie basée sur les courbes elliptiques aux images, il est nécessaire de transformer ces images en séquences de points sur des courbes elliptiques. Dans ce papier, nous décrivons une méthode de transformation d'une image en séquence de points d'une courbe elliptique.

ABSTRACT. Several transformations are performed on images (Discrete Cosine Transform, Discrete Wavelet Transform, etc.) to facilitate their processing and ensure their security. However, these transformations do not always offer better satisfaction when applying cryptographic algorithms such as crypto-compression. Nowadays, Elliptic Curve Cryptography (ECC) have demonstrated remarkable performances in cryptography. To apply ECC on images, it is necessary to transform these images into sequences of points of elliptic curves. In this paper, we describe a method to transform an image into sequence of points of an elliptic curve.

MOTS-CLÉS : Crypto-compression, courbe elliptique, image

KEYWORDS : Crypto-compression; elliptic curve; image.

1. Introduction

Images are digital data transferable through public channels, thus need to be secured. The main solutions to secure images are watermarking and encryption. While watermarking enables image authentication, encryption ensures its confidentiality [1]. Several works have been done on the security of images ([2, 3, 4, 5, 6]). Generally, before performing encryption algorithms to images, various image transformations (Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT),... are used to facilitate image processing. Then quantification (compression) can be applied to reduce the size of the image. So, to practically secure images, compression is joined to encryption to obtain a hybrid process called crypto-compression. Compression algorithms aim to reduce the size of images, and such facilitate the transfer, storage, encryption, etc. Some crypto-compression algorithms can be found in [4, 5, 7, 8, 9, 10, 11]. Compression algorithms consider image as bytes matrix (table of digits) and most encryption algorithms used for image security are based on Number Theory (RSA, AES, DES,). Consequently, crypto-compression systems on images require keys of very large size, which is a problem in practice. Moreover, cryptosystems based on Number Theory are exposed to quantum attacks. Elliptic Curves Cryptography (ECC) presents several advantages compared to Number Theory based Cryptography : it offers smaller key sizes (160-bit for 1024-bit with RSA for example), has faster and more efficient implementation issues [12], etc. ECC has not yet been applied to image security. For image to be secured by ECC, it has to be seen as points of an elliptic curve.

In this paper, we propose an algorithm to transform an image into sequence of points of an elliptic curve. Following Koblitzs algorithm presented in [13], which transforms a character (letter, digit, etc) to a point of an elliptic curve, we describe how a pixel value can be represented as point on an elliptic curve. Finally, we describe how a whole image can be represented and seen as a sequence of points of a elliptic curve.

2. Preliminaries

Let $\mathbb{K}$ be a field of characteristic different from 2 and 3. An elliptic curve E over $\mathbb{K}$ is the set of points

$$E = \{O\} \cup \{(x, y) \in \mathbb{K} \times \mathbb{K}, y^2 = x^3 + ax + b\}$$

where O is a specific point called the point at infinity.
For example, Figure 1 below shows the elliptic curve $y^2 = x^3 - 2x - 2$ on the real field $\mathbb{R}$.

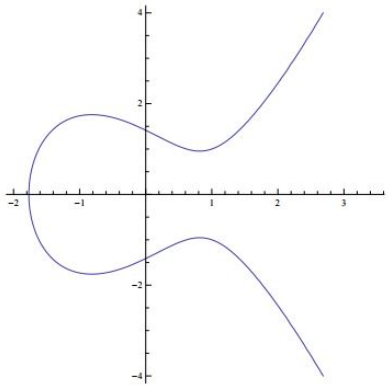


Figure 1. *Graph of the elliptic curve* $y^2 = x^3 - 2x - 2$ *on the real field* $\mathbb{R}$

More on elliptic curves and Elliptic Curve Cryptography (ECC) can be found in [12, 13, 14].

Below we will consider the field $\mathbb{K}$ to be the finite field $\mathbb{F}_q$, where $q = p^r$ for p prime and integer $r > 0$.

3. Transforming a character to a point of an elliptic curve

In [13], Koblitz described a process to transform a character to a point of an elliptic curve. A character c is represented as an integer m, such that $0 \leq m < M \in N$. For example, letters A to Z can be considered as integers between 0 and 25 ($M = 26$). For a given character c represented by an integer m, Algorithm 1 below computes a pair (x, y) which is a point of an elliptic curve, representing the given character.

Assume that we have a finite field $\mathbb{F}_q$ such that q is on the form $q = p^r$, p prime, $r > 0$; and $q \geq Mk + 1$, where k is generally set to 30 or 50. Given the curve $y^2 = x^3 + ax + b$ over the finite field $\mathbb{F}_q$ and given a character represented by an integer m.

Compute for each $j = 1, ...k$,

$$mk + j$$

Let x be the corresponding element of $mk + j$ in $\mathbb{F}_q$.

For such x, we compute $y^2 = f(x) = x^3 + ax + b$ and find a square-root for $f(x)$. If there exists a y such that $y^2 = f(x)$, the point of the elliptic curve representing m is $P_m = (x, y)$. If there is no square-root for $f(x)$ for the current j, we jump to the next j. With $k = 50$ the algorithm always return a good result [13]. This process is detailed in Algorithm 1.

From Algorithm 1, given a point (x, y) representing a character, this initial character m can be recovered by computing $\left\lfloor \frac{(\tilde{x}-1)}{k} \right\rfloor$, where $\lfloor v \rfloor$ represents the integer part of v and $\tilde{x}$ is the integer which corresponds to x in the equivalence between the integers and the elements of $\mathbb{F}_q$.

Algorithm 1 Transform a character to a point of an EC

Require: a character m, $\mathbb{F}_q$, k, a, b
Ensure: a pair $(x, y) \in \mathbb{F}_q \times \mathbb{F}_q$ representing m
1. j=1
2. while $j \leq k$
3. compute $\tilde{x} = mk + j$
4. write $\tilde{x}$ with r digits $m_{r-1}...m_1m_0$
5. compute $x = \sum_{i=0}^{r-1} m_i g^i \in \mathbb{F}_q$, where g is a generator of $\mathbb{F}_q$
6. compute $y^2 = f(x) = x^3 + ax + b$, and find a square-root for $f(x)$
7. if there exists a y such that $y^2 = f(x)$, then return $P_m = (x, y)$

else, increment j by 1.

4. Transforming image to points on elliptic curve

Pixel values are digits between 0 and 255. So, Algorithm 1 can be used to compute the point of the elliptic curve, representing each pixel value. Algorithm 2 shows how to compute the 256 points of an elliptic curve (E) representing the 256 possible pixel values. An illustration is presented below (Figure 2).

Algorithm 2 Transform pixel values to points on EC (PointsEC)

Require: an elliptic curve E over $\mathbb{F}_q$
Ensure: a sequence of points $(x, y) \in \mathbb{F}_q \times \mathbb{F}_q$ representing the 256 pixel values
1. points=[]
2. For each pixel value m between 0 and 255
2.1 execute Algorithm 1 to find P_m
2.2 add P_m to list
3. Return points

With this algorithm, for a given image, the sequence of points representing that image on the elliptic curve can be produced as presented in Algorithm 3.

At the end of this algorithm, of a given image I can be encrypted or processed as points of an elliptic curve.

When an image is so transformed to points of an elliptic curve, the original image can be recovered. The 256 points representing the 256 pixels values are also known as computed by Algorithm 2. Given a point representing a pixel value in an image, the index (rank) of that point in the list of 256 points computed in Algorithm 2 is the corresponding pixel value of the given point.

Finally, the original image can be reconstituted by substitution of each point by the corresponding pixel value.

Algorithm 3 Transform an image to points on EC

Require: an image I

Ensure: a sequence of points $(x, y) \in \mathbb{F}_q \times \mathbb{F}_q$ representing the given image

1. Define an elliptic curve E on the form $y^2 = x^3 + ax + b$ over $\mathbb{F}_q$
2. computes points=PointsEC(E)
3. imageEC=[]
4. For each pixel m in I
   4.1 add points[m] to imageEC
5. Return imageEC

5. Illustration

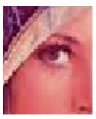
Figure 2 presents a list of the 256 points representing the 256 pixels values on the elliptic curve $y^2 = x^3 + x + 1$ over $\mathbb{F}_{7681}$. Implementation has been done using the computer algebra system SAGE [14]. With SAGE, the points are represented by triplets, which are their projective coordinates. For any point P, the projective coordinates $(X_P : Y_P : Z_P)$ correspond to the affine coordinates $(X_P/Z_P, Y_P/Z_P)$ if Z_P is non-zero, and O if Z_P is zero.

```
[(0 : 1 : 0), (0 : 1 : 1), (0 : 7680 : 1), (1 : 316 : 1), (1 : 7365 :
1), (2 : 196 : 1), (2 : 7485 : 1), (9 : 1621 : 1), (9 : 6060 : 1), (12 :
3734 : 1), (12 : 3947 : 1), (14 : 3674 : 1), (14 : 4007 : 1), (17 : 1331
: 1), (17 : 6350 : 1), (18 : 3831 : 1), (18 : 3850 : 1), (22 : 3483 :
1), (22 : 4198 : 1), (24 : 2822 : 1), (24 : 4859 : 1), (25 : 17 : 1),
(25 : 7664 : 1), (32 : 307 : 1), (32 : 7374 : 1), (33 : 538 : 1), (33 :
7143 : 1), (34 : 832 : 1), (34 : 6849 : 1), (35 : 3124 : 1), (35 : 4557
: 1), (37 : 388 : 1), (37 : 7293 : 1), (39 : 1254 : 1), (39 : 6427 : 1),
(40 : 1771 : 1), (40 : 5910 : 1), (50 : 1157 : 1), (50 : 6524 : 1), (51
: 1462 : 1), (51 : 6219 : 1), (55 : 648 : 1), (55 : 7033 : 1), (56 :
3810 : 1), (56 : 3871 : 1), (60 : 2050 : 1), (60 : 5631 : 1), (62 : 2594
: 1), (62 : 5087 : 1), (63 : 1532 : 1), (63 : 6149 : 1), (65 : 612 : 1),
(65 : 7069 : 1), (67 : 3784 : 1), (67 : 3897 : 1), (68 : 1928 : 1), (68
: 5753 : 1), (71 : 2015 : 1), (71 : 5666 : 1), (72 : 611 : 1), (72 :
7070 : 1), (73 : 71 : 1), (73 : 7610 : 1), (74 : 3411 : 1), (74 : 4270 :
1), (75 : 2416 : 1), (75 : 5265 : 1), (76 : 1693 : 1), (76 : 5988 : 1),
(78 : 3751 : 1), (78 : 3930 : 1), (79 : 96 : 1), (79 : 7585 : 1), (81 :
886 : 1), (81 : 6795 : 1), (82 : 1520 : 1), (82 : 6161 : 1), (83 : 3141
: 1), (83 : 4540 : 1), (84 : 2903 : 1), (84 : 4778 : 1), (85 : 3547 :
1), (85 : 4134 : 1), (87 : 2441 : 1), (87 : 5240 : 1), (88 : 259 : 1),
(88 : 7422 : 1), (92 : 766 : 1), (92 : 6915 : 1), (95 : 1404 : 1), (95 :
6277 : 1), (97 : 1865 : 1), (97 : 5816 : 1), (99 : 1771 : 1), (99 : 5910
: 1), (100 : 3290 : 1), (100 : 4391 : 1), (101 : 3334 : 1), (101 : 4347
: 1), (102 : 1526 : 1), (102 : 6155 : 1), (105 : 399 : 1), (105 : 7282 :
1), (107 : 186 : 1), (107 : 7495 : 1), (108 : 3335 : 1), (108 : 4346 :
1), (110 : 570 : 1), (110 : 7111 : 1), (111 : 2821 : 1), (111 : 4860 :
1), (112 : 902 : 1), (112 : 6779 : 1), (117 : 1100 : 1), (117 : 6581 :
1), (118 : 2446 : 1), (118 : 5235 : 1), (119 : 1110 : 1), (119 : 6571 :
1), (121 : 2338 : 1), (121 : 5343 : 1), (124 : 3398 : 1), (124 : 4283 :
1), (126 : 2602 : 1), (126 : 5079 : 1), (130 : 3080 : 1), (130 : 4601 :
1), (131 : 2201 : 1), (131 : 5480 : 1), (132 : 3767 : 1), (132 : 3914 :
1), (133 : 1274 : 1), (133 : 6407 : 1), (134 : 2357 : 1), (134 : 5324 :
1), (137 : 1479 : 1), (137 : 6202 : 1), (141 : 1003 : 1), (141 : 6678 :
1), (146 : 1876 : 1), (146 : 5805 : 1), (147 : 1547 : 1), (147 : 6134 :
1), (148 : 3350 : 1), (148 : 4331 : 1), (150 : 3769 : 1), (150 : 3912 :
1), (151 : 513 : 1), (151 : 7168 : 1), (153 : 1274 : 1), (153 : 6407 :
1), (155 : 649 : 1), (155 : 7032 : 1), (157 : 532 : 1), (157 : 7149 :
1), (161 : 2321 : 1), (161 : 5360 : 1), (162 : 3254 : 1), (162 : 4427 :
1), (163 : 532 : 1), (163 : 7149 : 1), (166 : 3404 : 1), (166 : 4277 :
1), (173 : 718 : 1), (173 : 6963 : 1), (174 : 3523 : 1), (174 : 4158 :
1), (175 : 2520 : 1), (175 : 5161 : 1), (178 : 307 : 1), (178 : 7374 :
1), (182 : 1658 : 1), (182 : 6023 : 1), (183 : 495 : 1), (183 : 7186 :
1), (185 : 2501 : 1), (185 : 5180 : 1), (189 : 1744 : 1), (189 : 5937 :
1), (190 : 2638 : 1), (190 : 5043 : 1), (191 : 3656 : 1), (191 : 4025 :
1), (192 : 3281 : 1), (192 : 4400 : 1), (193 : 1938 : 1), (193 : 5743 :
1), (198 : 2583 : 1), (198 : 5098 : 1), (201 : 3778 : 1), (201 : 3903 :
1), (203 : 128 : 1), (203 : 7553 : 1), (206 : 1847 : 1), (206 : 5834 :
1), (208 : 1517 : 1), (208 : 6164 : 1), (209 : 1648 : 1), (209 : 6033 :
1), (211 : 2594 : 1), (211 : 5087 : 1), (212 : 1897 : 1), (212 : 5784 :
1), (213 : 2089 : 1), (213 : 5592 : 1), (216 : 21 : 1), (216 : 7660 :
1), (217 : 2874 : 1), (217 : 4807 : 1), (218 : 3814 : 1), (218 : 3867 :
1), (220 : 1283 : 1), (220 : 6398 : 1), (221 : 2158 : 1), (221 : 5523 :
1), (222 : 3516 : 1), (222 : 4165 : 1), (224 : 2415 : 1), (224 : 5266 :
1), (225 : 3409 : 1), (225 : 4272 : 1), (226 : 172 : 1), (226 : 7509 :
1), (228 : 1956 : 1), (228 : 5725 : 1), (229 : 2492 : 1), (229 : 5189 :
1), (230 : 2784 : 1), (230 : 4897 : 1), (231 : 1800 : 1), (231 : 5881 :
1), (232 : 76 : 1), (232 : 7605 : 1), (237 : 2989 : 1), (237 : 4692 :
1), (239 : 2026 : 1), (239 : 5655 : 1), (240 : 3279 : 1), (240 : 4402 :
1), (242 : 1740 : 1), (242 : 5941 : 1), (243 : 3572 : 1), (243 : 4109 :
1), (245 : 3077 : 1), (245 : 4604 : 1), (248 : 3190 : 1), (248 : 4491 :
1), (250 : 3676 : 1), (250 : 4005 : 1), (251 : 2141 : 1), (251 : 5540 :
1), (253 : 1031 : 1), (253 : 6650 : 1), (254 : 1161 : 1), (254 : 6520 :
1), (255 : 3824 : 1)]
```

Figure 2. *List of points corresponding to the 256 pixels values on the elliptic curve* $y^2 = x^3 + x + 1$ *over* $\mathbb{F}_{7681}$

The next figures (Figure 3 and Figure 4) show an image and a sequence of points representing a part of that image.

Figure 3. *Original image "lena"*



[(68 : 91 : 1), (70 : 74 : 1), (79 : 197 : 1), (75 : 219 : 1), (70 : 74 : 1), (60 : 204 : 1), (54 : 202 : 1), (52 : 58 : 1), (48 : 204 : 1), (48 : 47 : 1), (69 : 32 : 1), (87 : 192 : 1), (101 : 198 : 1), (107 : 32 : 1), (98 : 36 : 1), (69 : 32 : 1), (57 : 55 : 1), (72 : 109 : 1), (68 : 160 : 1), (39 : 25 : 1), (48 : 204 : 1), (42 : 191 : 1), (20 : 95 : 1), (92 : 44 : 1), (121 : 80 : 1), (92 : 207 : 1), (87 : 59 : 1), (140 : 65 : 1), (167 : 122 : 1), (150 : 74 : 1), (143 : 47 : 1), (161 : 98 : 1), (173 : 152 : 1), (161 : 98 : 1), (180 : 14 : 1), (190 : 37 : 1), (167 : 129 : 1), (72 : 142 : 1), (31 : 74 : 1), (41 : 213 : 1), (36 : 209 : 1), (48 : 204 : 1), (47 : 31 : 1), (45 : 73 : 1), (45 : 178 : 1), (48 : 204 : 1), (47 : 220 : 1), (42 : 191 : 1), (41 : 213 : 1), (42 : 191 : 1), (48 : 204 : 1), (54 : 202 : 1), (54 : 202 : 1), (54 : 49 : 1), (60 : 204 : 1), (69 : 219 : 1), (69 : 219 : 1), (62 : 106 : 1), (59 : 240 : 1), (62 : 145 : 1), (69 : 219 : 1), (53 : 29 : 1), (68 : 160 : 1), (59 : 11 : 1), (52 : 58 : 1), (69 : 219 : 1), (81 : 174 : 1), (75 : 219 : 1), (69 : 219 : 1), (57 : 55 : 1), (45 : 178 : 1), (37 : 95 : 1), (87 : 192 : 1), (138 : 50 : 1), (122 : 194 : 1), (81 : 174 : 1), (110 : 73 : 1), (180 : 237 : 1), (160 : 208 : 1), (148 : 5 : 1), (155 : 114 : 1), (160 : 208 : 1), (164 : 170 : 1), (179 : 13 : 1), (190 : 214 : 1), (202 : 111 : 1), (119 : 124 : 1), (31 : 177 : 1), (36 : 42 : 1), (59 : 240 : 1), (58 : 12 : 1), (68 : 91 : 1), (63 : 93 : 1), (63 : 158 : 1), (63 : 93 : 1), (69 : 219 : 1), (73 : 215 : 1), (70 : 74 : 1), (60 : 204 : 1), (69 : 32 : 1), (80 : 215 : 1), (81 : 174 : 1), (75 : 219 : 1), (80 : 36 : 1), (92 : 207 : 1), (83 : 120 : 1), (81 : 77 : 1), (87 : 192 : 1), (98 : 36 : 1), (98 : 36 : 1), (82 : 249 : 1), (72 : 142 : 1), (59 : 11 : 1), (47 : 31 : 1), (59 : 11 : 1), (82 : 2 : 1), (81 : 77 : 1), (68 : 160 : 1), (51 : 210 : 1), (36 : 42 : 1), (37 : 95 : 1), (75 : 32 : 1), (122 : 194 : 1), (148 : 5 : 1), (107 : 219 : 1), (105 : 133 : 1), (140 : 65 : 1), (169 : 160 : 1), (159 : 227 : 1), (150 : 74 : 1), (147 : 39 : 1), (156 : 82 : 1), (169 : 160 : 1), (176 : 22 : 1), (173 : 152 : 1), (169 : 160 : 1), (91 : 101 : 1), (51 : 210 : 1), (41 : 38 : 1), (48 : 204 : 1), (54 : 202 : 1), (63 : 158 : 1), (73 : 215 : 1), (81 : 77 : 1), (82 : 249 : 1), (96 : 73 : 1), (101 : 53 : 1), (101 : 53 : 1), (98 : 215 : 1), (101 : 53 : 1), (108 : 110 : 1), (113 : 99 : 1), (112 : 199 : 1), (48 : 204 : 1), (47 : 31 : 1), (58 : 239 : 1), (62 : 106 : 1), (39 : 25 : 1), (81 : 77 : 1), (110 : 178 : 1), (96 : 178 : 1), (73 : 215 : 1), (69 : 219 : 1), (72 : 142 : 1), (97 : 117 : 1), (91 : 150 : 1), (70 : 74 : 1), (57 : 196 :

Figure 4. *Sequence of points representing a part of image "lena"*

6. Conclusion and perspectives

We presented how to transform an image into a sequence of points of an elliptic curve. With such representation, encryption schemes and various algorithms of Elliptic Curve Cryptography can be applied on images. Some other operations in image processing as watermarking, compression, can also be redefined on images seen as points on elliptic curve. These open doors for new interests in research on image processing and security.

7. Bibliographie

[1] Lafourcade P., *Security and Cryptography just by images*, Université Joseph Fourier, Verimag DCS, 2009

[2] Abdmouleh M.K., Bouhlel M.S., *Effective Crypto-compression Scheme for Medical Images*, International Journal of Signal Processing, 2017

[3] Miao Z., Xiaojun T., *Joint image encryption and compression scheme based on IWT and SPIHT*, Optics and Lasers in Engineering, vol 90, pp 254-274, 2017

[4] Xiaoyong J., Sen B., Guibin Z. and Bing Y., *Image encryption and compression based on the generalized knights tour, discrete cosine transform and chaotic maps*, Multimedia Tools and Applications, vol. 76, Issue 10, pp 1296512979, 2017

[5] Puech W. and Rodrigues J.M., *Crypto-Compression of medical images by selective encryption of DCT*, In EUSIPCO'05, Antalya, Turquie, Septembre 2005

[6] Puech W., Rodrigues J.M. et Develay-Morice J.E., *Transfert sécurisé d'images médicales par codage conjoint : cryptage sélectif par AES en mode par flot et compression JPEG*, Traitement du signal (TS), numéro spécial Traitement du signal appliqué à la cancérologie, vol. 23, n°5, 2006

[7] Waghmare A., Bhagat A., Surve A., Kalgutkar S., *Chaos Based Image Encryption and Decryption*, IJARCCE, vol 5, 2016

[8] Benabdellah M, Majid H.M., Zahid N., Regragui F. and Bouyakhf E.H., *Encryption-Compression of still images using the FMT transformation and the DES algorithm*, International Journal of Computer Sciences and Telecommunications, No. 4, 2006

[9] Benabdellah M, Majid H.M., Zahid N., Regragui F. and Bouyakhf E.H., *Encryption-compression of images based on FMT and AES algorithm*, International Journal Applied Mathematical Sciences, Vol. 1, 2007

[10] Jalel H., Mohamed A., Ben F., Mounir S. and Abdennaceur K., *Crypto-compression of images based on chaos*, IEEE, 2013

[11] Masmoudi A., Puech W., *Lossless chaos-based crypto-compression scheme for image protection*, ITE Image Processing, vol 8, 2014

[12] Bos J. W., Halderman J. A., Heninger N.,Moore J., Naehrig M.and Wustrow E., *Elliptic Curve Cryptography in Practice*, IACR, 2013

[13] Koblitz N., *A Course in Number Theory and Cryptography - 2nd ed.*, Springer-Verlag, 1994

[14] Stein W., *Sage Tutorial - Release 8.1*, The Sage Development Team, 2017

Novel approach to maximise the lifetime of Wireless sensor networks

Hierarchical protocols LEACH and PEGASIS

Fatima Es-sabery[1] — Abdellatif Hair[2]

Faculty of Sciences and Technology
Sultan Moulay Slimane University
B.P. 523, Beni Mellal
MOROCCO

[1]fatima.essabery@gmail.com

[2] abd_hair@yahoo.com

ABSTRACT. The hierarchical routing of data in WSNs is a specific class of routing protocols it encompasses solutions that take a restructuring of the physical network in a logical hierarchy system for the optimization of the consumption of energy. Several hierarchical routing solutions proposed, namely: the protocol LEACH (Low Energy Adaptive Clustering Hierarchy) consist of dividing the network in distributed clusters at one pop in order of faster data delivery and PEGASIS protocol (Power-Efficient Gathering in Sensor Information Systems) which uses the principle of constructing a chain's sensor node. Our contribution consists of a hierarchical routing protocol, which is the minimization of the energy consumption by reducing the transmission distance of data and reducing the data delivery time. Our solution combines the two hierarchical routing approaches: chain based approach and the cluster based approach. Our approach allows for multi-hop communications, intra- and inter- cluster, and a collaborative aggregation of data in each Cluster, and a collaborative aggregation of data at each sensor node.

KEYWORDS: Hierarchical routing, LEACH, Optimization of energy, PEGASIS, WSNs

1. Introduction

Wireless sensor networks (WSN) consist of a large number of devices known as sensors. These are equipped with the ability to collect physical quantities such as temperature, pressure, pH, etc. in a study area. Then, they perform a processing on the collected data before they cooperate among them to route it to a control center called base station. Due to the small size of the sensors and their low cost of production, WSN offer numerous practical applications; these applications may be sensitive especially in the military, medical, environmental, etc. [1].

Due to the miniaturization constraints [2], the nodes typically have very limited resources in terms of computing capacity, data storage space, transmission and energy flow. These limits are part of the research questions in the field of wireless sensor networks. In particular, the constraint linked to energy is a fundamental problem. Indeed, all elements need energy to operate; the control of energy consumption of a node remains a major problem for maximizing its lifetime [3, 4, 5].Hierarchical routing is considered as a powerful tool as regards to the minimization of the energy consumption compared to other types of routing. Our contribution is to propose a new hybrid approach based on hierarchical protocols [6, 7, 8, 9]. Hierarchical routing is considered a powerful tool for minimizing power consumption compared to other types of routing. Our contribution consists to propose a new hybrid approach based on hierarchical protocols. The rest of this article is structured as follows:

In the next section, we presented the energy consumption model. Then we present our hybrid approach in section 3. Finally we evaluate the performance of our approach in section 4, before concluding and presenting the perspectives of this work.

2. Model of Energy Consumption

The sensor node consumes energy to perform three main tasks: detecting, communication and data processing. The energy used for the detection of physical phenomena is not very important. As well as the one used for the treatment is lower than the energy of communication. For example, the necessary energy to transmit 1KB over a distance of 100m is approximately equivalent to the energy needed to run 3 million instructions with a speed of 100 million instructions per second. While the necessary energy for processing the data is calculated by applying the following formula :

$$E_{DA} = 5\,\text{nanojoule}/\,1\,\text{bit} \quad (1)$$

Since communications dissipate much more energy than other tasks, a power radio's consumption model is proposed by Heinzelman et al. [10] Thus, the necessary energies to emit E_{tx} and receive E_{rx} messages are given by:

To send a message of k bits over a distance of d meters, the transmitter consumes:

$$\begin{aligned} E_{tx}(k,d) &= E_{tx-elec}(k) + E_{tx-amp}(k,d) \\ &= k.E_{elec} + k.E_{fs}.d^2 \quad si\ d < d_0 \\ &= k.E_{elec} + k.E_{amp}.d^4 si\ d \geq d_0 \end{aligned} \tag{2}$$

To receive a message of k bits, the receiver consumes:

$$E_{rx}(k) = k.E_{elec} \tag{3}$$

Where $E_{tx-elec}(k)$ It is the energy transmission, $E_{tx-amp}(k,d)$ it is the amplification energy E_{elec} is the amount of energy consumed by a bit and E_{fs} is the signal amplification in a lower distance to the threshold distanced_0. If the distance transmission is superior to d_0 the amplification E_{amp} is used Such as:

$$d_0 = \sqrt{\frac{E_{fs}}{E_{amp}}} \tag{4}$$

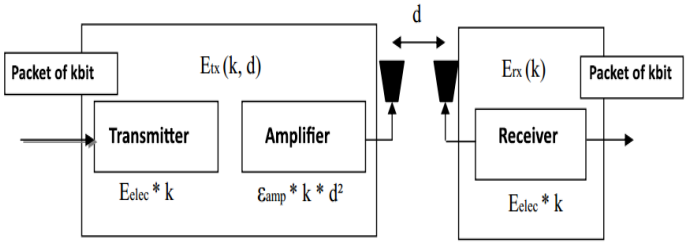


Figure 1. *A model of energy consumption*

3. Hybrid approach proposed

After analyzing the two algorithms (LEACH and PEGASIS) [3,6], we noticed that we can improve the first protocol (LEACH) by applying the concept of the second protocol (PEGASIS) within groups (cluster) and at the level of cluster heads, this leads us to propose a new hybrid protocol which combines the advantages of two broad approaches which are (clustered approach) and (chained approach).

3.1. Basic Concepts of our protocol

The proposed algorithm consists of combining the two protocols PEGASIS and LEACH according to two major steps:

Step 1: Application of PEGASIS within the cluster

The organization of the nodes those belong to the same group (cluster) in a chain can improve and regulate the energy dissipation, which reduces the load into cluster-head. Actually, the nodes communicate only with their close neighbors and not directly with their cluster-head, which saves the energetic consumption and offers better use of the bandwidth. The aggregation of data at each node between nodes and cluster-head, this is the consequence of preserving energy reserves in the nodes and cluster-head.

The figure below shows how the nodes are organized in groups (clusters), the C0 node transmits its data to its nearest neighbor C1, C1 aggregates the data received with its own and transmits them to its neighbor until they reach the leader node which transmits them to the CH(cluster-head). So in this first organizational step (chain group), all nodes in the cluster will transmit their data collected in their respective CHs (cluster-heads) by connecting them through the chain, while each CH receive the collected data by the leader node (the nearest node CH) of the chain.

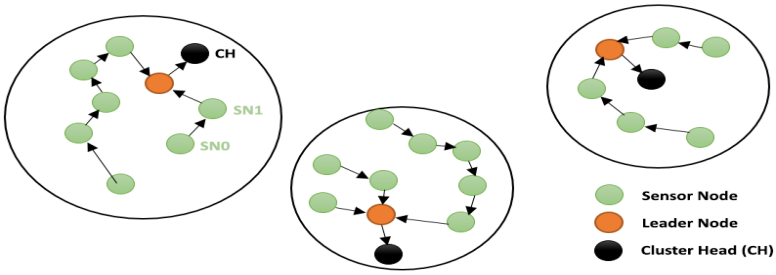


Figure 2. *Organization of nodes of a cluster as chain*

Step 2: Application of PEGASIS at Cluster heads level

The principle of this second step is to organize cluster head nodes as a closely neighbor chain. In order to prevent the farthest cluster-heads from the base station to die quickly, thereby, aggregating the data at each cluster-head reduces the number of transmissions to the base station to a single transmission carried out by the cluster-head leader of the chain, which also reduces the load on the base station. This allows to save and to regulate the energy consumption by cluster-heads.

The below figure shows how the cluster-heads nodes will be organized, the node CH_0transmits its data to its nearest neighbor CH_1, CH_1 aggregates the data received with its own and transmits them to its neighbor until the cluster-head leader which transmits them to the base station. So in this first organizational step (chain group), all cluster-heads nodes will transmit their collected data respectively to the base station by being connected to the chain, while the base station must receive the data collected by the leader node (the closest node base station) in the chain.

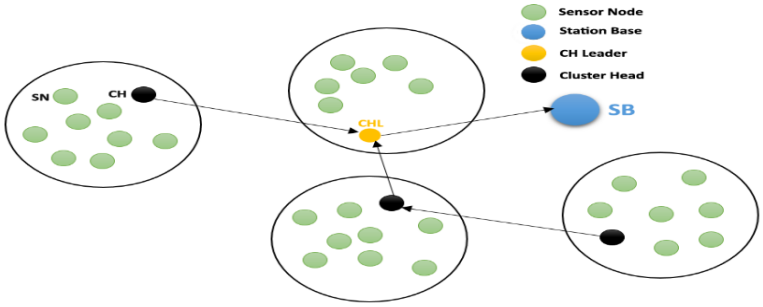


Figure 3. *Organization of cluster heads as chain*

To summarize, our approach is used to improve the LEACH protocol by using basic concepts of PEGASIS protocol, this improvement can change in the LEACH's topology as shown in the figure below :

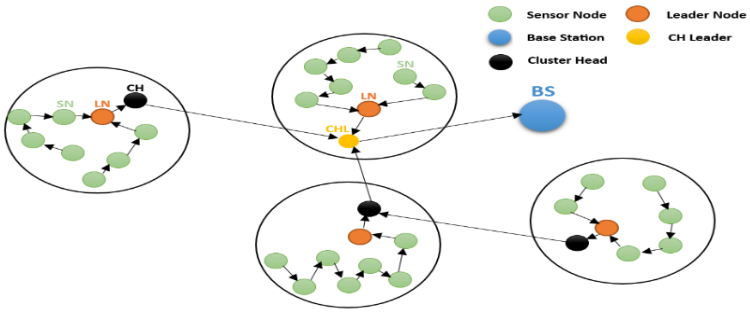


Figure 4. *Topology of our hybrid approach*

3.2. Major steps of our algorithm

The progress of our hybrid protocol is divided into several execution cycles (Figure 11). Each cycle begins with an initialization phase in which the chain clusters are formed. The CHs are elected and the CH chain is formed, followed by a transmission phase where the collected data is transferred through the chains to the CHs which in turn transmit them to the control center through the CH chain. Nodes must all be synchronized to participate in the initialization phase at the same time

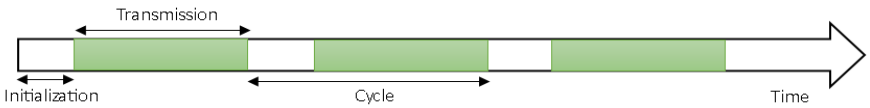


Figure 5. *Stages of execution of our hybrid approach*

3.2.1. Initialization step

The initialization step begins with the creation of the groups in which we adopt the same approach used in centralized LEACH-C, where the base station uses the simulated success to form groups. This approach provides a better result compared to the distributed approach, used in LEACH, in terms of forming groups and energy conservation. After the formation of groups, cluster-heads are selected in a simplified way where only the node that has the largest reserve of energy among the nodes of the same group, is elected. Then we approach the construction of two chains, chain linked the same cluster member nodes and other nodes linked the cluster-heads where a centralized method is followed in which the base station uses the information sent by the nodes to form chain using the chain forming algorithm proposed by PEGASIS algorithm.

3.2.2. Transmission step:

The transmission step is divided into several iterations in which nodes will transmit their collected data, through the chain, to the cluster-heads. In addition, these cluster-heads transmit in turn, their data through the chain they form to the base station. In each iteration, a node transmits at least one data packet during its time slot previously allocated by the base station. Knowing that the time slot allocated to each node is constant, the time for each iteration of transmission will obviously depend on the number of existing nodes in each cluster and the number of cluster-heads.

The transmission phase in our approach is divided into two stages: the first stage concerns the intra_cluster transmission and the second stage concerns the inter-cluster transmission. In the first stage, the members of each cluster node transmit their collected data through the chain to their cluster-head. After the cluster-heads receives data, the

transmission process proceeds to the second step of transmitting the collected data by the cluster-heads through the channel to the base station. We can summarize this phase by the following:

- Data collected through the sensor nodes.
- Transfer of aggregated and collected data from neighboring nodes through the chain to cluster-heads.
- The cluster-heads transmit the received data from the closest neighbor through the chain to the base station.

4. Evaluation of our approach

The simulation of our hybrid algorithm is the most important stage in our work since we can prove the improvements made in terms of energy savings and overall lifetime of the network using the results provided. The performance analysis of our hybrid routing algorithm is evaluated using MATLAB. The results from the simulation are compared with LEACH algorithms and PEGASIS in terms of network lifetime.

To compare the lifetime of the network between the two algorithms LEACH and PEGASIS and our algorithm, we measured the residual energy of sensor nodes for each iteration to determine the number of communication rounds when 1%, 20%, 50% and 100% of nodes die, we reused and reconfigured according to our parameters, simulation information LEACH and PEGASIS protocols provided in [11] and [12], and compared with the results of our simulation, the result is given as a graph in below Figures:

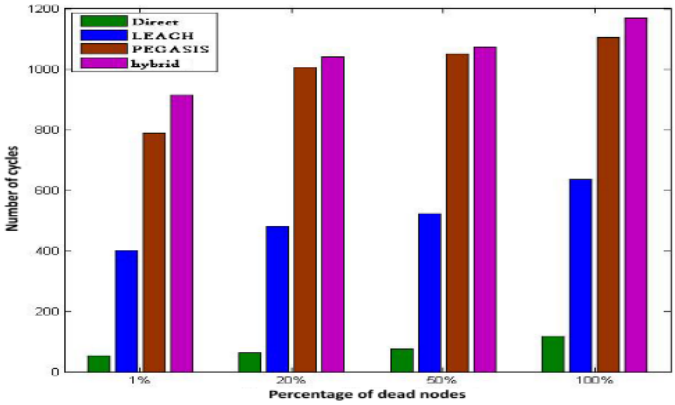


Figure 7. *The performance results of a network of 50m x 50m with an initial energy of 0.25j / node*

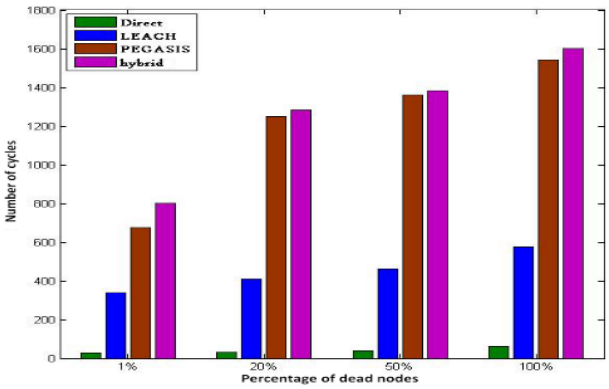


Figure 8. *The performance results of a network of 100m x 100m with an initial energy of 0.5J / node*

Based on simulation results, we have shown that our hybrid algorithm improves the energy dissipation inside and outside clusters, increases energy gain, and extends the lifetime of the network from 50% to 75% compared to the LEACH protocol and from 10% to 17% compared to PEGASIS protocol.It remains to be noted that our algorithm provided the best value, since it increases the lifetime of the network compared to LEACH protocol, and significantly reduces the extreme latency introduced by the protocol PEGASIS.

5. Conclusion

Motivated by the extreme latency introduced by the long nodes chain in the PEGASIS protocol and poor energy dissipation in the LEACH protocol, where cluster heads and farthest nodes die faster than others nodes, we tried to propose a new algorithm that combines the benefits of both protocols in order to reduce their disadvantages and provide a best value life / latency. To validate the improvements made by our protocol in terms of extending the network lifetime and the effective management of energy consumption, we simulated the operation of our algorithm with MATLAB and compared the results with those provided protocol LEACH and PEGASIS.

The results from the simulation show that our protocol offers better power management compared to LEACH and PEGASIS protocols. In addition, the degree of latency caused by the long chain in the PEGASIS protocol is significantly reduced.

REFERENCES

[1] Rajendra Prasad Mahapatra, Rakesh Kumar Yadav.,"*Descendant of LEACH Based Routing Protocols in Wireless Sensor Networks*", vol. 57, pp. 1005-1014, 2015.

[2] Vishal Kumar Arora, Vishal Sharma, Monika Sachdeva., *"A survey on LEACH and other's routing protocols in wireless sensor network"*, vol. 127, pp. 6590-6600, August 2016.

[3] Abbas Nayebi, Hamid Sarbazi-Azad., *"Performance modeling of the LEACH protocol for mobile wireless sensor networks"*, vol. 71, pp. 812-821, June 2011.

[4] Madhura Mahajan, K.T.V. Reddy, Manita Rajput., "Design and Simulation of a Blacklisting Technique for Detection of Hello Flood Attack on LEACH Protocol", vol. 79, pp. 675-682, 2016.

[5] V. Geetha, P.V. Kallapur, Sushma Tellajeera., *clustering in Wireless Sensor Networks: Performance Comparison of LEACH & LEACH-C Protocols Using NS2*, vol. 4, pp. 163-170, 2012.

[6] Rina Mahakud, Satyanarayan Rath, Minu Samantaray, BabySradha Sinha, Priyanka Priya, Ananya Nayak, Aarti Kumari., "*Energy Management in Wireless Sensor Network Using PEGASIS*", vol. 92, pp. 207-212, 2016.

[7] Young-Long Chen, Jia-Sheng Lin., "*Energy efficiency analysis of a chain-based scheme via intra-grid for wireless sensor Networks*", vol. 35, pp. 507-516, 2012.

[8] Vishal Kumar Arora, Vishal Sharma, Monika Sachdeva.,"*A survey on LEACH and other's routing protocols in wireless sensor network*", vol. 127, pp. 6590-6600, 2016.

[9] Abbas Nayebi, Hamid Sarbazi-Azad., *"Performance modeling of the LEACH protocol for mobile wireless sensor networks"*, vol. 71, pp. 812-821, 2011.

[10] Geetha, P.V. Kallapur, Sushma Tellajeera*., "Clustering in Wireless Sensor Networks: Performance Comparison of LEACH & LEACH-C Protocols Using NS2*", vol. 4, pp. 163-170, 2012.

[11] Jian Shen, Anxi Wang, Chen Wang, Yongjun Ren, Jin Wang*., "Performance Comparison of Typical and Improved LEACH Protocols in Wireless Sensor Network*", pp. 161-166, 2016.

[12] Binkal S Ahir, Rohan Parmar, Bintu Kadhiwala*.,"Energy efficient clustering algorithm for data aggregation in wireless sensor network",* pp. 683-688, 2015.

Interfaces of Roles

in Distributed Collaborative Systems

Eric Badouel[(a)] and Rodrigue Aimé Djeumen Djatcha[(b)]

(a) Inria Rennes-Bretagne Atlantique, Irisa, University of Rennes I,
Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France
(b) Faculty of Sciences, University of Douala, BP 24157 Douala, Cameroon
eric.badouel@inria.fr djeumenr@yahoo.fr
This work was partially supported by ANR Headwork.

ABSTRACT. We address the problem of component reuse in the context of service-oriented programming and more specifically for the design of user-centric distributed collaborative systems modelled by Guarded Attribute Grammars. Following the contract-based specification of components we developp an approach to an interface theory for the roles in a collaborative system in three stages: we define a composition of interfaces that specifies how the component behaves with respect to its environement, we introduce an implementation order on interfaces and finally a residual operation on interfaces characterizing the systems that, when composed with a given component, can complement it in order to realize a global specification.

RÉSUMÉ. Nous abordons le problème de la réutilisation des composants dans le contexte de la programmation orientée services et plus spécifiquement pour la conception de systèmes collaboratifs distribués centrés sur l'utilisateur modélisés par des grammaires attribuées gardées. En suivant la démarche de la spécification contractuelle des composants, nous développons une approche de la théorie des interfaces pour les rôles d'un système collaboratif en trois étapes: on définit une composition d'interfaces qui spécifie comment le composant se comporte par rapport à son environnement, on introduit un ordre d'implémentation sur les interfaces et enfin une opération de résidus sur les interfaces qui caractérise les systèmes qui, lorsqu'ils sont composés avec un composant donné, peuvent le compléter afin de réaliser une spécification du système global.

KEYWORDS : Component Based Design, Service Oriented Programming, Interface, Role, Collaborative System, Guarded Attribute Grammars

MOTS-CLÉS : Conception à base de composants, Programmation orientée services, Interface, rôle, systèmes collaboratif, grammaires attribuées gardées

1. Introduction

We address the problem of component reuse in the context of service-oriented programming and more specifically for the design of user-centric distributed collaborative systems. The role of a specific user is given by all the services he or she offers to the environment. A role can be encapsulated by a module whose interface specifies the provided services the module exports and the required external services that it imports. Usually the modules in a service-oriented design are organized hierarchically. In contrast, modules in a distributed collaborative systems would often depend on each other (even though cyclic dependencies between services should be avoided). Moreover services that are currently activated can operate as coroutines and a service call can activate new services in a way that may depend on the user's choice of how to provide the service. We thus need a richer notion of interface for roles in a distributed collaborative system. In this paper we consider a very simple extension of the concept of interface obtained by adding a binary relation on the set of services indicating for each of the provided services the list of services that are potentially required to carry it out. This relation gives only *potential dependencies* because a user can provide a service in various ways and relying on a variety of external services. We motivate our presentation in the context of systems modelled by Guarded Attribute Grammars [4]. Possible extensions of this basic model of interface are mentioned in the concluding section. They would provide finer descriptions of the behaviour of a module in a Guarded Attribute Grammar specification.

Even if the objectives differ (service-oriented design versus verification) as well as the models used (user-centric collaborative systems versus reactive systems) we are largely inspired by the works that have been carried out on behavioural interfaces of communicating processes. Three main ingredients have been put forward in these studies which will serve as our guideline.

First, an interface is mainly used to formalize a contract-based reasoning for components. The idea is that a component of a reactive system [5] is required to behave correctly only when its environment does. The correctness of composition is stated in terms of a contract given by *assume-guarantee* conditions: the component should guarantee some expected behaviour when plugged into an environment that satisfies some properties. The principle of composition is however made subtle by the fact that each component takes part in the others' environment [1]. Safety and liveness properties, which are not relevant in our case, are crucial issues in this context and largely contribute to the complexity of the resulting formalisms. The underlying models of a component range from process calculi [2] to I/O automata and games [3]. These interface theories have also been extended to take some qualitative aspects (time and/or probability) into account.

Second, an interface is viewed as an abstraction of a component, a so-called *behavioural type*. Thus we must be able to state when a component satisfies an interface, viewed as an abstract specification of its behaviour. A relation of refinement, given by a pre-order $I_1 \leq I_2$, indicates that any component that satisfies I_2 also satisfies I_1. In the context of service-oriented programming we would say that interface I_2 *implements* interface I_1.

Third, a notion of *residual specification* has also proved to be useful. The problem was first stated in [6] as a form of equation solving on specifications. Namely, given a specification G of the desired overall system and a specification C of a given component we seek for a specification X for those systems that when composed with the component satisfies the global property. It takes the form of an equation $L \bowtie X \approx G$ where $\bowtie$

stands for the composition of specifications and $\approx$ is some equivalence relation. If $\approx$ is the equivalence induced by the refinement relation the above problem can better be formulated as a Galois connection [9] $G/L \leq X \iff G \leq L \bowtie X$ stating that the residual specification G/L is the smallest (i.e. less specific or more general) specification that when composed with the local specification is a refinement of the global specification. Since $L \bowtie -$ is monotonous (due to Galois connection) it actually entails that a component is an implementation of the residual specification if and only if it provides an implementation of the global specification when composed with an implementation of the local specification .

2. Roles in a Collaborative System Modelled by a Guarded Attribute Grammar

Guarded Attribute Grammars (GAG) [4] is a user centric model of collaborative work that puts emphasis on task decomposition and the notion of user's workspace. We assume that a workspace contains, for each service offered by the user, a repository that contains one artifact for each occurrence of a service call (that initiates a so-called *case* in the system). An artifact is a tree that records all the information related to the treatment of the case. It contains open nodes corresponding to pending tasks that require user's attention. In this manner, the user has a global view of the activities in which he or she is involved, including all relevant information needed for the treatment of the pending tasks.

Each *role* (played by some users) is associated with a grammar that describes the dynamic evolution of a case. A production of the grammar is given by a left-hand side, indicating a non-terminal to expand, and a right-hand side, describing how to expand this non-terminal. We interpret a production as a way to decompose a task, the symbol on the left-hand side, into sub-tasks associated with the symbols on the right-hand side. The initial tasks are symbols that appear in some left-hand side (they are *defined*) but do not appear on right-hand side of rules (they are not *used*). They correspond to the *services* that are *provided* by the role. Conversely a symbol that is used but not defined (i.e it appears on some right-hand side but on no left-hand side) is interpreted as a call to an external service. It should appear as a service provided by another role. Symbols that are both used and defined are internal tasks and their names are bound to the role.

$$
\begin{array}{ll}
p_1: & A \to \varepsilon \\
p_2: & A \to BC \\
p_3: & B \to \varepsilon \\
p_4: & B \to D
\end{array}
$$

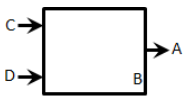


Figure 1. *A grammar for a role that provides a service A and uses the external services C and D. B is an internal task, bound to the role, and whose name can henceforth be changed.*

In order to solve a task A, that appears as a pending task in his workspace, the user may choose to apply production p_1 (which corresponds to a certain action or activity) and this decision ends the performance of task A (since the right-hand side is empty). Alternatively production p_2 may be chosen. In that case, two new (residual) tasks of respective sort B and C are created and A will terminate as soon as B and C have terminated.

The GAG model also attach (inherited and synthesized) information to a task as well as a guard (condition bearing on the inherited information) that specifies when the production is enabled. In this paper we restrict our attention to the dynamic evolution of tasks (the grammar) and forget about extra information and guards.

Our purpose is to define some abstraction of the grammar, called the *interface of the role*, whose aim is to specify what services are provided, which external services are required to carry them out and an over-approximation of the dependencies between required and provided services (the potential dependencies). In particular the interface disregards internal tasks. As a first attempt one considers that the provided service A potentially relies on external service B if a derivation $A \rightarrow^* u$ exists where word u contains an occurrence of B.

The interface of the role given in Figure 2 is relation $R = \{(C, A), (D, A)\}$. It is an over-approximation of the dependencies since it may happen that A uses none of the services C and D (using derivation $A \rightarrow^* \varepsilon$) or only C (using derivation $A \rightarrow^* C$). But an external user invoking service A does not know how the service will be carry out and therefore he must assume the availability of all external services that may potentially be used.

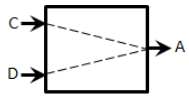


Figure 2. *An interface*

We assume that the grammars are *non-recursive* in the sense that no symbol can derive from itself. Namely we exclude the situation where a derivation $X \rightarrow^* u$ exists in which u is a word that contains an occurrence of X. This condition is very generally verified in the examples we have encountered in practice, for instance when modeling epidemiological surveillance [7]. [1]

Definition 2.1. *Let Ω denote a fixed set of services. An* interface *$({}^\bullet R, R, R^\bullet)$ consists of a finite binary relation $R \subseteq \Omega \times \Omega$ and disjoint subsets ${}^\bullet R$ and $R^\bullet$ of Ω, such that ${}^\bullet R = R^{-1}(\Omega) = \{A \in \Omega \mid \exists B \in \Omega\ (A, B) \in R\}$ and $R^\bullet \supseteq R(\Omega) = \{B \in \Omega \mid \exists A \in \Omega\ (A, B) \in R\}$. The set $R^\bullet$ stands for the services* provided *(or* defined*) by the interface and ${}^\bullet R$ for the* required *(or* used*) services. The relation $(A, B) \in R$ indicates that service B potentially depends upon service A. Thus $A \in R^\bullet \setminus R(\Omega)$ is a service provided by the interface that requires no external services. An interface is* closed *(or* autonomous*) if relation R (and thus also ${}^\bullet R$) is empty. Thus a closed interface is given by the set of services that it (autonomously) provides.*

Note that since ${}^\bullet R$ is the domain of relation R, the set of required services may be left implicit. The same is not true for the set of provided services since it can strictly encompass the codomain of the relation. Still, we shall by abuse of notation use the same symbol to denote an interface and its underlying relation. We extend the following notations from binary relations to interfaces:

1) The empty interface that renders no service at all is $\emptyset = (\emptyset, \emptyset, \emptyset)$.

1. However, it can sometimes be useful to model situations where a task A derives into an arbitrary number of tasks B. Such a situation can be presented by the recursive grammar with rules $A \rightarrow B\ A$ and $A \rightarrow \varepsilon$ which may equivalently be given by the (generalized) production: $A \rightarrow B^*$. Hence, one may be tempted to use non-recursive but generalized grammars (whose right-hand sides are given by regular expressions). However, as we are interested only in the dependencies between services, one can w.l.o.g. replace any regular expression on a right-hand side by the sequence of symbols (without repetition) that occur in it and therefore obtaining an ordinary non-recursive grammar with the same interface.

2) $R_1; R_2 = \{(A,C) \in \Omega^2 \mid \exists B \in \Omega\ (A,B) \in R_1 \wedge (B,C) \in R_2\}$ is the sequential composition with ${}^\bullet(R_1;R_2) = {}^\bullet R_1$ and $(R_1;R_2)^\bullet = R_2^\bullet$.

3) the restriction $R \restriction O$ of interface R to $O \subseteq \Omega$ is given by $R|O = \{(A,B) \in R \mid B \in O\}$ with $(R|O)^\bullet = O \cap R^\bullet$ and ${}^\bullet(R|O) = R^{-1}(O \cap R^\bullet)$.

3. The Composition of Interfaces

The union of interfaces is an interface if none of the services defined by an interface is used by another one. In the general case $R = (\cup_i {}^\bullet R_i, \cup_i R_i, \cup_i R_i^\bullet)$ satisfies the conditions in Definition 2.1 but ${}^\bullet R \cap R^\bullet = \emptyset$. If relation R^* is acyclic we say that it is a *quasi-interface* since it induces an interface given by the following definition.

Definition 3.1. *If $R = ({}^\bullet R, R, R^\bullet)$ is a quasi-interface, i.e. satisfies ${}^\bullet R = R^{-1}(\Omega)$, $R^\bullet \supseteq R(\Omega)$, and its transitive closure R^* is acyclic, then we let $\langle R\rangle = R^* \cap (I \times O)$, where $I = {}^\bullet R \setminus R^\bullet$ and $O = R^\bullet$. It is an interface with ${}^\bullet\langle R\rangle = I$ and $\langle R\rangle^\bullet = O$.*

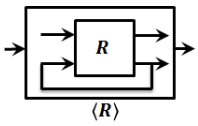


Figure 3. *Interface induced by a quasi-interface*

For instance if $R_1 = (\emptyset, \emptyset, \{A\})$ is the autonomous interface that provide service A and $R_2 = (\{A\}, \{(A,B)\}, \{B\})$ uses A to define another service B, then they jointly provide an autonomous interface $\langle R_1 \cup R_2\rangle = (\emptyset, \emptyset, \{A,B\})$ that provides services A and B. Note that the information that B requires A is lost: the meaningful information is that the interface exports A and B and has no imports. If we assume that interface R_1 rather produces service B from A, namely $R_1 = (\{B\}, \{(B,A)\}, \{A\})$, then the computation of the composition would also give $\langle R_1 \cup R_2\rangle = (\emptyset, \emptyset, \{A,B\})$ even though these two interfaces when combined together cannot render any service. This is the rationale for assuming that a quasi-interface must be acyclic.

Definition 3.2. *Two interfaces R_1 and R_2 are said to be composable if $(R_1 \cup R_2)^*$ is acyclic and $R_1^\bullet \cap R_2^\bullet = \emptyset$. Then we let $R_1 \bowtie R_2 = \langle R_1 \cup R_2\rangle$ denote their composition.*

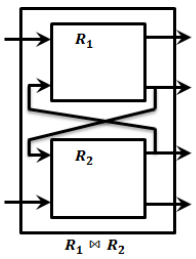


Figure 4. *The composition of two interfaces.*

Note that $(R_1 \bowtie R_2)^\bullet = R_1^\bullet \cup R_2^\bullet$.
Moreover, since ${}^\bullet R_i \cap R_i^\bullet = \emptyset$ for $i = 1, 2$ one gets

$$ {}^\bullet(R_1 \bowtie R_2) = ({}^\bullet R_1 \setminus R_2^\bullet) \cup ({}^\bullet R_2 \setminus R_1^\bullet) $$

It follows also directly from the definition that the composition of interfaces is commutative and has the empty interface as neutral element. Note that we may have ${}^\bullet R_1 \cap {}^\bullet R_2 \neq \emptyset$, thus both interfaces may require some common external services. The following example shows that the composition is not associative if we do not require that composable interfaces have disjoints outputs.

Exemple 3.3. *Let R_1, R_2, and R_3 the three interfaces given in Figure 5. If $R_1 \bowtie (R_2 \bowtie R_3) = (R_1 \bowtie R_2) \bowtie R_3$ we would expect this interface to be given by $R = \langle R_1 \cup R_2 \cup R_3\rangle$ hence $R = \{(A,D),(C,D),(A,E),(B,E),(C,E)\}$. Note that service D may be produced*

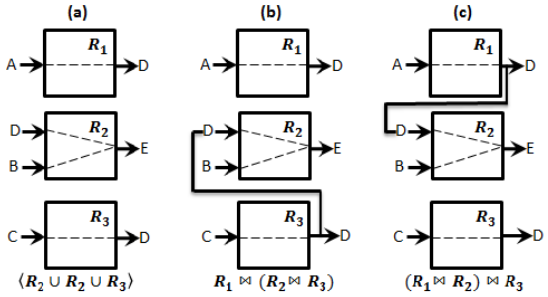


Figure 5. *A counter-example showing that associativity of composition does not hold if interfaces shared some provided services.*

by either R_1 *or* R_3 *so that we find both* (A, D) *and* (C, D) *as dependencies in* R*. It follows that E potentially depends on both A, B, and C. However if we compute* $R_1 \bowtie (R_2 \bowtie R_3)$ *we get* $R_r = \{(A, D), (C, D), (B, E), (C, E)\}$ *because in* R_2 *the required service D is no longer an input in* $R_2 \bowtie R_3$*. Symmetrically* $R_l = (R_1 \bowtie R_2) \bowtie R_3 = \{(A, D), (C, D), (A, E), (C, E)\}$.

Proposition 3.4. *The composition of interfaces is associative. More precisely, if* $R_1 \cdots R_n$ *are pairwise composable interfaces, then* $\bowtie_{i=1}^{n} R_i = \langle R_1 \cup \cdots \cup R_n \rangle$.

The following two cases of composition are noteworthy:

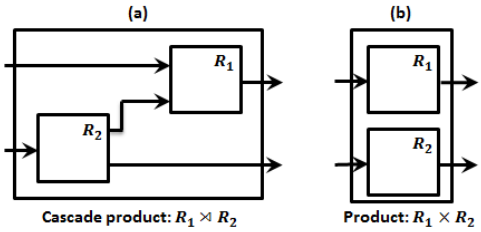


Figure 6. *Cascade product and (direct) product*

Cascade product If $R_1^\bullet \cap {}^\bullet R_2 = \emptyset$ we denote $R_1 \rtimes R_2$ their composition (or $R_2 \ltimes R_1$ since this operation as a particular case of $\bowtie$ is still commutative). Then ${}^\bullet(R_1 \rtimes R_2) = ({}^\bullet R_1 \setminus R_2^\bullet) \cup {}^\bullet R_2$, and $(R_1 \rtimes R_2)^\bullet = R_1^\bullet \cup R_2^\bullet$.

(Direct) product If both $R_1^\bullet \cap {}^\bullet R_2 = \emptyset$ and $R_2^\bullet \cap {}^\bullet R_1 = \emptyset$ hold we say that the composition is the product of R_1 and R_2, denoted as $R_1 \times R_2$. Note that $R_1 \times R_2 = R_1 \cup R_1$ and thus ${}^\bullet(R_1 \times R_2) = {}^\bullet R_1 \cup {}^\bullet R_2$ and $(R_1 \times R_2)^\bullet = R_1^\bullet \cup R_2^\bullet$.

Definition 3.5. R_1 *is a* component *of* R*, in notation* $R_1 \sqsubseteq R$*, if there exists an interface* R_2 *such that* $R = R_1 \bowtie R_2$*.* R_1 *is a* strict component *of* R*, in notation* $R_1 \sqsubset R$*, if there exists an interface* R_2 *such that* $R = R_1 \times R_2$.

4. Implementation Order

An environment for an interface is any component that provides all the services required by the interface and uses for that purpose only services that are provided by it.

Definition 4.1. *An interface E is an* admissible environment *for an interface R if the two interfaces are composable and the resulting composition is a closed interface, namely ${}^\bullet(R \bowtie E) = \emptyset$. We let $\mathbf{Env}(R)$ denote the set of admisible environments of interface R.*

Definition 4.2. *An interface R_2 is an* implementation *of interface R_1, in notation $R_1 \leq R_2$, when $R_2^\bullet = R_1^\bullet$ and $R_2 \subseteq R_1$.*

Thus R_2 is an implementation of R_1 if it renders the same services as R_1 using only services already used by R_1 and with less dependencies.[2] The following proposition shows that R_2 is an implementation of interface R_1 if and only if it can be substituted to R_1 in any admissible environment for R_1.

Proposition 4.3. *$R_1 \leq R_2$ if and only if $\mathbf{Env}(R_1) \subseteq \mathbf{Env}(R_2)$.*

5. Residual Specification

Proposition 5.1. *If $R_1 \sqsubseteq R$ then $R = R_1 \rtimes (R \swarrow R_1)$ where $R \swarrow R_1$, called the* strict residual *of R by R_1, is given as the restriction of R to $R^\bullet \setminus R_1^\bullet$. If $R = R_1 \rtimes R_2$ then $R{\restriction}R_2^\bullet = R \swarrow R_1 = R_2$ and $R = (R \swarrow R_2) \times (R \swarrow R_1)$.*

Corollary 5.2. *If $R^\bullet = O_1 \cup O_2$ with $O_1 \cap O_2 = \emptyset$ then $R = (R{\restriction}O_1) \times (R{\restriction}O_2)$ and $R{\restriction}O_i = R \swarrow (R{\restriction}O_j)$ for $\{i, j\} = \{1, 2\}$ and the following conditions are equivalent:*

1) R_1 is a strict component of R: $\exists R_2 \cdot R = R_1 \times R_2$,

2) R_1 is a left component in a cascade decomposition of R: $\exists R' \cdot R = R' \rtimes R_2$,

3) R_1 is a restriction of R: $R_1 = R{\restriction}(R_1^\bullet)$, and

4) R_1 is a strict residual of R: $\exists R_2 \cdot R_1 = R \swarrow R_2$.

Proposition 5.3. *If R_1 is a component of R and R' is an interface then*

$$R \swarrow R_1 \leq R' \iff R \leq R_1 \rtimes R'.$$

By Corollary 5.2 the above proposition implies that an implementation of a strict residual $R \swarrow R_1$ is a strict component of R and therefore it cannot capture all the components of an implementation of R, i.e. all interfaces R' such that $R \leq R_1 \bowtie R'$. For that purpose we need to add in the residual all the dependencies between the respective ouputs of the component and of the residual that do not contradict dependencies in R:

Definition 5.4. *If $R_1 \sqsubseteq R$ the* residual *R/R_1 of R by R_1 is given by $(R/R_1)^\bullet = R^\bullet \setminus R_1^\bullet$ and $R/R_1 = R \swarrow R_1 \cup R \nearrow R_1$ where*

$$R \nearrow R_1 = \left\{ (A, B) \in R_1^\bullet \times (R^\bullet \setminus R_1^\bullet) \;\middle|\; R^{-1}(\{A\}) \subseteq R^{-1}(\{B\}) \right\}.$$

2. In pratice an interface used as an implementation may define additional services: R_2 is a *weak implementation* of interface R_1, in notation $R_1 \leq_w R_2$, if $R_2^\bullet \supseteq R_1^\bullet$ and $R_1 \leq R_2{\restriction}(R_1^\bullet)$. However the additional services provided by R_2 should be hided so that they cannot conflict with services of any environment compatible with R_1.

Proposition 5.5. *If R_1 is a component of R and R' is an interface then*

$$R/R_1 \leq R' \iff R \leq R_1 \bowtie R'.$$

Hence the residual R/R_1 characterizes those interfaces that, when composed with R_1, produce an implementation of R.

6. Conclusion

This work is a first attempt to develop an interface theory for distributed collaborative systems in the context of service-oriented programming. We intend to use it to define and structure the activities of crowdsourcing system operators. The residual operation can be used to identify the skills to be sought in the context of existing services in order to achieve a desired overall behaviour. Such a system can be implemented by Guarded Attribute Grammars and interfaces can be used to type applications. However, the notion of interface presented in this paper is still a somewhat rudimentary abstraction of the roles described by a GAG specification. In particular, we would like to be able to take into account the non-determinism resulting from the choices of users in their ways to solve a given task. This could be done by replacing the relation $R \subseteq \Omega \times \Omega$ by a map $R : \Omega \to \wp(\wp(\Omega))$ that associates each service $A \in \Omega$ with a finite number of alternative ways to carry it out, and each of these with the set of external services that it requires. Then we would have $R^\bullet = \{A \in \Omega \mid R(\Omega) \neq \emptyset\}$ and ${}^\bullet R = \cup\{R(A) \mid A \in \Omega\}$. The composition, implementation (pre-)order, and residual would have to be adapted in this context. It would then be possible to define some new operations like the corestriction $R\lfloor I$ of an interface R to a set of services $I \subseteq \Omega$, where $(R\lfloor I)(A) = \{X \cap I \mid X \in R(A)\}$ states how a role can be used when the set of services actually provided by the environment is known (to be I). Now it might be possible that we have only a partial knowledge of the set of available services in the form of a believe function [10] or a possibilistic distribution [11]. Then we should enrich an interface with qualitative information and viewed it as a believe function transformer that updates the knowledge on the services rendered by the environment when a new role enters the system. Finally, one can also enrich the interface with information on time execution.

7. References

[1] MARTÍN ABADI, LESLI LAMPORT, "Composing Specifications", *ACM Transactions on Programming Languages and Systems*, vol. 15, 1993:73–132.

[2] MARTÍN ABADI, GORDON D. PLOTKIN, "A logical view of composition", *Theoretical Computer Science*, vol. 114, 1993:3–30.

[3] LUCA DE ALFARO, THOMAS A. HENZINGER, "Interface Automata", *Foundation of Software Engineering (ESEC/FES-9)*, 2001: 109–120.

[4] ERIC BADOUEL, LOÏC HÉLOUËT , GEORGES-EDOUARD KOUAMOU, CHRISTOPHE MORVAN , ROBERT FONDZE JR. NSAIBIRNI, "Active Workspaces; Distributed Collaborative Systems based on Guarded Attribute Grammars", *ACM SIGAPP Applied Computing Review*, vol. 15, num. 3, 2015:6–34.

[5] DAVID HAREL, AMIR PNUELI, "On the development of reactive systems", *Logics Models of Concurrent Systems*, NATO ASI Series, vol. F13, Springer Berlin, 1984: 477–498.

[6] PHILIP M. MERLIN, GREGOR VON BOCHMANN, "On the construction of submodule specifications and communication protocols", *ACM Transactions on Programming Languages and Systems*, vol. 5, 1983:1–25.

[7] ROBERT FONDZE JR NSAIBIRNI, ERIC BADOUEL, GAËTAN TEXIER , GEORGES-EDOUARD KOUAMOU, "Active Workspace: A Dynamic Collaborative Business Process Model for Disease Surveillance Systems", *Health Informatics and Medical Systems*, Las Vegas, USA, 2016: 58–64.

[8] VAUGHAN R. PRATT, "Origins of the Calculus of Binary Relations", *IEEE Logic in Computer Science (LICS'92)*, Santa Cruz, California, USA, 1992: 248–254.

[9] JEAN-BAPTISTE RACLET, "Residual for Component Specifications", *Electronic Notes in Theoretical Computer Science*, vol. 215, 2008:93–110.

[10] GLENN SHAFER, *A mathematical theory of evidence*, Princeton University Press, 1976.

[11] LOFTI ZADEH, "Fuzzy Sets as the Basis for a Theory of Possibility", *Fuzzy Sets and Systems*, vol. 1, 1978:3–28.

Appendix: Proofs of Results

The theory of interfaces that we consider is mainly a calculus of relations [8] even though we put stress on the (concurrent) composition rather than on the usual (sequential) composition of relations. As a result we have introduced a residuation operation for the composition in place of the left and right residuals for sequential composition. Similarly our implementation order is mostly given by the set-theoretical inclusion of relations. In order to ease computation we identify a set $X \subseteq \Omega$ with the interface $(X = (X, \{(A,A) \in \Omega^2 \mid A \in X\}, X)$. By doing so, one can for instance express the condition $B \in Y \wedge (\exists C \in X\ (A,C) \in R \wedge (C,B) \in Y)$ for $R; S \subseteq \Omega \times \Omega$ and $X, Y \subseteq \Omega$ as $(A,B) \in R; X; S; Y$. One can also express the cascade product as a sequential composition:

Remark 7.1. *$R_1 \rtimes R_2 = (I_1 \times R_2)\,;(R_1 \times O_2)$ where $I_1 = {}^\bullet R_1 \setminus R_2^\bullet$ and $O_2 = R_2^\bullet \setminus {}^\bullet R_1$.*

Note moreover that with this convention one has $R\restriction X = R; X$ and $X \cap Y = X; Y$ for R an interface and X and Y subsets of Ω.

1. Associativity of the Composition of Interfaces

Remark 7.2. *$\langle R\rangle = \{(A,B) \in R^* \mid \neg\,(\exists C \in \Omega.\ (C,A) \in R)\}$. Hence any $(A,B) \in \langle R\rangle$ is associated with a path in the graph of R that leads to $B \in R^\bullet$ and cannot be extended on the left. Note that such a path is of the form $A = A_0 \to A_1 \to \cdots \to A_n = B$, with $A \in {}^\bullet R \setminus R^\bullet$ and $\forall 1 \le i \le n\ (A_{i-1}, A_i) \in R$. Note that $\forall 1 \le i \le n\ A_i \in R^\bullet$, i.e. all elements in this path but the first one, namely A, belongs to $R^\bullet$.*

Proposition. *The composition of interfaces is associative. More precisely, if $R_1 \cdots R_n$ are pairwise composable interfaces, then $\bowtie_{i=1}^n R_i = \langle R_1 \cup \cdots \cup R_n\rangle$.*

Proof. Using the commutativity of composition, the proposition follows by induction on n as soon as it has been verified that $(R_1 \bowtie R_2) \bowtie R_3 = \langle R_1 \cup R_2 \cup R_3\rangle$ for pairwise composable interfaces R_1, R_2 and R_3. Hence we have to show $\langle\langle R_1 \cup R_2\rangle \cup R_3\rangle = \langle R_1 \cup R_2 \cup R_3\rangle$ or, more generally, that $\langle\langle R\rangle \cup R'\rangle = \langle R \cup R'\rangle$ where $R \subseteq \Omega \times \Omega$ is a finite binary relation with possibly ${}^\bullet R \cap R^\bullet \neq \emptyset$, and R' is an interface such that $(R \cup R')^*$ acyclic, and $R^\bullet \cap (R')^\bullet = \emptyset$. First, note that $\langle R\rangle^\bullet = R^\bullet$ and $(R \cup R')^\bullet = R^\bullet \cup (R')^\bullet$ and

thus $\langle\langle R\rangle \cup R'\rangle^\bullet = \langle R \cup R'\rangle^\bullet$. By condition $R^\bullet \cap (R')^\bullet = \emptyset$ we deduce $R \cap R' = \emptyset$. More precisely a transition $(A, B) \in R \cap R'$ belongs (exclusively) either to R or to R' depending respectively on $B \in R^\bullet$ or $B \in (R')^\bullet$. According to Remark 7.2, let $\pi = A_0 \to A_1 \to \cdots \to A_n$ be a path in $R \cup R'$ (i.e. $\forall 1 \leq i \leq n\ (A_{i-1}, A_i) \in R \cup R'$ and $A_0 \in {}^\bullet(R \cup R') \setminus (R \cup R')^\bullet$) witnessing that $(A_0, A_n) \in \langle R \cup R'\rangle$. Let $\pi' = A_i \to \cdots \to A_j$ be a maximal sub-path of π made of R transitions only (i.e., $\forall i \leq k \leq j\ A_k \in R^\bullet$). Then either $A_i = A_0$ or $(A_{i-1}, A_i) \in R'$. In both cases $A_i \in {}^\bullet R \setminus R^\bullet$ and thus π' is a path witnessing that $(A_i, A_j) \in \langle R\rangle$ from which it follows that π is a path witnessing that $(A_0, A_n) \in \langle\langle R\rangle \cup R'\rangle$, showing $\langle R \cup R'\rangle \subseteq \langle\langle R\rangle \cup R'\rangle$ and hence $\langle\langle R\rangle \cup R'\rangle = \langle R \cup R'\rangle$ since the converse inclusion is immediate. □

2. Implementation Order

Proposition. $R_1 \leq R_2$ *if and only if* $\mathbf{Env}(R_1) \subseteq \mathbf{Env}(R_2)$.

Proof. We first show that the condition is necessary. For that purpose let us assume $R_1 \leq R_2$ (which means that $R_2^\bullet = R_1^\bullet$ and $R_2 \subseteq R_1$) and prove that any admissible environment E for R_1 is an admissible environment for R_2. Since E is composable with R_1 we get $R_1^\bullet \cap E^\bullet = \emptyset$ and $(E \cup R_1)^*$ is acyclic. Then we also have $R_2^\bullet \cap E^\bullet = \emptyset$ and $(E \cup R_2)^*$ is acyclic since $R_2^\bullet = R_1^\bullet$ and $R_2 \subseteq R_1$. Hence E is composable with R_2. Moreover, for the same reasons, ${}^\bullet(E \bowtie R_2) = ({}^\bullet E \setminus R_2^\bullet) \cup ({}^\bullet R_2 \setminus E^\bullet) \subseteq ({}^\bullet E \setminus R_1^\bullet) \cup ({}^\bullet R_1 \setminus E^\bullet) = {}^\bullet(E \bowtie R_2) = \emptyset$. Henceforth $E \in \mathbf{Env}(R_2)$. We show that the condition is sufficient by contradiction. Since $R_1 \leq R_2$ implies $R_2^\bullet = R_1^\bullet$ one has to construct $\mathcal{E} \in \mathbf{Env}(R_1) \setminus \mathbf{Env}(R_2)$ under the assumption that $R_1 \not\supseteq R_2$. Let $(A, B) \in R_2 \setminus R_1$ then the interface we are looking for is E such that ${}^\bullet E = \{B\}$, $E^\bullet = {}^\bullet R_2$, and $E = \{(B, A)\}$. Indeed, E is composable with R_1 but not with R_2 because of the cycle $B \to A \to B$ in $(R_1 \cup \{(B, A)\})^*$. Moreover the composition of E with R_1 gives a closed interface. □

3. Residual specification

Proposition. *If* $R_1 \sqsubseteq R$ *then* $R = R_1 \rtimes (R \swarrow R_1)$ *where* $R \swarrow R_1$*, called the* strict residual *of* R *by* R_1*, is given as the restriction of* R *to* $R^\bullet \setminus R_1^\bullet$*. If* $R = R_1 \rtimes R_2$ *then* $R \restriction R_2^\bullet = R \swarrow R_1 = R_2$ *and* $R = (R \swarrow R_2) \times (R \swarrow R_1)$.

Proof. One has to show that if R_1 and R_2 are two composable relation with $R = R_1 \bowtie R_2$ then $R = R_1 \rtimes R \swarrow R_1$ and $R = (R \swarrow R_1) \times (R \swarrow R_1)$ where $R \swarrow R_i = R \restriction R_j^\bullet$ for $\{i, j\} = \{1, 2\}$. By remark 7.2 $R_1 \bowtie R_2$ is the (unique)[3] solution of the system of equations

$$\begin{array}{rrl} & R_1 \bowtie R_2 = & (A \cup I_1); R_1 \cup (B \cup I_2); R_2 \\ where & I_1 = & {}^\bullet R_1 \setminus R_2^\bullet \\ & I_2 = & {}^\bullet R_2 \setminus R_1^\bullet \\ & O_1 = & R_1^\bullet \cap {}^\bullet R_2 \\ & O_2 = & R_2^\bullet \cap {}^\bullet R_1 \\ & A = & (B \cup I_2); R_2; O_2 \\ & B = & A \cup I_1); R_1; O_1 \end{array}$$

3. Unicity comes from the fact that one considers only finite paths due to acyclicity.

Then it is immediate (see Figure 7) that $R_1 \rtimes (R_1 \bowtie R_2) \lceil Out(R_2)$ is solution of the same system of equations and thus the two relations coincide. The same system of equations is associated with $(R_1 \swarrow R_2) \times (R_1 \swarrow R_1)$ as shown in Figure 8.

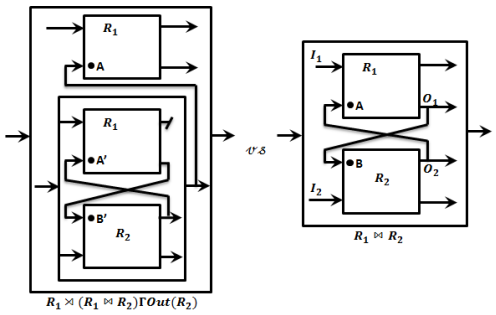


Figure 7. $R = R_1 \rtimes (R_1 \swarrow R_1)$ *when* $R = R_1 \bowtie R_2$

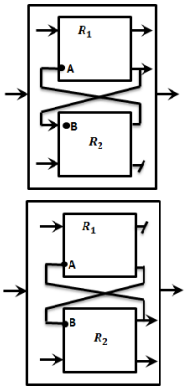


Figure 8. $(R_1 \swarrow R_2) \times (R_1 \swarrow R_1)$

It remains to show that if $R = R_1 \rtimes R_2$ then $R_1 \swarrow R_1 = R \restriction R_2^\bullet$ coincides with R_2, and indeed $R \restriction R_2^\bullet = ((^\bullet R_1 \setminus R_2^\bullet) \cup R_2)\,; (R_1 \cup R_2^\bullet) \restriction R_2^\bullet = ((^\bullet R_1 \setminus R_2^\bullet) \cup R_2) \restriction R_2^\bullet = R_2$ by Remark 7.1 and because $R_1^\bullet \cap R_2^\bullet = \emptyset$. □

Lemma 7.3. *$R_1 \leq R_2$ implies $R \bowtie R_1 \leq R \bowtie R_2$ whenever R_1 and R_2 are both components of R.*

Proof. By Remark 7.2 $(A, B) \in R \bowtie R_i$ if and only if there exists a finite sequence $A_0, \ldots, A_n$ such that $A = A_0 \in {}^\bullet R \setminus R_i^\bullet \cup {}^\bullet R_i \setminus R^\bullet$, $(A_{k-1}, A_k) \in R \cup R_i$ for all $1 \leq k \leq n$, and $B = A_n \in R^\bullet \cup R_i^\bullet$. Monotony of $R \bowtie -$ then follows from the fact that $R_1^\bullet = R_2^\bullet$. □

Lemma 7.4. *$R_1 \leq R_2$ implies $R_1 \swarrow R \leq R_2 \swarrow R$ whenever R is a component of both R_1 and R_2.*

Proof. $R_1 \leq R_2$ means that $R_1^\bullet = R_2^\bullet$ and $R_2 \subseteq R_1$. Then $R_1 \swarrow R = R_1 \restriction (R_1^\bullet \setminus R^\bullet) \leq R_2 \restriction (R_2^\bullet \setminus R^\bullet)$ because $R_1^\bullet \setminus R^\bullet = R_2^\bullet \setminus R^\bullet$ and $R_2 \subseteq R_1$. □

Proposition. *If R_1 is a component of R and R' is an interface then*

$$R_1 \swarrow R_1 \leq R' \iff R \leq R_1 \rtimes R'.$$

Proof. By Proposition 5.1 and Lemma 7.3 we get $R_1 \swarrow R_1 \leq R' \implies R = R_1 \rtimes (R_1 \swarrow R_1) \leq R \rtimes R'$. The converse direction follows by Lemma 7.4 and Proposition 5.1: $R \leq R_1 \rtimes R' \implies R_1 \swarrow R_1 \leq (R \rtimes R') \swarrow R_1 = R'$. □

Lemma 7.5. *If R_1 is a component of R then $R_1 \bowtie (R / R_1) = R$*

Proof. Since $(R/R_1)^\bullet = R^\bullet \setminus R_1^\bullet = (R_1 \swarrow R_1)^\bullet$ and $R/R_1 \supseteq R_1 \swarrow R_1$ one has $R/R_1 \leq R_1 \swarrow R_1$ and by Lemma 7.3 $R_1 \bowtie (R/R_1) \leq R_1 \bowtie (R_1 \swarrow R_1) = R_1 \rtimes (R_1 \swarrow R_1) = R$. We are left to prove that $R_1 \bowtie (R/R_1) \subseteq R$. Let $(A, B) \in R_1 \bowtie (R/R_1)$ then by Remark 7.2 there exists a sequence $A_0, \ldots, A_n$ such that $A = A_0 \in {}^\bullet (R_1 \bowtie (R/R_1))$,

$B = A_n \in (R_1 \bowtie (R/R_1))^\bullet = R^\bullet$, and $(A_{i-1}, A_i) \in R_1 \cup (R/R_1)$ for all $1 \leq i \leq n$. One has ${}^\bullet(R_1 \bowtie (R/R_1)) = {}^\bullet R_1 \setminus (R^\bullet \setminus R_1^\bullet) \cup {}^\bullet(R/R_1) \setminus R_1^\bullet$. Thus $A \in {}^\bullet R$ because ${}^\bullet R_1$ and ${}^\bullet(R/R_1)$ are subsets of ${}^\bullet R$. There are three possibilities for each transition (A_{i-1}, A_i): *(i)* $(A_{i-1}, A_i) \in R_1$ if $A_i \in R_1^\bullet$, *(ii)* $(A_{i-1}, A_i) \in R_{\swarrow} R_1$ if $A_i \in R^\bullet \setminus R_1^\bullet$ and $A_{i-1} \in {}^\bullet R \setminus R_1^\bullet$, or *(iii)* $(A_{i-1}, A_i) \in R \nearrow R_1$ if $A_i \in R^\bullet \setminus R_1^\bullet$ and $A_{i-1} \in R_1^\bullet$, Note that if the sequence contains no transition of the latter category then it witnesses that $(A, B) \in R$ due to the fact that $R_1 \bowtie (R_{\swarrow} R_1) = R_1 \rtimes (R_{\swarrow} R_1) = R$. We're going to gradually eliminate all transitions of type *(iii)*. For doing so let us consider the leftmost transition of this latter category if it exists. Thus i is the smallest index such that $(A_{i-1}, A_i) \in R \nearrow R_1$. Since $R_1^\bullet$ is a subset of $R^\bullet$ and thus is disjoint of ${}^\bullet R$ we deduce that $A_{i-1} \neq A$ and thus $i - 1 \geq 1$. Now the sequence $\sigma : A = A_0 \to \cdots \to A_{i-1}$, which contains only transitions of types *(i)* or *(ii)*, witnesses that $A \in R^{-1}(\{A_{i-1}\})$. Since $(A_{i-1}, A_i) \in R \nearrow R_1$ we deduce that $A \in R^{-1}(\{A_i\})$. Thus by replacing sequence σ by transition (A, A_i) we get a sequence with one less transition in $R \nearrow R_1$ and thus we end up with a sequence with no transition in $R \nearrow R_1$ witnessing that $(A, B) \in R$. □

Lemma 7.6. *If R_1 and R_2 are composable then $(R_1 \bowtie R_2)/R_1 \leq R_2$.*

Proof. Let R_1 and R_2 be composable interfaces, in particular $R_1^\bullet \cap R_2^\bullet = \emptyset$, and $R = R_1 \bowtie R_2$. Then $(R/R_1)^\bullet = (R_1^\bullet \cup R_2^\bullet) \setminus R_1^\bullet = R_2^\bullet$. $(A, B) \in R_2 \setminus (R_{\swarrow} R_1) = R_2 \setminus (R \restriction R_2)$ if and only if $(A, B) \in R_2$ (hence $B \in R_2^\bullet$, and $A \in {}^\bullet R_2 \cap R_1^\bullet$. Then necessarily $R^{-1}(\{A\}) \subseteq R^{-1}(\{B\})$ and therefore $(A, B) \in R \nearrow R_1$. It follows that $R/R_1 = R_{\swarrow} R_1 \cup R \nearrow R_1 \supseteq R_2$ and thus $R/R_1 \leq R_2$. □

Lemma 7.7. *$R_1 \leq R_2$ implies $R_1/R \leq R_2/R$ whenever R is a component of both R_1 and R_2.*

Proof. Recall that $R_i \nearrow R = \left\{(A, B) \in R^\bullet \times (R_i^\bullet \setminus R^\bullet) \;\middle|\; R^{-1}(\{A\}) \subseteq R^{-1}(\{B\})\right\}$ and $R_i/R = R_{i\swarrow} R \cup R_i \nearrow R$. $R_1 \leq R_2$ means that $R_1^\bullet = R_2^\bullet$ and $R_2 \subseteq R_1$ from which it follows that $R^\bullet \times (R_1^\bullet \setminus R^\bullet) = R^\bullet \times (R_2^\bullet \setminus R^\bullet)$ and thus $R_2 \nearrow R \subseteq R_1 \nearrow R$. The result then follows from Lemma 7.4 and $(R_1/R)^\bullet = R_1^\bullet \setminus R^\bullet = R_2^\bullet \setminus R^\bullet = (R_2/R)^\bullet$. □

Proposition. *If R_1 is a component of R and R' is an interface then*

$$R/R_1 \leq R' \iff R \leq R_1 \bowtie R'.$$

Proof. By Lemma 7.5 and Lemma 7.3 we get $R/R_1 \leq R' \implies R = R_1 \bowtie (R/R_1) \leq R \bowtie R'$. The converse direction follows by Lemma 7.7 and Lemma 7.6: $R \leq R_1 \bowtie R' \implies R/R_1 \leq (R \bowtie R')/R_1 \leq R'$. □

ε-TPN: definition of a Time Petri Net formalism simulating the behaviour of the timed grafcets

Médésu Sogbohossou — Antoine Vianou

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

ABSTRACT. To allow a formal verification of timed GRAFCET models, many authors proposed to translate them into formal and well-reputed languages such as timed automata or Time Petri nets (TPN). Thus, the works presented in [Sogbohossou, Vianou, Formal modeling of grafcets with Time Petri nets, IEEE Transactions on Control Systems Technology, 23(5)(2015)] concern the TPN formalism: the resulting TPN of the translation, called here ε-TPN, integrates some infinitesimal delays (ε) to simulate the synchronous semantics of the grafcet. The first goal of this paper is to specify a formal operational semantics for an ε-TPN to amend the previous one: especially, priority is introduced here between two defined categories of the ε-TPN transitions, in order to respect strictly the synchronous hypothesis. The second goal is to provide how to build the finite state space abstraction resulting from the new definitions.

RÉSUMÉ. Afin de permettre la vérification formelle des grafcets temporisés, plusieurs auteurs ont proposé de les traduire dans des langages formels de réputation tels que les automates temporisés et les réseaux de Petri temporels (TPN). Ainsi, les travaux présentés dans [Sogbohossou, Vianou, Formal modeling of grafcets with Time Petri nets, IEEE Transactions on Control Systems Technology, 23(5)(2015)] concernent le formalisme des TPN: le réseau résultant de la traduction, dénommé ici ε-TPN, intègre des délais infinitésimaux (ε) pour simuler la sémantique synchrone du grafcet. Le premier objectif de cet article est de définir la sémantique opérationnelle d'un ε-TPN afin d'améliorer l'ancienne définition: spécifiquement, une priorité est introduite ici entre deux catégories de transitions définies pour ces réseaux, dans l'optique de respecter rigoureusement l'hypothèse synchrone. Le second but est de fournir une méthode de calcul de l'espace d'état fini qui découle des nouvelles définitions.

KEYWORDS : Time Petri Net, timed grafcet, state class, partial order execution, synchronous modelling

MOTS-CLÉS : Réseau de Petri temporel, grafcet temporisé, classe d'état, exécution ordre partiel, modélisation synchrone

1. Introduction

Formal specification of a critical system at the early stage of conception is often needed to achieve their reliability in working, by means of languages allowing simulation or formal verification on the established model of this system [6]. Graphical state-transition modeling formalisms in engineering are appreciated because of their intuitiveness. They are based on the automata theory, ensuring an unambiguous description of the behaviours of a system. Petri nets (PN) are one of these formalisms, used to model in a compact and explicit way the concurrency and the synchronization between the dynamic components of the so-called discrete-event systems [5]. In PNs, firing of transitions (with possibly multiple concurrent firings in the same instant) changes the state and express the dynamics of the modeled system. Time Petri nets (TPN) [1] are one of its extensions, suitable when quantitative time analyses are required for the real-time specifications.

Otherwise, the engineering pratices often promote less formal graphical languages, because of their increased semantic richness (for instance, litteral formulae and hierarchical modeling do not exist in the ordinary PNs) favoring more compact and fluent modeling to the detriment of unambiguous interpretations. These are the cases of formalisms derived from PNs, such as GRAFCET [1] (IEC 60848 standard) [8] and SFC (Sequential Function Chart, IEC 61131-3 standard) [9], used mainly in the world of the manufacturing control. Whereas simultaneous fireable transitions are always done by their total interleaving with PNs, the semantics of these two IEC standards considers only synchronous firings; a consequence is that the notion of transitions in conflict does not exist in GRAFCET and SFC formalisms. GRAFCET is intended for specification purposes (event-driven modeling), contrary to SFC for implementation uses (clock-driven modeling), and is considered in the sequel.

To allow a formal verification of GRAFCET (or SFC) models integrating quantitative time informations, many authors proposed to translate them into formal and well-reputed languages such as timed automata [7] or TPN [12, 11]. The work in [11] is focused on defining some transformation rules which are used to translate the entities composing a timed and not necessarily sound grafcet chart (steps, transitions, literal variables, actions) into connected blocks to obtain the resulting TPN. The method exploits the similarity between TPN and GRAFCET to avoid exponential size of the translation, and implicitly relies on a clear choice about the GRAFCET semantics.

To deal with synchronous firings inherent to GRAFCET formalism, the authors [11] introduced transitions with infinitesimal ε delays, however without redefining formally the resulting extended TPN. The first goal of this paper is to palliate this lack, by specifying a formal semantics for the so-called ε-TPN; the slight differences with the definition in [11] are also presented. Basing on this new definition, the second goal is to provide how to build the state space abstraction of an ε-TPN (which is just sketched in [11]) with the ε delays. Particularly, it is shown how to take advantage from this kind of TPN to cope with the state-space explosion problem, by avoiding useless interleaving of concurrent firings and by abstracting some state classes during the state-space construction.

The three next sections are organized as follows. In Section 2 are given definitions about TPN with ε delays on some transitions: especially, the formal operational semantics of ε-TPN is presented and compared with [11]. In Section 3, the algorithm for generating the finite abstraction (derived from the state class construction for classic TPN [1, 3])

1. Acronym in French: *GRAphe Fonctionnel de Commande Etape Transition*.

taking into account the explosion problem is provided. At last, the conclusion section summarizes the contribution of this paper, and sketches its perspectives.

2. Syntax and semantics of ε-TPN

2.1. The context

The works [11] proposed to translate a timed (and non-hierarchical) grafcet into TPN for formal verification purposes. They give some transformation rules to convert the elements of a grafcet into modules of TPN which are connected, as one goes along a suitable order of translation. An extra module named *phase sequencer* (Fig. 1(a)) is necessary to allow a transient evolution without modification of inputs as external events: it forces alternation between the reaction phase (called *evolution* with grafcets) and an external event production in a stable situation. After adding this first module, the generation of the complete TPN is done by translating sequentially (see figures in Annex B): the steps, the inputs, the timed variables, the outputs, the counter variables, the continuous and conditional actions, the stored actions and the grafcet transitions.

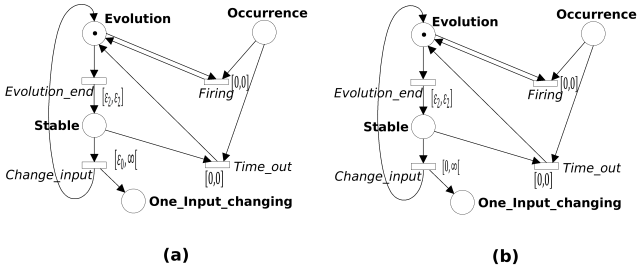


Figure 1. *Phase sequencer: former version (a), new version (b)*

All the transitions in the resulting TPN bear intervals of the form $[\delta, \delta]$ (δ is a delay), except only one: that is called $Change\_input$ in the phase sequencer (Fig. 1), of which firing allows an input event to occur at any time during a stable situation.

An illustration of a grafcet translation into TPN is given in Annex C.

2.2. Syntax

Lets ε_0, an infinitesimal constant (a delay comparable to 0^+). It follows that $\varepsilon_n \stackrel{\text{def}}{=} \varepsilon_0 \times n$ for any $n \in \mathbb{N}^*$, and $\mathcal{E} \stackrel{\text{def}}{=} \{\varepsilon_n \mid n \in \mathbb{N}^*\}$. It is assumed, for any $\varepsilon_n \in \mathcal{E}$ and any $d \in \mathbb{R}^{+*}$, that $0 < \varepsilon_n < d$ and $d + \varepsilon_n \approx d$ hold. By extension, $\mathcal{E}_0 \stackrel{\text{def}}{=} \mathcal{E} \cup \{0\}$.

Définition 1. *A* Time Petri net *(TPN) is a tuple* $(P, T, W, W_I, W_R, ED, LD, M_0)$ *such as:*

1) *the nodes:* P *is the set of* places *and* T *is the set of* transitions *(*$P \cap T = \emptyset$*);*
2) $W : P \times T \cup T \times P \longrightarrow \mathbb{N}$ *defines the* regular arcs *between nodes, and their weights;*
3) $W_R : P \times T \longrightarrow \mathbb{N}$ *defines the* read arcs*;*
4) $W_I : P \times T \longrightarrow \mathbb{N}^* \cup \{\infty\}$ *defines the* inhibitor arcs*;*

5) *the initial marking* $M_0 : P \longrightarrow \mathbb{N}$*;*

6) *the earliest firing delays* $ED : \ T \longrightarrow \mathbb{Q}^+ \cup \mathcal{E}$*;*

7) *the latest firing delays* $LD : \ T \longrightarrow \mathbb{Q}^+ \cup \mathcal{E} \cup \{\infty\}$*;*

8) *the set* T *is a partition of three subsets* $T_{\mathcal{E}_0}$*,* T_T *and* T_∞ *such as:*

- ***a)*** $t \in T_{\mathcal{E}_0}$ *iff* $ED(t) = LD(t) \in \mathcal{E}_0$ *;*
- ***b)*** $t \in T_T$ *iff* $ED(t) = LD(t) \in \mathbb{Q}^{+*}$ *;*
- ***c)*** $t \in T_\infty$ *iff* $ED(t) = 0$ *and* $LD(t) = \infty$*.*

Items 1 to 5 correspond to the classic definition about ordinary Petri nets. Let $p \in P$ and $t \in T$. Graphically, an arc ((p,t) or (t,p)) may be qualified regular, inhibitor or read. By default, the weight is 0 when no regular arc links two nodes (one place and one transition), and is 1 when any arc is represented without its weight. Also, no read arc (resp. no inhibitor arc) directed from a place p to a transition t means the weight $W_R(p,t) = 0$ (resp. $W_I(p,t) = \infty$).

A marking M ($M : P \longrightarrow \mathbb{N}$) enables a transition t iff: $\forall p \in P, (M(p) \geq W(p,t) \wedge M(p) \geq W_R(p,t)) \wedge \nexists p \in P, W_I(p,t) \leq M(p)$. $En(M)$ designates the set of transitions enabled by the marking M. In the next subsection, the notation ${}^\bullet t$ (resp. $t^\bullet$) indicates the multiset [2] of the input (resp. output) places for a given transition t, by the relation W.

The following items (6 to 8) of the definition 1 extend the classic definition about TPN, by integrating the infinitesimal delays and the specific constraints on the static firing interval form of a transition. In practice, transitions with fixed delay in $T_\mathcal{E}$ are aimed at modeling synchronous firings, mainly useful during a reaction phase of the control part. External events to the control part are modeled by transitions T_∞ (input events) and T_T (delay events for the timed variables): it is assumed here that two transitions of these kinds may never occur simultaneously in the same instant. Then, only one firing in the set $T_\infty \cup T_T$ should trigger a reaction, that is (a sequence of) synchronous firings in the set $T_{\mathcal{E}_0}$. The transitions with interval $[0,0]$ may be used as well in a reaction phase as for an external event production to update some informations instantaneously.

In the sequel, the notation $\delta(t)$ is related to a transition t with a fixed delay as the static interval: $\delta(t) = ED(t) = LD(t)$.

The definition 1 is more general than the informal presentation in [11] where: T_∞ is made of a single transition [3] t such as $ED(t) = \varepsilon_0$ and $LD(t) = \infty$; and $\mathcal{E} \stackrel{\text{def}}{=} \{\varepsilon_0, \varepsilon_1, \varepsilon_2\}$.

2.3. Semantics

The chosen operational semantics of TPN is the *standard semantics* (as in the references [1, 4]). Moreover, among the enabled transitions, $T_{\mathcal{E}_0}$ transitions always have priority over those in $T_\infty \cup T_T$ (unlike [11] which considers no such priority), according to the synchronous hypothesis: the reaction time to an external event is always smaller than the delay before any next occurrence of external events. Then, non-infinitesimal elapsing of time is only possible when no transition in $T_{\mathcal{E}_0}$ is enabled: time elapse is considered dense in such a *stable* state. A next firing in $T_\infty \cup T_T$ may allow entering in a *reaction* phase where only transitions in $T_{\mathcal{E}_0}$ are fired until reaching a new stable state.

Among the well-known two characterizations of TPN state [3], interval state and clock state, the second one (which is the more general) is used to define the semantics of ε-TPN.

2. Given a set X, a multiset on X is a function $Y : X \to \mathbb{N}$.

3. This transition represents the delay observed, waiting for some external event to be produced to allow leaving from the stable state.

The semantics of a TPN may be defined as a transition system. Let be a vector v of size $n = |T|$ and with coefficients in $\mathbb{R}^+ \cup \mathcal{E}$: $v \in (\mathbb{R}^+ \cup \mathcal{E})^T$. $v(t_i)$ (for $i \in [1, n]$) represents a quantity of elapsed time related to the transition $t_i \in T$ (*local clock*[4] for t_i). The nil vector is $v_0 \in (0)^T$. A state q of the transition system is a pair $< M, v >$.

Définition 2. *The* timed transition system $< Q, \{q_0\}, \Sigma, \rightarrow >$ *of a marked ε-TPN is defined by:*

1) $q_0 =< M_0, v_0 > \in Q$*: the initial state;*

2) $Q \subseteq (\mathbb{N})^T \times (\mathbb{R}^+ \cup \mathcal{E})^T$*: the set of states (reachable from q_0);*

3) $\Sigma = T$*: the alphabet of the discrete transitions;*

4) $\rightarrow \subseteq Q \times (T \cup \mathbb{R}^{+*} \cup \mathcal{E}) \times Q$*: the relation of the timed and instantaneous transitions:*

a) *a timed transition is such as:*

i) $\exists d \in \mathbb{R}^{+*}$ *(non-infinitesimal time elapse),* $< M, v > \xrightarrow{d} < M, v' >$ *iff :*

$$\begin{cases} \nexists t \in En(M) \cap T_{\mathcal{E}_0} \\ v' \stackrel{\text{def}}{=} v + d \\ \forall t_k \in En(M) \Rightarrow v'(t_k) \leq LD(t_k) \end{cases}$$

ii) $\exists d \in \mathcal{E}$ *(infinitesimal time elapse),* $< M, v > \xrightarrow{d} < M, v' >$ *iff :*

$$\begin{cases} \exists t \in En(M) \cap T_{\mathcal{E}} \\ \forall t_k \in T, \begin{cases} \textit{if } t_k \in T_{\mathcal{E}_0} \textit{ then } v'(t_k) \stackrel{\text{def}}{=} v(t_k) + d, \\ \quad \textit{else } v'(t_k) \stackrel{\text{def}}{=} v(t_k) \\ t_k \in En(M) \Rightarrow v'(t_k) \leq LD(t_k) \end{cases} \end{cases}$$

b) *an instantaneous firing* $t_i \in En(M)$, $< M, v > \xrightarrow{t_i} < M', v' >$ *iff:*

$$\begin{cases} t_i \in T_{\mathcal{E}_0} \vee (\nexists t_j \in En(M) \cap T_{\mathcal{E}_0}) \wedge (t_i \in T_T \cup T_\infty) \\ M' \stackrel{\text{def}}{=} M \backslash {}^\bullet t_i \cup t_i^\bullet \\ ED(t_i) \leq v(t_i) \leq LD(t_i) \\ \forall t_k \in T,\ v'(t_k) \stackrel{\text{def}}{=} \begin{cases} 0 \textit{ if } t_k \in En(M') \\ \quad \wedge (t_k \notin En(M \backslash {}^\bullet t_k) \vee t_k = t_i) \\ v(t_k) \textit{ else} \end{cases} \end{cases}$$

According to the definition 2, in a given state q, when some transitions in the set $T_{\mathcal{E}_0}$ are enabled, one is fired instantaneously if its clock reached its fixed delay, or the least infinitesimal delay is observed among the enabled transitions in $T_{\mathcal{E}_0}$ to make a transition fireable; meanwhile, clocks for enabled transitions in $T_T \cup T_\infty$ do not change. Otherwise, only transitions in $T_T \cup T_\infty$ are enabled, and the common semantics for TPN is applied: a transition t must be fired instantaneously if its clocks reaches $LD(t)$, otherwise a wait delay d may be observed for each transition provided that it does not increase some clock beyong its LD, or any transition is fired if its clock value is inside the static interval. After a firing, the clocks value of the transitions $En(M')$ are updated in accordance with the standard semantics.

4. The clock of an unenabled transition does not change (and does not mind). Indeed, such a clock value is not taken into account to decide equality between two states.

During a reaction phase, it happens that parallel firings and variable updatings are computed in the same instant, constituting a *step* of fired transitions in the set $T_{\mathcal{E}_0}$. In the sequel, such a step is considered in only one interleaving of its firings, in order to reduce state explosion when computing the state space.

This semantics definition is different from the one adopted in [11] where priority is not given to firings in $T_{\mathcal{E}_0}$.

2.4. Enhancements on definitions in [11]

With the new definitions about ε-TPN in this section, some few changes have to be considered about the definitions and interpretations given in [11]. First, as a minor change, in the former phase sequencer (Fig. 1(a)), $ED(Change\_input)$ is now replaced by 0 (Fig. 1(b)): this trick was used to avoid possible firing of this transition concurrently with those in $T_{\mathcal{E}_0}$, which is useless now since this transition of T_∞ type has lesser priority.

Second, in [11], when only transitions in $T_T \cup T_\infty$ are fireable, it is possible of the occurrence of an input event in T_∞ to be followed by a firing in T_T concurrently with some transitions in $T_{\mathcal{E}_0}$, meaning some possible interference between the beginning of a reaction phase and an external event occurrence [5]. The new semantics of ε-TPN avoids such a case, by always giving the priority to the reaction phase.

3. State space construction

For reminder, the specificity of an ε-TPN is simulating the behaviour of a grafcet as a synchronous modeling language: the execution is an infinite alternation of stimulus (external event) followed by a reaction phase (called *evolution*) while no deadlock occurs, and a reaction may consist of iterated firings (which are sequential and/or concurrent) constituting a *firing stage*.

Ideally, a reaction must be made of a finite number of firings (meaning no livelock), and the possible interleavings of their concurrent and synchronous firings should lead up to the same state. The state space construction should cash in on this peculiarity to limit the potential state explosion, while providing an unexpensive way to check this expectation (see Subsection 3.2).

The state space construction of a TPN is classically based on abstractions of the timed transition system (definition 2), in shape of transition systems (without time on the transitions) called state class graphs (SCG): the nodes are state classes (which are generally agglomerations of an infinite number of states) with dense time. According to the purposes of state space generation, there is several kind of abstractions [3]. Here, the focus is on the linear SCG [2], knowing that the other types of abstraction may be deduced without difficulty from this one.

3.1. Computing a state class of an ε-TPN

A state class C is a couple of a marking M and a clock domain D (meaning a conjunction of time constraints characterizing the clock values of the enabled transitions). We denote by τ (resp. τ_i), the clock variable of a transition t (resp. t_i) appearing in a domain. The initial class is $C_0 = (M_0, D_0)$, such as $D_0 = \bigwedge_{t \in En(M_0)} \tau = 0$.

5. At the same time [11], when a timed event in T_T occurs firstly, it cannot be followed by the event $Change\_input$ (with interval $[\varepsilon_0, \infty[$) in concurrency with a firing in $T_{\mathcal{E}}$, which is a bit contradictory.

The algorithm 1 (in Annex A) describes the construction of the SCG for an ε-TPN.

For computation of class bounds, with ϵ delays, the supplementary arithmetic about time quantities is obvious. For $i, j \in \mathbb{Z}^*$ ($\varepsilon_n = -\varepsilon_{-n}$ if $n < 0$) and $d \in \mathbb{R}^*$: $d \pm \varepsilon_i = d$, $\infty \pm \varepsilon_i = \infty$, $\varepsilon_i \pm 0 = \varepsilon_i$, and $\varepsilon_i \pm \varepsilon_j = \varepsilon_{i \pm j}$.

For a transition enabled in a current class C (line 6), two scenarios are possible: either $En(M) \cap T_{\mathcal{E}_0}$ is empty (lines 8-15), or not (lines 17-27).

The first scenario which is the classic case, is reminded in Annex A.

The second scenario is specific to ε-TPN. The fireability check does not change just for the first firing (line 17). Computing D' (in four steps) changes slightly for every firing $t_f \in T_{\mathcal{E}}$ in this scenario. Indeed, an adjustement is necessary to express that such a firing is only possible after a discrete time : $d = \delta(t_f) - \tau_f$ will replace $d \geq 0$ at the first of the four steps. Moreover, for all $(t_i, t_j) \in T_{\mathcal{E}_0} \times (T_T \cup T_\infty)$ (with $t_i, t_j \in En(M)$), the constraint $\tau_i \leq \tau_j$ (conjunctive with the previous one) should be added at the second step, to express that an enabled transition not in $T_{\mathcal{E}_0}$ may not occur before anyone in $T_{\mathcal{E}_0}$.

After the first firing of the second scenario, iterated firings (lines 19-21 of the algorithm 1) are applied, supposing that synchronous firings lead to the same state whatever is the interleaving used. Thus, the fireability check may be simplified for each of the iterated firings: having $\tau_f = \delta(t_f)$ is obviously sufficient to infer a positive test. For each reached class C'' (line 20), the domain D'' is computed as after the first firing (line 18). Since intermediate state classes of iterated firings should not be stored (those states being just particular to the current interleaving), only the final reached class C' is saved (lines 23-26) and the transition to reach C' is the multiset of the iterated firings (line 27).

More informations are given about the algorithm 1 in Annex A. An application is given in Annex C.

3.2. Consistency check of an iterated firing stage

The line 22 of the algorithm 1 checks (maybe by a function) if the state reached by the *stage* T_t (firings done in the same instant) does not depend on the particular interleaving which was executed. Else, the problem should be reported and it means that the model has to be improved: for instance, contradictory orders from concurrent parts of the system controller may exist (set and reset the same output in two concurrent stored actions for example). Concurrency is a factor of explosion: for n concurrent firings, there is 2^n states covered by the potential $n!$ interleavings.

In case of strong concurrency, partial order techniques may be employed to cope with the explosion. Especially, the unfolding method [10] which will not interleave concurrent events forming a stage, may be used from the state just before the first firing (line 17) as the initial state. This is eased by the finiteness of the executions of a firing stage (in spite of the presence of inhibitor and read arcs). Knowing the expected unique final state (from the current interleaving), and knowing the transitions which may be fired in the same instant (i.e. verifying $\tau = \delta(t)$), the unfolding algorithm may be adapted (and based on the solution to the coverability problems described in [10]) to answer if any other final state may appear.

4. Conclusion

In this paper, the syntax and the operational semantics of ε-TPN are formally defined, and the construction of the corresponding state class graph taking into account the ε delays

is presented. That is complementary to the works in [11], by now preventing an overlap between an evolution phase and the occurrence of some external event, and by admitting only one external event before the subsequent reaction. This is achieved by introducing priority: the transitions for reaction ($T_{\mathcal{E}}$) have priority on the transitions for external events (T_T for quantitative timing events and T_∞ for input events).

An advantage of the proposed state space construction is the efficient elimination of the explosion due to concurrency during reactions, where no branching is displayed. To cope with the explosion caused by a multiplicity of input events, the tracks may consist in abstracting the states of inputs with no incidence on the subsequent reactions (while computing the SCG), and/or in modelling theirs dynamics to restrain the possible input changings.

Other perspectives are conceivable. One is to extend the translation rules to take into account hierarchy (macrostep, enclosure and forcing) in the grafcet: we hope that the more general definitions given in this paper will help to achieve this goal. Another one is to propose a model-checking framework for the grafcets, suitable to the specificity of the translation into ε-TPN.

5. References

[1] B. Berthomieu and M. Diaz. Modeling and verification of time dependent systems using time Petri nets. *IEEE Trans. Software Eng.*, 17(3):259–273, 1991.

[2] B. Berthomieu and F. Vernadat. State class constructions for branching analysis of time Petri nets. In *TACAS*, pages 442–457, 2003.

[3] H. Boucheneb and R. Hadjidj. CTL* model checking for time Petri nets. *Theor. Comput. Sci.*, 353(1):208–227, 2006.

[4] G. Bucci and E. Vicario. Compositional validation of time-critical systems using communicating time Petri nets. *IEEE Trans. Softw. Eng.*, 21(12):969–992, 1995.

[5] C. G. Cassandras and S. Lafortune. *Introduction to Discrete Event Systems*. Springer Publishing Company, Incorporated, 2nd edition, 2010.

[6] E. M. Clarke, Jr., O. Grumberg, and D. A. Peled. *Model checking*. MIT Press, Cambridge, MA, USA, 1999.

[7] D. L'Her, P. Le Parc, and L. Marcé. Proving sequential function chart programs using timed automata. *Theoretical Computer Science*, 267(1-2):141–155, 2001.

[8] IEC 60848. Grafcet specification language for sequential function charts. Technical report, International Electrotechnical Commission, 2013.

[9] IEC 61131-3. Programmable controllers - part 3: Programming languages. Technical report, International Electrotechnical Commission, 2013.

[10] K. L. McMillan. A technique of state space search based on unfolding. *Form. Methods Syst. Des.*, 6(1):45–65, 1995.

[11] M. Sogbohossou and A. Vianou. Formal modeling of grafcets with Time Petri nets. *IEEE Transactions on Control Systems Technology*, 23(5):1978–1985, Sept 2015.

[12] N. Wightkin, U. Buy, and H. Darabi. Formal modeling of Sequential function Charts with Time Petri nets. *IEEE Transactions on Control System Technology*, 19(2):455–464, 2011.

A. Algorithm to generate the state class graph of a ε-TPN

The algorithm 1 describes the construction of the state class graph for an ε-TPN.

```
1  Input: marked ε-TPN;
2  Output: sets Classes and Transitions representing the state class graph;
3  Classes := {C_0}; Transitions := {};
4  Stack the initial class C_0;
5  while the stack is not empty do
6      Unstack (LIFO) the class C = (M, D);
7      if ∄t ∈ En(M) ∩ T_{ε_0} then
8          foreach transition t fireable from the class C do
9              Compute the successor class C';
10             if C' ∉ Classes then
11                 Stack C';
12                 Classes := Classes ∪ {C'};
13             end
14             Transitions := Transitions ∪ {(C, {{t}}, C')};
15         end
16     else
17         Find any fireable transition t from C; T_t := {{t}};
18         Compute the successor class C';
19         while ∃t' ∈ En(M') ∩ T_{ε_0} fireable in the same instant as t do
20             Compute the successor class C'' from C'; T_t := T_t ∪ {{t'}}; C' := C'';
21         end
22         Make consistency check of the stage (T_t) between C and C';
23         if C' ∉ Classes then
24             Stack C';
25             Classes := Classes ∪ {C'};
26         end
27         Transitions := Transitions ∪ {(C, T_t, C')};
28     end
29 end
```

Algorithm 1. *Construction of the state class graph of an ε-TPN*

For a transition enabled in a current class C (line 6), two scenarios are possible: either $En(M) \cap T_{\varepsilon_0}$ is empty (lines 8-15), or not (lines 17-27).

In the first scenario which is the classic case [3], fireability of each enabled transition is tested, and when it is fireable, the successor class C' is computed. A fireability check of a transition $t_f \in (T_T \cup T_\infty)$ to fire includes the test of the current *domain consistency* achieved this manner: $D \wedge (d \geq 0) \wedge (ED(t_f) \leq \tau_f + d) \wedge (\bigwedge_{t \in En(M)} (\tau + d \leq LD(t)))$. Then, the new domain D' is computed in the following steps:

1) Initially, D' is $D \wedge d \geq 0$; then, replace each variable τ in D' by $\tau - d$;

2) Add the constraints: $ED(t_f) \leq \tau_f \wedge (\bigwedge_{t \in En(M)} \tau \leq LD(t))$;

3) Eliminate τ_f, d and all τ_c such as $t_c \in En(M) \wedge t_c \notin En(M \backslash {}^\bullet t_c)$;

4) Add the constraint $\tau = 0$ for each t newly enabled by M' (that is: $t \in En(M') \wedge (t \notin En(M \backslash {}^\bullet t) \vee t = t_f)$).

For this classic scenario where no transition in $T_{\mathcal{E}}$ is enabled, time progress is only dense and ε delays don't mind: each bound in a class domain with this kind of value should be replaced by the value 0.

Line 17 (in the second scenario) supposes that a transition will be always found to continue the execution of the program; it is guaranteed by the phase sequencer (Fig. 1) which is an infinite loop execution (no dead state is possible), even if an evolution may be *void*: the transition $Evolution\_end$ may be fired without firing any other reaction transition (when the previous occurred external event does not cause a real reaction phase).

The **while** loop (line 19) may potentially be infinite when a reaction phase is not finite. This part of the algorithm can be easily modified by saving apart the states C'' to detect and report the infinite loop in order to fix the problem in the model.

It should be noted that, by abstracting the **if** line 7 and the **else** block from the lines 16 to 27 which are specific to an ε-TPN, the classic algorithm is ensued.

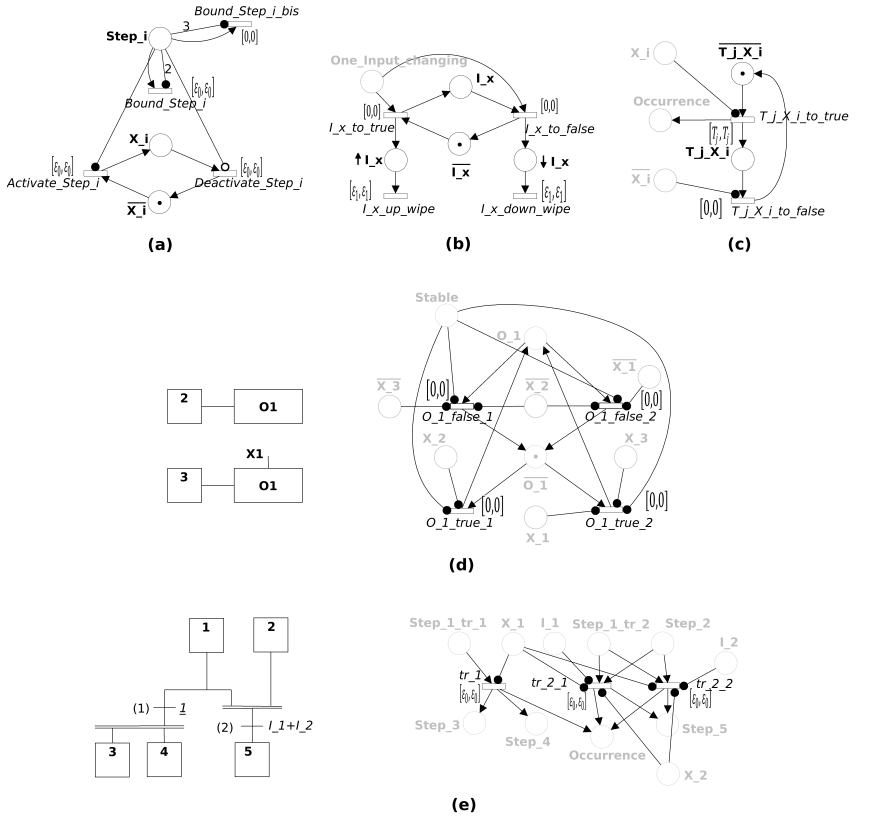


Figure 2. *Examples of translated elements of grafcet (according to [11])*

B. Translation of elements of grafcet into TPN [11]

The given elements of translation here are respectively (Fig. 2): a step i with its X_i state variables (Fig. (a)), an input with its rising and falling edges (Fig. (b)), a timed variable T_j/X_i (Fig. (c): the transition $T\_j\_X\_i\_to\_true$ is the only one type expressing an external event, with the transition $Change\_input$ Fig. 1(a) in a phase sequencer), an example of continuous and conditional actions (Fig. (d)), and an example of transitions with simultaneous divergence and convergence (Fig. (e)). Let notice that the elements in gray are added in a previous phase of the translation.

In this paper, no change concerns the translation rules proposed in [11], except slighty the phase sequencer as developped in the subsection 2.1.

C. Application

As an illustration of the newly defined semantics, an example of grafcet is provided Fig. 3(a), its translation into ε-TPN is provided Fig. 3(b) and the construction of its state class graph (SCG) is sketched Fig. 4.

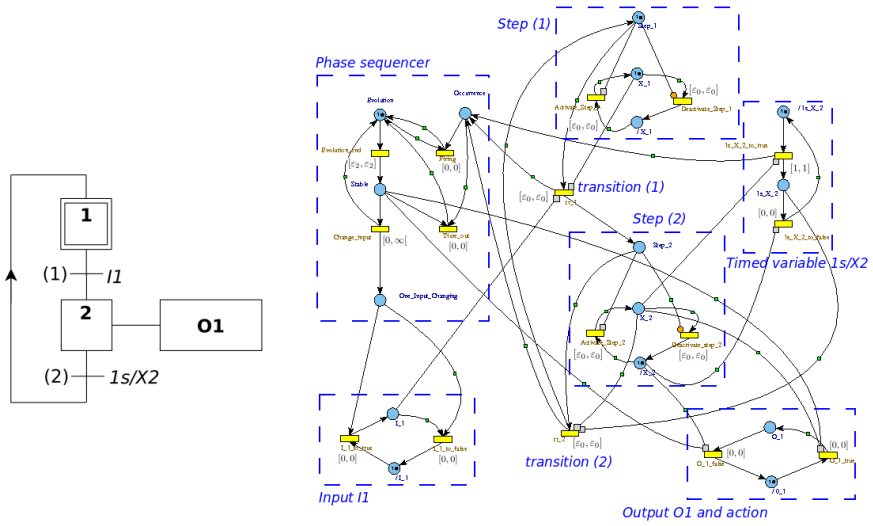


Figure 3. *An example of grafcet (a) and its translation into ε-TPN (b)*

The resulting ε-TPN (edited with Roméo [6] with superimposed images to specify intervals with ϵ delays) gets some simplified modules since the grafcet example is a safe model, and all litterals (such as the edges about the input I_1) are not useful. In general, the spatial complexity of the translation is globally polynomial with the number of nodes (steps and transitions), variables or literal terms of a grafcet [11].

6. Roméo, a tool for Time Petri Nets analysis: `romeo.rts-software.org`

The ε-TPN Fig. 3(b) is a safe model. Fig. 4 is just the initial part of the SCG (the twelve first classes). Each class is represented in a box with the upper part enumerating the marked places and the lower part giving the enabled transitions with their clock interval (the clock domain is so simplified).

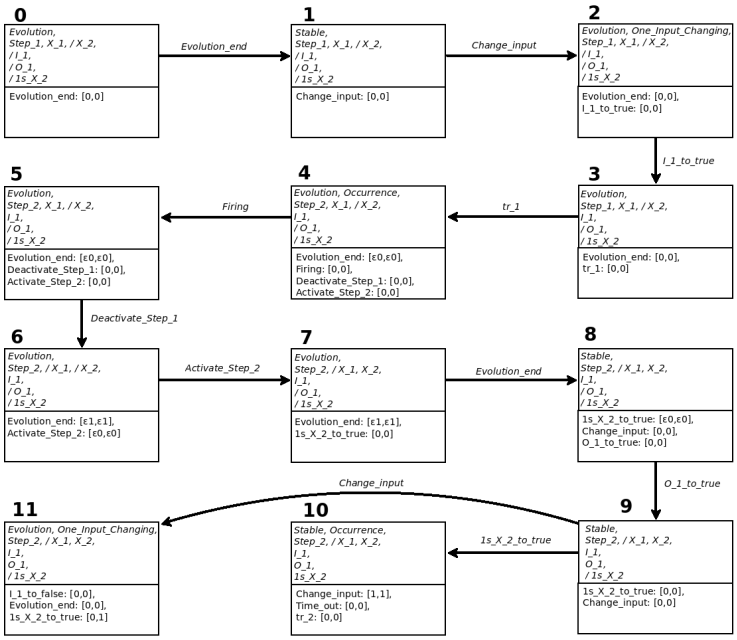


Figure 4. *The initial part of the SCG of the example Fig. 3(b)*

Altogether, 35 classes are covered (with 38 transitions) by the algorithm 1. Abstracting the intermediate classes during iterated firings eliminates 12 ones; for the given initial part: the class $\{C_4\}$ (resp. $\{C_6\}$, $\{C_8\}$) for the stage $\{\{tr\_1, Firing\}\}$ (resp. $\{\{Deactivate\_Step\_1, Activate\_Step\_2\}\}$, $\{\{Evolution\_end, O\_1\_to\_true\}\}$) do not appear in the set $Classes$. Abstracting the class C_6 avoids one of the two interleavings of the concurrent firings $Deactivate\_Step\_1$ and $Activate\_Step\_2$.

The SCG abstraction generated by Algorithm 1 holds however more informations than the first abstraction proposed in [11]: according to this last one, the class C_2 and the classes $\{C_4, C_5, C_6\}$ have to be abstracted too, and in the resulting macro-transition, only the firing $Change\_input$ (resp. tr_1) needs to be displayed for a behaviour verification, since other relevant informations are contained in the marking of the reached class C_3 (resp. C_7). With the same idea, the transition between C_7 and C_9 after the abstraction of C_8 only needs to carry the information $Evolution\_end$. Algorithm 1 is easily modifiable to enhance the abstraction in this way.

At last, in [11] was proposed a second and more compact abstraction [7] to only display the stable states of the grafcet. It will result in a SCG with only 4 classes and 7 transitions.

7. The goal of the first and the second abstractions in [11] was to only preserve the informations contained in the source grafcet.

Model-checking on grafcets through translation into time Petri net

Médésu Sogbohossou[1] — Rodrigue Yehouessi[1] — Bernard Berthomieu[2]

[1]Département Génie Informatique et Télécommunications
École Polytechnique d'Abomey-Calavi (EPAC), 01 BP 2009 Cotonou, BENIN
medesu.sogbohossou@epac.uac.bj, yehouessi_rodrigue@yahoo.fr

[2]LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, FRANCE
Bernard.Berthomieu@laas.fr

ABSTRACT. The conception of a critical automated system goes through its formal specification in order to proceed to its validation. One of the well-known formalisms to specify the behaviour of such a system is the GRAFCET standard (IEC 60848). GRAFCET being just a semi-formal language, we choose to use an intermediate language to translate without any ambiguity a grafcet model: the time Petri nets (TPN), which take into account quantitative time in a model. In this paper, we propose some verification formulas on GRAFCET charts, via the generated intermediate TPN model: CTL and SE-LTL temporal logics are used to express properties being about situations and actions of the GRAFCET chart. Then, we provide a procedure of implementation by using JGrafchart (a grafcet editor) and the model-checkers in TINA software, namely SELT (for SE-LTL properties) and MUSE (for CTL properties).

RÉSUMÉ. La conception d'un système automatisé critique passe par sa spécification formelle afin de procéder à sa validation. Un des formalismes répandus pour spécifier le comportement d'un tel système est la norme GRAFCET (IEC 60848). GRAFCET n'étant qu'un langage semi-formel, nous choisissons de passer par un langage intermédiaire vers lequel le modèle est traduit sans ambiguité: les réseaux de Petri temporels (TPN), qui prennent en compte le temps quantitatif dans un modèle. Dans cet article, nous proposons des formules de vérification sur les grafcets, via le modèle TPN intermédiaire généré: les logiques temporelles CTL et SE-LTL sont utilisées pour exprimer des propriétés portant sur les situations et les actions du diagramme grafcet. Ensuite, nous proposons une procédure de mise en œuvre passant par l'éditeur de grafcet JGrafchart et les model-checkers du logiciel TINA, à savoir SELT (pour les propriétés SE-LTL) et MUSE (pour les propriétés CTL).

KEYWORDS : grafcet (IEC 60848), Time Petri Net (TPN), model-checking, CTL, SE-LTL

MOTS-CLÉS : grafcet (IEC 60848), réseau de Petri temporel, model-checking, CTL, SE-LTL

1. Introduction

The formal verification of the automated systems [9] is essential before their realization because they are often critical systems and require important costs. There are several techniques for checking such discrete event systems: theorem proof, test, simulation and model-checking [8]. Model-checking is a computer-assisted method for the analysis of systems that can be modeled by state-transition formalisms. The model-checking software takes as input the model (an automaton) and the property to check on it. When a property is not satisfied for the studied system, model-checking may provide a counter-example.

The GRAFCET [1] formalism [11] is widely used by the automation specialists to describe the behavior of the sequential control part of an automated system, by the means of charts in the system specification phase. This standard should not be confused with the SFC one [12, 13] intended for implementation purposes.

However, GRAFCET (and SFC) formalisms are not mathematically defined, and so they contain some ambiguities to clarify via a translation of the chart into a formal representation, such as SMV textual language [15, 1, 14], timed automata [10] or time Petri nets (TPN) [16]. TPNs are structurally (and historically) closer to the GRAFCET than the other formal models.

Thus, the novelty of the present work is to propose a procedure for doing model-checking on a GRAFCET chart (or *grafcet* for short) translated into time Petri net (TPN) according to [16]. Based on the state space construction of a classic TPN [3, 5], two algorithms are implemented to obtain sufficiently compact abstractions which will be the inputs of a model-checker. The first algorithm only preserves informations in the original grafcet (by abstracting extra informations appearing after the translation). The second one enhances the abstraction and displays only states where the grafcet is in a stable situation. Futher, taking into account the specificities of the translation into TPN, some expressions of properties are proposed in CTL and LTL temporal logics, and concern situations and actions of the grafcet. Thanks to SE-LTL [7], it is specially possible to integrate transitions in a property formula.

For the practical experiences, the grafcet editor called JGrafchart [2] is used, and after implementing translation and abstractions, the model-checking is applied by means of two components of TINA software [3] [4], namely SELT (for SE-LTL model-checking) and MUSE (for CTL model-checking).

The remainder of this article is organized as follows. Section 2 shortly presents the used modeling formalisms, and introduces CTL and SE-LTL model-checking fragments. In Section 3 a set of formulas is proposed about the situations and actions of a grafcet. Section 4 describes the different practical steps to achieve model-checking of a grafcet, and contains a case study to illustrate our approach. Finally, Section 5 concludes this paper and gives some outlooks.

1. Acronym in French: *GRAphe Fonctionnel de Commande Etape Transition*.
2. JGrafchart, `http://www.control.lth.se/Research/tools/grafchart.html`
3. TIme petri Net Analyzer (TINA), `http://projects.laas.fr/tina`

2. Modeling formalisms and model-checking

2.1. GRAFCET charts

A GRAFCET chart [11] is a graphical representation modelling the behavior of the control part of an automated system. This representation consists of two parts:

– the *structure* describes the possible evolutions between the situations. It consists of the following basic elements: step, transition and directed link. A situation is the set of active steps at a given instant;

– the *interpretation* enables the relationship between the literal variables (inputs, outputs, delays, internal variables, ...) and the structure. It is done through the transition conditions (containing inputs, rising/falling edges of boolean inputs, delays, ...) and the actions (continuous actions, stored actions).

Figure 2 in Annex B shows an example of a grafcet edited with JGrafchart. It should be noticed that JGrafchart respects only partially the syntax (and the semantics) of the GRAFCET standard. For instance, a continuous action and a stored action on activation are defined respectively with qualifiers N and S (like in the SFC standard), and a timed variable T_j/X_i on a step i (with the value T_j in the second unit) is denoted by $S_i.s > T_j$.

2.2. Translation of grafcet into Time Petri net

Definition 1. *A* Time Petri Net *(TPN) [16] is a tuple* $(P, T, W, W_I, W_R, \downarrow SI, \uparrow SI, M_0)$ *such as:*

– *the nodes:* P *is the set of* places *and* T *is the set of* transitions *(*$P \cap T = \emptyset$*);*

– $W : P \times T \cup T \times P \longrightarrow \mathbb{N}$ *defines the* regular arcs *between nodes (and their weights);*

– $W_R : P \times T \longrightarrow \mathbb{N}$ *defines the* read arcs*;*

– $W_I : P \times T \longrightarrow \mathbb{N}^+ \cup \{\infty\}$ *defines the* inhibitor arcs*;*

– $\downarrow SI : \ T \longrightarrow \mathbb{Q}^+$ *(resp.* $\uparrow SI : \ T \longrightarrow \mathbb{Q}^+ \cup \{\infty\}$*) defines the lower (resp. upper) bound of the static interval of the transitions;*

– *the initial marking* $M_0 : P \longrightarrow \mathbb{N}$.

A marking M may enable some transitions in the set T. A transition firing is also conditionned by time information of all the enabled transitions, depending on their static intervals. A firing sequence expresses a behaviour of the modelled system. The standard semantics is used here and is more precisely recalled in a reference such as [6].

The works [16] have proposed a procedure of translating a grafcet into a TPN model, of which syntax is extended by ε infinitesimal delays as bounds on some transitions, allowing to simulate the synchronous semantics of GRAFCET. A extra module (called *phase sequencer*) is necessary to allow a transient evolution without modification of inputs as external events: it forces alternation between the reaction phase (called *evolution* with grafcets) and an external event production (an input change or some timed variable becomes true) in a stable situation. After adding this first module, the generation of the complete TPN is done by translating sequentially: the steps, the inputs, the timed variables, the outputs, the counter variables, the continuous and conditional actions, the stored actions and the grafcet transitions. These grafcet elements (steps, transitions, input variables, actions, ...) correspond to different but connected blocks in the resulting TPN.

The spatial complexity of the translation is polynomial with the number of nodes (steps and transitions), variables or literal terms of the grafcet.

2.3. Model-checking

A model-checking software takes as input an abstraction of the system behavior (a transition system such as a TPN state space [5]) and a property (expresses in Temporal Logic [8]) to check on the model, and answers if the abstraction satisfies or not this property. There are several types of temporal logic including : LTL (Linear Temporal Logic) to express properties on each path of the transition system and CTL (Computational Tree Logic) to express properties taking into account the branching of the different possible futures of the transition system.

A property p is formulated by means of a logical proposition (or formula), of which interpretation (i.e. true or false value) depends on a model $\mathcal{M}$ on which this property is expressed. Thus, the property p verified for the model $\mathcal{M}$ is denoted: $\mathcal{M} \models p$. For temporal properties about a discrete event system, the commonly used model is called *labeled Kripke structure* (LKS): it is a kind of state graph of which each state is labeled with some atomic propositions (in a set AP) true in this state; a transition between two states is labeled by a subset A of events in Σ. In our context, the model $\mathcal{M}$ is the state class graph (SCG) [4] obtained from the translation of a grafcet into an equivalent TPN [16], and the propositions concerns the marking of the different places in the TPN. Here, CTL and LTL temporal logics are used to express properties about situations and actions of the GRAFCET chart.

A path $\pi = (s_0, A_0, s_1, A_1, s_2, A_2, ...)$ of a LKS is an alternating infinite sequence of states ($s_0, s_1, ...$ with s_0 the initial state) and events ($A_0, A_1, ...$ with A_i a set of TPN firings from the state s_i). Notation π^i stands for the suffix of π starting in the state s_i.

The syntax of SE-LTL (State-Event LTL [7]) path formula is given by (where p ranges over AP and a ranges over Σ) :

$$\varphi := p \mid a \mid \neg\varphi \mid \varphi \vee \varphi \mid \varphi \wedge \varphi \mid \mathbf{X}\,\varphi \mid \mathbf{F}\,\varphi \mid \mathbf{G}\,\varphi \mid \varphi\,\mathbf{U}\,\varphi$$

For the SE-LTL semantics, a path-satisfaction of formulas is defined inductively as follows ($\mathcal{L}(s_0)$ is a subset of AP labeling s_0):

1) $\pi \models p$ iff $p \in \mathcal{L}(s_0)$, and $\pi \models a$ iff $a \in A_0$,
2) $\pi \models \neg\varphi$ iff $\pi \not\models \varphi$,
3) $\pi \models \varphi_1 \vee \varphi_2$ iff $\pi \models \varphi_1$ or $\pi \models \varphi_2$,
4) $\pi \models \varphi_1 \wedge \varphi_2$ iff $\pi \models \varphi_1$ and $\pi \models \varphi_2$,
5) $\pi \models \mathbf{X}\,\varphi$ iff $\pi^1 \models \varphi$,
6) $\pi \models \mathbf{F}\,\varphi$ iff $\exists k \geq 0$ s.t. $\pi^k \models \varphi$,
7) $\pi \models \mathbf{G}\,\varphi$ iff $\forall k \geq 0$, $\pi^k \models \varphi$,
8) $\pi \models \varphi_1\,\mathbf{U}\,\varphi_2$ iff $\exists k \geq 0$ s.t. $\pi^k \models \varphi_2$ and $\forall\, 0 \leq j < k$, $\pi^j \models \varphi_1$.

LTL is the restriction of SE-LTL without labels on transitions (i.e. just State LTL). Here, CTL does not consider events, like simple LTL.

The syntax of CTL state formula is given by (φ is a path sub-formula):

$$\phi := p \mid \neg\phi \mid \phi \vee \phi \mid \phi \wedge \phi \mid \mathbf{E}\,\varphi \mid \mathbf{A}\,\varphi$$
$$\varphi := \mathbf{X}\,\phi \mid \mathbf{F}\,\phi \mid \mathbf{G}\,\phi \mid \phi\,\mathbf{U}\,\phi$$

For the CTL semantics, a state-satisfaction of formulas is defined inductively as follows:

1) $s_0 \models p$ iff $p \in \mathcal{L}(s_0)$,

2) $s_0 \models \neg\phi$ iff $s_0 \not\models \phi$,

3) $s_0 \models \phi_1 \vee \phi_2$ iff $s_0 \models \phi_1$ or $s_0 \models \phi_2$,

4) $s_0 \models \phi_1 \wedge \phi_2$ iff $s_0 \models \phi_1$ and $s_0 \models \phi_2$,

5) $s_0 \models \mathbf{E}\ \varphi$ iff $\exists \pi = (s_0, A_0, \dots)$ s.t. $\pi \models \varphi$ for φ as a path sub-formula,

6) $s_0 \models \mathbf{A}\ \varphi$ iff $\forall \pi = (s_0, A_0, \dots),\ \pi \models \varphi$ for φ as a path sub-formula.

A path sub-formula φ in CTL is only in the form of: $\mathbf{X}\ \phi$, $\mathbf{F}\ \phi$, $\mathbf{G}\ \phi$ or $\phi_1\ \mathbf{U}\ \phi_2$, according to the same semantics of LTL (items 5-8), and where ϕ, ϕ_1 and ϕ_2 are state sub-formulas.

3. Model-checking on grafcets

We distinguish the properties according to the objects of the grafcet that they handle : situations or actions.

Some properties depend on the specificities of the approach of translation proposed in [16]. Some elements in the resulting TPN report the evolution states and the stability states of a grafcet: it is the case of the places called $Stable$ and $Evolution$, and the transition $Evolution\_end$.

3.1. Properties on the structural aspect

Let S be the set of steps of the considered grafcet.

The LTL properties on the structural aspect are the followings:

1) To find out whether the step X_i is permanently active: $\mathbf{G}\ X_i$

2) To test the existence of a step X_i permanently active: $\bigvee_{X_i \in S} \mathbf{G}\ X_i$

3) To verify that a step X_i is never active: $\mathbf{G}\ \neg X_i$

The CTL properties about the structural aspect are the followings:

1) To find out whether the step X_i is permanently active: $\mathbf{AG}\ X_i$

2) To test the existence of a step X_i permanently active: $\bigvee_{X_i \in S} \mathbf{AG}\ X_i$

3) To verify that a step X_i is never active: $\mathbf{AG}\ \neg X_i$

4) To know if in the future the step X_i could be permanently active: $\mathbf{EF\ EG}\ X_i$

5) To test the existence of any step active permanently in the future: $\bigvee_{X_i \in E} \mathbf{EF\ EG}\ X_i$

6) To check[4] whether the activity of the step X_j is reachable since the one of the step X_i: $\mathbf{AG}\ (X_i \Rightarrow \mathbf{AF}\ X_j)$

7) To check whether active step X_k is reachable from active step X_i through the active step X_j: $\mathbf{EF}\ (X_i \Rightarrow \mathbf{EF}\ (X_j \wedge (X_j \Rightarrow \mathbf{EF}\ X_k)))$

8) To check that it is possible to find a grafcet execution where the steps X_i, ..., X_n are activated simultaneously: $\mathbf{EF}\ ((\neg X_i \wedge ... \wedge \neg X_n) \wedge \mathbf{EX}\ (X_i \wedge ... \wedge X_n))$

9) To check whether it is possible to return to step X_i or to verify that a step X_i is accessible from all grafcet situations: $\mathbf{AG\ EF}\ X_i$

4. $\phi \Rightarrow \varphi$ is equivalent to $\neg\phi \vee \varphi$.

10) To check if there is a grafcet situation where there is a deadlock (that is to say a situation that can no longer be left): **EF EG** $(Evolution \Rightarrow Evolution\_end)$

11) To check if there may be total instability in the system: **EF EG** $\neg Evolution\_end$

In fact, the two last properties are not valid in CTL since $Evolution\_end$ is an event. To make such properties valid in a State-Event CTL such as UCTL [2], any classical proposition $Prop$ only made with events (i.e. transition firings) should be replaced by $\mathbf{AX}_{Prop}$ **true**; so, $Evolution\_end$ will become here $\mathbf{AX}_{Evolution\_end}$ **true**.

3.2. Properties on the actions

Let S_j be the set of steps associated with the action $action_j$, T_{j_1} (resp. T_{j_2}) the set of succeeding (resp. preceding) transitions of the steps associated with the action $action_j$. The possible forms of $action_j$ are:

- $action_j$ is a continuous action: $(\bigvee_{X_i \in S_j} X_i) \wedge Stable$
- $action_j$ is a conditioned action by $condition_j$ (a logical expression): $(\bigvee_{X_i \in S_j} X_i) \wedge Stable \wedge condition_j$
- $action_j$ is a stored action (translatable into $\mathbf{AX}_{action_j}$ **true** in UCTL):
  - on activation : $\bigvee_{tr_i \in T_{j_2}} tr_i$
  - on deactivation : $\bigvee_{tr_i \in T_{j_1}} tr_i$

These different forms are used to check the following LTL and CTL properties :

1) LTL property : to show that an action $action_1$ always follows an action $action_2$: $action_2 \Rightarrow \mathbf{F}\ action_1$

2) CTL properties:

a) To show that an action $action_1$ always follows an action $action_2$: **AG** $(action_2 \Rightarrow$ **AF** $action_1)$

b) To show that an action $action_1$ is launched simultaneously with an action $action_2$: **EF** $((\neg action_1 \wedge \neg action_2) \Rightarrow$ **EX** $(action_1 \wedge action_2))$

Naturally, some more general property may mix up both action and step propositions.

4. Implementation of the model-checking

4.1. Procedure

To make model-checking on the grafcet, we proceed as follows:

1) The grafcet to be verified is edited under the JGrafchart software (as shown the figure 2 of the case study in Appendix B). This software generates an XML file containing information on the elements of the grafcet;

2) From the XML file, our Java program generates a **.net** extension file containing information about the elements of the TPN equivalent to the edited grafcet;

3) The implementation of the algorithms 1 and 2 (Annex A) allows us to obtain respectively from the file **.net**, a file **.aut** containing the information on the elements of

the automaton (with unstable and stable states) of the grafcet and another one containing only the information on the stable states of the grafcet (by disregarding unstable states);

4) The TINA **ktzio** tool takes the **.aut** file as input to generate the Kripke structure (**.ktz** extension file on which the verifications are made);

5) Finally, the tools SELT and MUSE (examples in Appendix C) of TINA are used to check the LTL and CTL properties on the grafcet from the Kripke structure.

4.2. Application

The illustration is based on the grafcet as shown in Figure 2 (Annex B). This grafcet models two traffic lights located respectively on a track A and a track B. It contains a transient mode (orange lights blink three times) and a steady state. From the JGrafchart XML file, we generated the equivalent TPN and the two automata of figures 3 and 4 (in Annex B, edited from the generated **.aut** files).

The following examples of properties are checked on the grafcet.

Verification of a LTL property with SELT tool:

• The system leaves the transient mode (firing of transition 13): TRUE. The result of this verification [5] is shown in Figure 1.

```
C:\tina-3.4.4\bin>selt feuxTricolore.ktz
Selt version 3.4.4 -- 01/05/16 -- LAAS/CNRS
ktz loaded, 42 states, 42 transitions
0.000s

- [] (tr_13 => () (S0 /\ S4));
TRUE
0.016s
```

Figure 1. *Verification of the exit from the transient mode on the first automaton (with stable and unstable states).*

Verification of some CTL properties with MUSE tool:

• The street A light can stay permanently green: FALSE. Cf. Figure 5 (Annex C).

• The counter that allows blinking of the orange lights in the transient mode can reach the value 4: FALSE. Cf. Figure 6 (Annex C).

• Lights can become green or orange, simultaneously for streets A and B: FALSE. Cf. Figure 7.

• The same lights can pass simultaneously to two different colors (green and red for example): FALSE. Cf. Figure 8.

5. Conclusion

Through these works, we have shown the possibility to check properties (SE-LTL and CTL respectively with the tools SELT and MUSE of the software TINA) on a grafcet after translating it into an equivalent TPN, and subsequently into an automaton representing the state-space. This automaton is as compact as possible by abstracting much information in the TPN and by avoiding multiple interleavings due to the concurrent firings in the TPN.

5. With SELT, operators **G**, **X** and $\wedge$ are denoted resp. [ ], ( ) and /\.

Contrary to the grafcet translation into Timed Automata [10] or TSMV [14], TPNs do not allow model-checking on quantitative time properties with TCTL logic. To introduce timed properties, a perspective to our approach is to integrate observers into the TPN of the translation, to take into account delay events while model-checking the grafcet. Another perspective is the creation of a software implementing all steps of our approach: from editing a grafcet (in full conformity with the IEC60848 standard) until the verification phase. Finally, the extension of CTL to Action/State-Based Temporal Logic UCTL [2] will be an asset to generalize the expression of some properties including events of firing.

6. References

[1] N. Bauer, S. Engell, R. Huuck, S. Lohmann, B. Lukoschus, M. Remelhe, and O. Stursberg. Verification of PLC programs given as sequential function charts. In *Integration of Software Specification Techniques for Applications in Engineering*, pages 517–540, 2004.

[2] M. H. Ter Beek, A. Fantechi, S. Gnesi, and F. Mazzanti. An action/state-based model-checking approach for the analysis of communication protocols for service-oriented applications. In *FMICS*, pages 133–148. Springer, 2007.

[3] B. Berthomieu and F. Vernadat. State class constructions for branching analysis of time Petri nets. In *TACAS*, pages 442–457, 2003.

[4] B. Berthomieu and F. Vernadat. Time Petri nets analysis with TINA. In *Quantitative Evaluation of Systems, QEST 2006.*, pages 123–124. IEEE, 2006.

[5] H. Boucheneb and R. Hadjidj. CTL* model checking for time Petri nets. *Theor. Comput. Sci.*, 353(1):208–227, 2006.

[6] G. Bucci and E. Vicario. Compositional validation of time-critical systems using communicating time Petri nets. *IEEE Trans. Softw. Eng.*, 21(12):969–992, 1995.

[7] S. Chaki, E. M. Clarke, J. Ouaknine, N. Sharygina, and N. Sinha. State/event-based software model checking. In *IFM*, vol. 2999, pages 128–147. Springer, 2004.

[8] E. M. Clarke, T. A. Henzinger, and H. Veith. *Introduction to Model Checking*, pages 1–26. Springer International Publishing, Cham, 2018.

[9] D. Darvas, I. Majzik, and E. B. Viñuela. PLC program translation for verification purposes. *Periodica Polytechnica, Electrical Engineering and Computer Science*, 61:151–165, 2017.

[10] D. L'Her, P. Le Parc, and L. Marcé. Proving sequential function chart programs using timed automata. *Theoretical Computer Science*, 267(1-2):141–155, 2001.

[11] IEC 60848. Grafcet specification language for sequential function charts. Technical report, International Electrotechnical Commission, 2013.

[12] IEC 61131-3. Programmable controllers - part 3: Programming languages. Technical report, International Electrotechnical Commission, 2013.

[13] A. Karatkevich. *Petri Nets in Design of Control Algorithms*, pages 1–14. Springer International Publishing, Cham, 2016.

[14] N. Markey and P. Schnoebelen. TSMV: A symbolic model checker for quantitative analysis of systems. In *QEST*, vol. 4, pages 330–331, 2004.

[15] T. Ovatman, A. Aral, D. Polat, and A. O. Ünver. An overview of model checking practices on verification of PLC software. *Softw. Syst. Model.*, 15(4):937–960, October 2016.

[16] M. Sogbohossou and A. Vianou. Formal modeling of grafcets with time petri nets. *IEEE Transactions on Control Systems Technology*, 23(5):1978–1985, Sept 2015.

A. Algorithms

Two algorithms are proposed and implemented (in Java) to obtain sufficiently compact abstractions.

Algorithm 1. *SCG (LTL): first abstraction*

```
1  Save and stack (LIFO) the initial state;
2  while (the Stack is not empty) do
3      Unstack a state (or state class);
4      if (a grafcet transition is fireable) then
5          Fire all simultaneously fireable grafcet transitions;
6          Fire all fireable transitions for modifying literals;
7          if (the last reached state is new) then Save and stack it;
8      else if (Evolution_End is fireable) then
9          Fire transition Evolution_End;
10         Fire all fireable transitions for continuous actions;
11         if (the last reached state is new) then Save and stack it;
12     else if (Change_input or a delay transition of some timed variable model are fireable) then
13         for each fireable transition do
14             Fire all fireable transitions until a grafcet transition or Evolution_End is fireable;
15             if (the last reached state is new) then Save and stack it;
16         end
17     end
18 end
```

Algorithm 2. *SCG (LTL): second abstraction*

```
1  Stack (LIFO) the initial state;
2  while (the Stack is not empty) do
3      Unstack a state (or state class);
4      if (a grafcet transition is fireable) then
5          Fire all simultaneously fireable grafcet transitions;
6          Fire all fireable transitions for modifying literals;
7          if (the last reached state is new) then Stack it;
8      else if (Evolution_End is fireable) then
9          Fire transition Evolution_End;
10         Fire all fireable transitions for continuous actions;
11         if (the last reached state is new) then Stack it;
12     else if (Change_input or a delay transition of some timed variable model are fireable) then
13         for each fireable transition do
14             Fire all fireable transitions until a grafcet transition or Evolution_End is fireable;
15             if (the last reached state is new) then Save and stack it;
16         end
17     end
18 end
```

Algorithm 1 only preserves informations in the original grafcet, by abstracting extra informations appearing after the translation: for instance, firings related to the synchronous updatings (step states and literal variables) are abstracted.

Algorithm 2 is the same as Algorithm 1, except that only state classes corresponding to the stable situations of the grafcet (line 15) are saved, and only $Change\_input$ and delay transition firings from theses classes are displayed.

The two algorithms assume that all possible interleavings of firings which symbolize a grafcet evolution between two stable situations lead to the same global state. This assumption (and other prerequisites) were discussed in [16].

B. The case study

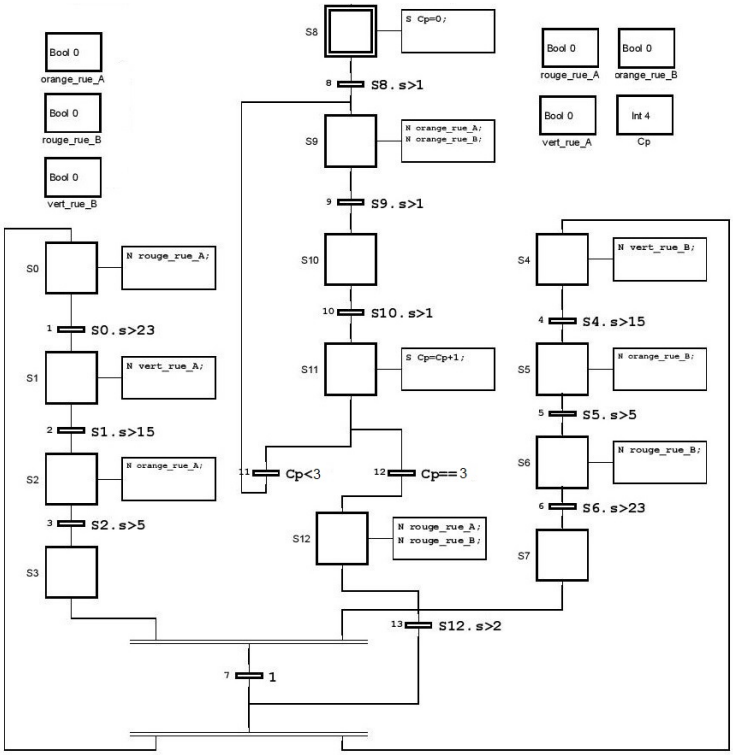


Figure 2. *Case study*

From the JGrafchart XML file of Figure 2, we have generated the equivalent TPN and the two automata of figures 3 and 4 (edited graphically with the tool ND of TINA taking as input the generated **.aut** files). To summarize:

– the TPN obtained is made of 75 places, 96 transitions, 205 regular arcs, 110 read arcs and 20 inhibitor arcs;

– the first abstraction (figure 3) is made of 42 states and 42 transitions;

– the second abstraction (figure 4) is made of 12 states and 12 transitions.

Figure 3. *Automaton based on algorithm 1 (stable and unstable states of the grafcet)*

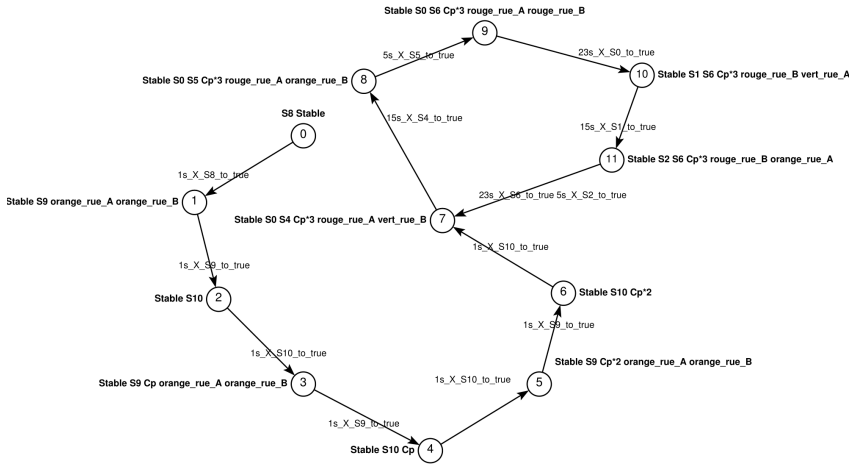


Figure 4. *Automaton based on algorithm 2 (stable states of the grafcet)*

C. Some TINA results of the case study

These results concern CTL properties tested with MUSE tool, only by using the second automaton (with only stable states).

```
C:\tina-3.4.4\bin>muse feuxTricolore2.ktz -prelude ctl.mmc -b
Muse version 3.4.4 -- 01/05/16 -- LAAS/CNRS
ktz loaded, 12 states, 12 transitions
0.000s
- EF EG vert_rue_A;
it : states
FALSE
0.000s
```

Figure 5. *Checking the permanent activation of the green light of the street A*

```
C:\tina-3.4.4\bin>muse feuxTricolore2.ktz -prelude ctl.mmc -b
Muse version 3.4.4 -- 01/05/16 -- LAAS/CNRS
ktz loaded, 12 states, 12 transitions
0.000s
- EF (Cp=4);
it : states
FALSE
0.000s
```

Figure 6. *Checking the state of the counter*

```
C:\tina-3.4.4\bin>muse feuxTricolore2.ktz -prelude ctl.mmc -b
Muse version 3.4.4 -- 01/05/16 -- LAAS/CNRS
ktz loaded, 12 states, 12 transitions
0.016s
- EF ((vert_rue_A /\ vert_rue_B) \/ (vert_rue_A /\ orange_rue_B) \/ (vert_rue_B
/\ orange_rue_A) \/ (S1 /\ S5));
it : states
FALSE
0.000s
```

Figure 7. *Checking a safety property*

```
C:\tina-3.4.4\bin>muse feuxTricolore2.ktz -prelude ctl.mmc -b
Muse version 3.4.4 -- 01/05/16 -- LAAS/CNRS
ktz loaded, 12 states, 12 transitions
0.016s
- EF ((vert_rue_A /\ rouge_rue_A) \/ (vert_rue_A /\ orange_rue_A) \/ (rouge_rue_
A /\ orange_rue_A));
it : states
FALSE
0.000s
```

Figure 8. *Checking no two lights on at the same place*

Fig. 5 shows that a green light will never stay indefinitely turned on. Fig. 6 shows that the orange lights in the transient mode will not blink more than three times. Fig. 7 displays that the lights for the two crossing streets will never allow all road users to pass through simultaneously. Finally, Fig. 8 shows that two lights may not turn on simultaneously for some user.

Adjustment module

to give auto-adaptiveness behavior to flood forecasting systems

Joel TANZOUAK*, Idrissa SARR+, Blaise YENKE*, Ndiouma BAME+, Serigne FAYE+

*The University of Ngaoundere (Cameroon),
+The University of Cheikh Anta Diop (Senegal)

RÉSUMÉ. La prévision est aujourd'hui un facteur clé dans la minimisation des dégâts causés par les inondations. En effet, les systèmes de prévision d'inondations (FFS) fonctionnent pour la plupart dans les pays developpés et utilisent des modèles hydrauliques pour fournir des prévisions du niveau et/ou du débit des rivières en se basant sur les prévisions meteo (NWP). Ces données de prévision sont utilisées pour fournir des alertes d'inondations; Il est donc important d'utiliser de bons modèles hydrauliques pour obtenir des données précises. De nombreux modèles hydrauliques ont été construits pour les FFS. Cependant, la différence entre les paramètres environnementaux et climatologiques entre les régions rend très difficile l'utilisation de ces FFS dans d'autres régions. De plus, l'évolution constante au cours du temps de l'environnement, causée par des facteurs anthropiques, nécessite un processus de recallage fréquent des modèles hydrauliques pour qu'ils s'ariment aux changements environnementaux. Par conséquent, il est nécessaire de construire des FFS qui s'adaptent dynamiquement aux changements environnementaux sans processus de recallage. L'objectif de cet article est de proposer une extension des FFS en introduisant un module d'ajustement qui utilise des données collectées en temps réel à partir de réseaux de capteurs combinées avec des données prévisionnelles issues des modèles hydrauliques, pour donner une capacité d'auto-adaptation dynamique aux FFS. Les résultats obtenus à partir d'expériences empiriques montrent les avantages de notre mécanisme d'ajustement dans l'auto-adaptation des FFS.

ABSTRACT. Forecasting is now a key factor in minimizing the damages caused by Flood. Indeed, Flood forecasting systems (FFS) operate mostly in developed countries and use hydraulic models to provide forecasts of river levels and / or flow, based on numeric weather predictions (NWP). These forecast data are used to provide flood alerts, so it is therefore important to use good hydraulic models to obtain accurate flood forecast. Many hydraulic models have been built for FFS. However, the difference between environmental and climatological parameters between regions makes very difficult the use of these FFS in other regions. Moreover, the constant evolution over time of the environment, caused by anthropic factors, need a frequent process updates of hydraulic models so that they can be adapted to environmental changes. Therefore, it is necessary to build FFS that dynamically adapt to environmental changes without a recall process. The purpose of this article is to propose an extension of FFS by introducing an adjustment module that uses real-time data collected from sensor networks combined with predictive data from hydraulic models, to provide FFS with a dynamic self-adaptation ability. The results obtained from empirical experiments show the advantages of our adjustment mechanism in the self-adaptation of FFS

MOTS-CLÉS : Modèle hydraulique, Réseau de Capteurs, Module d'ajustement, Auto-adaptabilité

KEYWORDS : Hydraulic model,Sensors network, Adjustment module, Auto-adaptiveness

1. Introduction

According to the statistics in recent decades, floods represent almost 40% of natural disasters in the world [6], and most of these flood are caused by rivers overflow. This issue is very challenging for governments in terms of on-time prevention. Many solutions used to face flood are usually built for developed countries, and are not well suited for African countries because of geographical and climatological differences. Concerning developing countries, mainly in Africa, a high number of victims and damages observed could be explain by these reasons :

– low data accuracy of numeric weather prediction to have good flood forecast quality,

– lack of meteorological information in many regions that makes the flood forecasting difficult,

– Low forecast accuracy of FFS

Concerning the first two problems, we proposed two solutions in [2] and [3]. It is important to say that the resolution of NWP data accuracy does not necessary resolve the FFS forecast accuracy since the data accuracy is one of the FFS accuracy problem but not the only problem. The goal of this work is to resolve the third problem, by adding auto-adaptability to FFS and reducing the inaccuracy of forecast. In this respect, we target the main output hydraulics model parameters of FFS such as river's level forecast assuming that they are the source of flood forecast inaccuracy when there are environmental changes. In others words, we suppose that the forecast system is working correctly and if ever an inaccuracy is observed in the forecast results, then, one of the main reasons are linked to the hydraulics models that provides output river's level. The problem are then probably caused by environmental changes over space or time.

The main contribution of this paper is the proposition of an adjustment module that gives the auto-adaptiveness ability to FFS. Basically, we developed an algorithm that uses past forecast and observed data to assess the error in the future FFS forecast and correct the forecast. The observed data are collected from sensor networks.

The rest of the document is organized as follows : Section II presents hydraulic flood forecasting systems features. The scope of this work is given in Section III. In Section IV, the adjustment algorithm designed to improve auto-adaptiveness is presented. Section V provides a discussion on the evaluation of the proposed algorithm. Conclusion and future directions ended this work.

2. FFS based on Hydraulic models

FFS predict the upcoming of flood in a certain delay. According to [11], advances in flood forecasting have been slackened by the ability to assess rainfall continuously over space. According to [2] FFS can be categorized into two groups : the sensors-based systems [8, 10, 13, 1, 9, 7] and hydraulic models based systems [15, 14, 12]. In the remaining of this section, we aim to portray FFS based on hydraulic models.

Medium and long term flood warning system are more required for densely populated areas in order to give enough time for reducing damages. To predict medium or long term flood, they rely on both Numeric weather prediction(NWP) and Hydraulic Models. Like the forecasting of weather where models are used to simulate atmosphere behavior, hydraulic models can be used to simulate rivers behavior based on rain forecast in order to estimate whether a flood may occur.

2.1. How FFS based on hydraulic models work ?

FFS based on hydraulic models usually need recalibration and updates when they are used in a region other than the one for which they were designed for (eg. a FFS built in Europe can not properly work in Africa without some recalibration process). In fact, NWP data are received from meteorological stations and other data like topography, vegetation, and so on are collected from other systems such as GIS (Geographical Information Systems). The entire data are sent to the Hydraulic model. The output of the hydraulic model is then used by FFS to evaluate flood risk. This output could be for example, the river's level or river's flow.

2.2. The problems of FFS based on Hydraulics models

FFS based on hydraulic models are able to predict water levels at any location in the modeled area for a given date if ever the NWP data are available. According to [4] one of World's best practice in FFS incorporates hydraulics models and data assimilation. Nevertheless there are some problem that we can observed in these FFS :

– Existing FFS are built for specifics region, and it could have some inaccuracy problem if these FFS are installed to another regions because of environmental differences. So Existing FFS need to be updated and re-calibrated if we want to use them in some other regions. This recalibration task are not usually easy and in certain cases, it is needed to rebuild the hydraulic model used in the FFS.

– There are some important parameters in the environment which are able to change over time and which are not taken into account. So if the environment meet some changes, existing FFS will not consider these changes, and this situation could seriously affect FFS forecast accuracy.

We realized that this type of systems has a main problem : auto-adaptiveness regarding environmental and regions changes. Our target here is to build a module which is able to give auto-adaptiveness ability to FFS in order to increase forecast accuracy. The next section presents the proposed solution.

3. Auto-adaptiveness ability of FFS based on hydraulics models

The main challenge here is to improve the coherence of hydraulic model forecast data when environment parameters are changed or have involved. We consider that as much as hydraulic model are auto-adaptive regarding a given region, the system will be geographically portable. To get our goal, we rely on a basic hydraulic model described above to propose a new approach.

3.1. Auto-adaptiveness hydraulic FFS

We leverage the basic model presented before by including others modules. Basically, observed data are collected from sensors network and they are used with forecast data produced by the hydraulic model, in the system to realize the adjustment task. These forecast and observed data are used as input in a learning program in charge of errors prediction.

An error is the difference between a prediction P(t) at a time t and the observation Obs(t) at a time t as shown in (1).

$$error(t) = P(t) - Obs(t) \tag{1}$$

The learning program uses a set of past errors, to make the future error prediction. After, the learning program sends the predicted error to the adjustment program, which applies the required update (Figure1) before output of the new adjusted forecast.

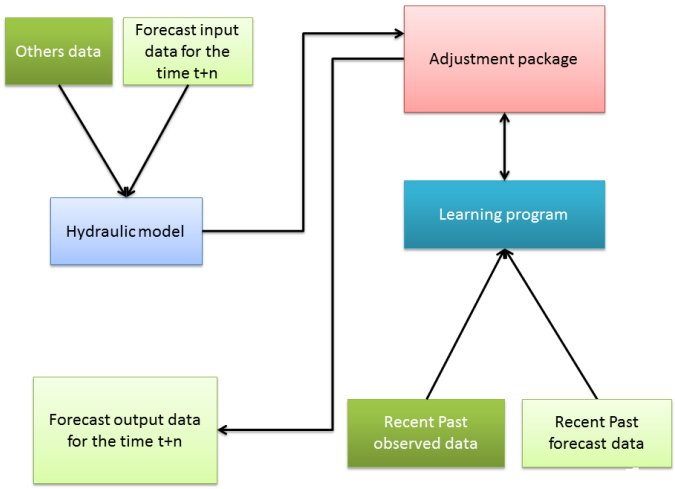


Figure 1 : Proposed hydraulic FFS architecture

3.2. Adjustment module description

3.2.1. Basic concepts

We consider a sensor network as a set of sensors at the risk zone that collect real-time data. Thus, the system collects forecast data from hydraulic model and observed data collected from sensors network installed across the river. A learning program evaluates error between forecast and observed data. When a new input forecast data is received, the adjustment program relies on the learning program to adjust them regarding future predicted error. These data are afterwards used by the system to evaluate flood risk.

3.2.2. Adjustment formalization

The learning program uses the function (2) to make prediction of the future error using past errors. Function (2) is a basic exponential smoothing [5] method of time series data used to make short term predictions. In this paper, forecast and observed data, are daily data.

$$\varepsilon(t) = \alpha * e(t) + (1 - \alpha) * \varepsilon(t - 1) \tag{2}$$

where

– **e(t)** represents the error (1) measured between a prediction and an observation at the time **(t)**

– $\varepsilon(t)$ represents the prediction of the error at a time **(t)**

– α represents the auto-adaptiveness speed. $\alpha \in [0; 1]$

Regarding (2), if $\varepsilon(t-1) > 0$ it will means that the prediction P(t) made by the FFS will be greater than the waiting observation Obs(t), so to adjust the predicted value P(t), we need to apply $A(t) = P(t) - \varepsilon(t-1)$ and consider A(t) instead of P(t).

But if in other hand, $\varepsilon(t-1) < 0$ it will means that the prediction P(t) made by the FFS will be lesser than the waiting observation Obs(t), so to adjust the predicted value P(t), we need to apply $A(t) = P(t) + |\varepsilon(t-1)|$

Based on the two formulations above we can define (3) as a general expression of the adjustment equation.

$$A(t+1) = P(t+1) - \alpha * e(t) - (1-\alpha) * \varepsilon(t-1) \tag{3}$$

where :

– **P(t+1)** is the prediction received from the hydraulic model at a time **(t+1)**

– ε**(t-1)** represents the prediction of the error at a time **(t-1)**

– **A(t+1)** represents the adjusted prediction value at the time **(t+1)**

The role of algorithm 1 is to evaluate the error after each prediction regarding to the observations. Obs(t) is the collected value at time t sent by a sensor. Algorithm 2 aims to adjust the prediction received from hydraulic model according to a predicted error.

Data: New observation data Obs(t)
Result: evaluated error e(t) between P(t) and Obs(t)
while *New observation Obs(t) is received* **do**
    – e(t)=P(t)-Obs(t)
    – Save (t, e(t))
end

Algorithm 1: Error evaluation

Data: New prediction data P(t)
Result: Adjusted value A(t) of the prediction
initialization (α);
while *New prediction P(t) is received* **do**
    – Predict the potential error on the prediction received using (2)
  if *($\varepsilon(t-1) > 0$)* **then**
    A(t)=P(t)-$\varepsilon(t-1)$;
    Save (t, P(t), A(t));
    Use A(t) instead of P(t) for the flood alert evaluation;
  else
    A(t)=P(t)+$\varepsilon(t-1)$;
    Save (t, P(t), A(t));
    Use A(t) instead of P(t) for the flood alert evaluation;
  end
end

Algorithm 2: Data adjustment

4. Adjustment algorithm validation

4.1. Data sets

To evaluate our adjustment module, we used measured groundwater level data collected daily during 1 hydraulic year in a small town in Senegal and we also used groundwater forecast data made by an hydraulic model for the region during the same time.

4.2. Scope of the test

The scope of this test is to show how the algorithm uses past errors made between the forecast data and observation data, to adjust the future forecast data in the aim of reducing the difference between hydraulic prediction and real observation even if the model is not suitable for the zone.

We want to remind that if we used a model which is not suitable for the region, it is to materialize the fact that the environment has changed or the model has been taken to another regions instead of the region where it was designed. So if the model is in another region or if the environment of the model has changed, it will be not suitable for the new condition and will need recalibration. We want to show by this simulation how the adjustment module can reduce the inaccuracy of the model due to these changes.

In others words, this test shows how, if an hydraulic model designed and calibrated for a given region is taken to another region, the forecast error done by the hydraulic model in the new region will be adjusted gradually until the system reach to a stable mode where the error between prevision of the hydraulic model and the observation realized on the groundwater will be minimized.

4.3. Experimentation

This part presents the evaluation method and the different tools used to validate the algorithm efficiency. To implement and run our solution, we used a MySql database, and the python language.

4.3.1. Evaluation

To characterize the efficiency of our algorithm, let us consider δ, the accuracy indicator defined in (4).

$$\delta = \frac{1}{N} \sum_{i=0, Y_i \neq 0}^{N-1} \mid \frac{Y_i - X_i}{Y_i} \mid \tag{4}$$

where Y represents observed value, X the forecast value and N the number of data observed.

4.4. Result and discussions

Figure 2 presents the measured data collected, the forecast data provided by the hydraulic model and the data provided by hydraulic model coupled with the adjustment module. All these data are for a period of approximatively 1 year for the same region. According to this figure, it can be observed that the difference between forecast of the hydraulic model and the observation is significant, as the accuracy indicator for forecast data is δ=0.08. This difference between the hydraulic model and the observation could be

explained by the fact that the hydraulics model was not suitable for the zone and needed to be calibrated to work well. This situation could illustrate the potential behavior of an hydraulic model when the initial environment has changed.

It can be also observed that the difference between forecast of the hydraulic model coupled with adjustment module and the observation is very low, as the accuracy indicator for the hydraulic model coupled with the adjustment module is δ=0.003 which is largely smaller than the accuracy indicator of the hydraulic model alone (0.003 over 0.08). This low difference between the hydraulic model coupled with adjustment module and the observation could be explained by the auto-adaptiveness behavior that the adjustment module gives to the hydraulic model. At each time the adjustment module tries to anticipate the error on the future forecast and use it to adjust the prediction received from the hydraulic model.

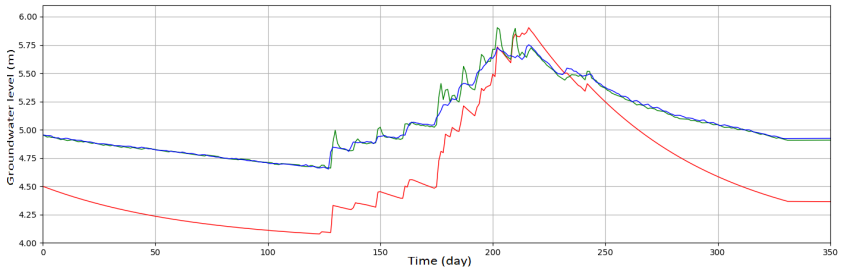


Figure 2 : Forecast realized by (hydraulic model + adjustment module) in Green, Observations in Blue and Forecast realized by model without adjustment module in red

This figure portrays in a very simple way the efficacy of the adjustment module in the improvement of the forecast capacity of the hydraulic Model. The difference between the two models materialized by the different values of δ shows that the adjustment module has a great impact on the forecast, over 96.25% of accuracy improvement. However, some fluctuations can be observed at the beginning of each regime. This is because at each time the adjustment module tries to adapt itself to the new regime and if the regime is changing continuously over the time the fluctuation will continue, because the module will always try to adapt itself to the new regime.

5. Conclusion

We presented a module which is able to improve the forecast quality of hydraulic models when the model is not suitable or well calibrated for a given region. The results obtained show that this module improves the accuracy of forecast data and can play a significant role in the improvement of the auto-adaptiveness of flood forecasting systems capacity. For future works, we plan to propose a module that manages the integration of observed and forecast data collected from sensor networks and meteorological stations.

Références

[1] Elizabeth Basha and Daniela Rus. Design of early warning flood detection systems for developing countries. *IEEE*, 2012.

[2] Joel TANZOUAK. Ndiouma BAME. Blaise YENKE, Idrissa SARR. Sytem to imporve numeric wheather predictions for flood forecasting systems. *IEEE-SITIS conference*, 2017.

[3] Joel TANZOUAK. Ndiouma BAME. Blaise YENKE, Idrissa SARR. System to extend forecast capacity of meteorological stations. *ACM-ICGDA conference*, 2018.

[4] Blake Boulton and Terry van Kalken. Hydraulic models needed for flood forecasting. *Water Engeneering Australia*, June 2011.

[5] Jr. Gardner. Exponential smoothing : The state of the art. *Journal of Forecasting*, 4, 1985.

[6] http ://www.emdat.be/database. Emergency database. 2016.

[7] Dany Hughes and Phil Greenwood. An intelligent and adaptable grid-based flood monitoring and warning system. *Computing Department, Infolab21, Lancaster University, UK*, 2006.

[8] Kabita Sahoo Indira Priyadarshinee and Chandrakant Mallick. Flood prediction and prevention through wireless sensor networking (wsn) : A survey. *International Journal of Computer Applications*, 113(9) :7, march 2015.

[9] Sunkpho J. and Ootamakorn. Real-time monitoring and warning system. *Songklanakarin Journal of Science and Technology*, 2(33) :8, 2011.

[10] Kavi Kumar Khedo. Real-time flood monitoring using wireless sensor networks. *The Journal Of The Institution Of Engineers Mauritius*, page 11, 2014.

[11] Robert J. Moore and Steven J. Cole. Issues in flood forecasting : Ungauged basins, extreme floods and uncertainty. *Frontiers in Flood Research*, 2006.

[12] F. Pappenberger, K.J Beven, and al. Cascading model uncertainty from medium range weather forecast (10 days) through a rainfall-runoff model to flood inundation predictions within the european flood forecasting system (effs). *Hydrology and heart system Science*, 4(9), 2005.

[13] Victor Seal and Arnab Raha. A simple flood forecasting scheme using wireless sensor networks. *International Journal of Ad hoc, Sensor and Ubiquitous Computing*, 3(1) :16, 2012.

[14] J. Thielen, J. Bartholmes, and al. The european flood alert system – part 1 : Concept and development. *Hydrology and Heart System Science*, 13, 2009.

[15] V. Thiemig1 and B. Bisselink1. A pan-african medium-range ensemble flood forecast system. *Hydrology and heart system science*, 19 :21, 2015.

Building ego-community based on a non-closed neighborhood

Ahmed OULD MOHAMED MOCTAR and Idrissa SARR

Department of Mathematics and Computer Science
Cheikh Anta Diop University
Dakar - Fann BP, 5005, Senegal
amed.mohameden@gmail.com, idrissa.sarr@ucad.edu.sn

RÉSUMÉ. La détection de communautés attire, de plus en plus, l'attention de la communauté scientifique vu son rôle central dans la recherche des caractéristiques des réseaux, de même pour en exploiter les informations. Toutefois, les travaux existants s'intéressent plus au partitionnement des réseaux en plusieurs communautés, appelées souvent "communautés plates". Une autre façon consiste à détecter les communautés à partir des noeuds d'intérêt. Ainsi, nous parlons de "communautés ego-centrées". Elles permettent de comprendre, outre les interactions entre un élément et ses voisins, les interactions qui se passent entre les voisins eux mêmes. Dans un travail précédent, nous avons proposé un algorithme de détection de communautés ego-centrées qui se limite au voisinage direct. Ceci ne reflète pas la réalité de communication, car les amis des amis d'une personne peuvent avoir un impact significatif sur son comportement. L'objectif de ce papier est de proposer une extension de l'algorithme précédent permettant la détection de communautés ego-centrées au-dela du voisinage direct. Une évaluation de la nouvelle solution avec un jeu de données réel montre son gain sur la précédente approche.

ABSTRACT. Community detection arouses, more and more, the interest of scientific community given its central role in research on network characteristics and in the mining of network information. However, most existing works are interested in partitioning networks into communities, called "flat communities". Another way is to detect communities from nodes of interest, such communities are called "ego-communities". They allow to better understand not only the interactions between an element and their neighbors, but also the interactions that are happening between neighbors themselves. In a previous work, we proposed an algorithm for detecting one-step ego-centered communities. That is, the approach is limited only to close neighborhood (*i.e.* nodes directly linked to the *ego*). However such an approach does not reflect the communication real state since friends of a person's friends can impact on its behavior. The purpose of this paper is to extend our previous algorithm by enlarging the neighborhood range. We also implement our algorithm and use adolescent friendship network to demonstrate the effectiveness of our algorithm.

MOTS-CLÉS : détection de communauté, fonction de qualité, communauté ego-centrée

KEYWORDS : community detection, quality function, ego-community

1. Introduction

Social networks analysis has shown benefits of studying and/or analyzing the structural properties of social networks. One of the main topics of social network analysis is community detection, which is a process of decomposing the network into groups that satisfy a set of characteristics. In fact, detecting communities is splitting the network into groups that can be helpful for several use cases. For instance, identifying the group of people who has been in contact (by any communication device) with an Ebola's patient give a huge boost on the target to be monitored or to be informed about the risk they incur.

In short, community detection studies can be divided into two main categories, namely, global community detection (*i.e.*, macroscopic vision of the network) and local community detection (*i.e.*, microscopic vision of some nodes of the network). The first category seeks to partition the network into several communities without any focus on nodes and their role. In contrary, the second category builds communities from a few nodes of the network that we call nodes of interest or "ego". Communities obtained from nodes of interests are named ego-centered communities. Basically, an ego-centered community can be built from an "ego" and its closed neighborhood (alters) as well as it can be defined based on a larger neighborhood (the neighbors of the alters).

In [8], we proposed an algorithm for detecting ego-communities based on closed neighborhood. This paper aims to extend the previous algorithm in order to detect ego-communities based on a larger neighborhood. Moreover, we revised one of the metric, called affinity degree, we used in [8] to select node candidates that form a community. The main contributions of our work can be summarized as follows :

– An improvement of the affinity degree in order to prevent evaluating relevances of nodes that doe not have outgoing links from the ego-community ;

– An algorithm for detecting ego-communities while considering both the alters and their neighbors, the neighbors of their neighbors and so forth. Actually, the algorithm works in two phases :*1*) a selection-phase that helps choosing nodes and *2*) a removing-phase that defines step-by-step the community structure ;

– A validation of our algorithms through experiments conducted over a real data sets that shows the feasibility of our approach and its efficiency.

For this purpose, our paper is organized as follows : firstly, we define, in section 2, the preliminary concepts needed to understand the rest of this article. Then, we present, in section 3, the main existing ego-centered community detection approaches. In Section 4, we present the weakness of the affinity degree that we proposed in a previous paper and we explain how we correct these weaknesses. Section 5 is used for describing the process of building ego-communities beyond direct neighborhood. Finally, we evaluate the efficiency of our solution in section 6 and conclude in section 7.

2. Definitions

In this section, we define the basic concepts needed to fully understand the rest of the article.

Définition 1 (Eccentricity)**.** *Eccentricity of a node u, denoted $e(u)$, designates the number of links used to connect u to the most distant node in network. Otherwise, $e(u)$ repre-*

sents the maximum distance that can exist between u and any other node in network. The eccentricity of a node u is given by the following formula :

$$e(u) = max_{\forall v \in V(G)} d(u, v) \quad [1]$$

Définition 2 (Node of interest)**.** *We call a node of interest or "ego" any node that, by its status, can influence the behavior of other nodes in a given network.*

Définition 3 (k-neighborhood)**.** *We define a k-neighborhood of an "ego" e as all nodes that can be reached with a path length lesser or equal to k. Formally, a node j is in the k-neighborhood of e if $d(e, j) \leq k$. The value of k is between 1 and the eccentricity of u. Ego's neighbors can be linked to each other.*

The figure 1 shows an example. The nodes colored in orange are direct neighbors while pinks are in 2-neighborhood as well as greens are 3-neighborhood. For sake of simplicity, the "ego" with its k-neighborhood is called "ego-network" in the remainder.

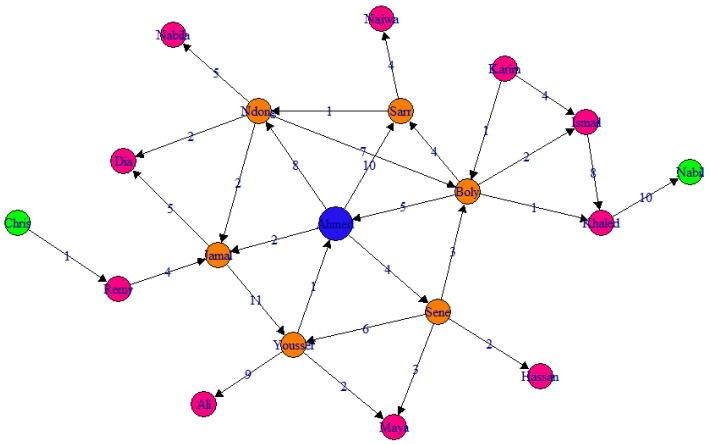


Figure 1 – Illustration of a directed and weighted ego-neighborhood.

Définition 4 (Ego-centered community)**.** *An ego-centered community is a group of nodes made of an ego and its k-neighbors so that the interactions number within the group is predominant compared to the one between group elements and the rest of network.*

3. Related work

To our best knowledge, there is not yet an algorithm for detecting ego-centered community beyond the direct neighborhood. However, we present in this section the existing algorithms for direct ego-community detection. Among the existing works, two approaches have attracted our attention.

3.1. Approach of quality function optimization

Considered as the most used approach [9, 3, 7], it consists in expressing the evaluation criteria of community relevance in the form of metric, called "quality function". The metric associates any community with a quality score. But, since applying a one-time quality

function does not guarantee community relevance, the result of the quality function can be optimized. The goal of this optimization is to minimize/maximize the quality score in order to detect a more relevant community than the original one. There are several ways to express a quality function that depends on how the algorithm defines a community.

The optimization of a quality function is done in three steps. It consists of starting with the node(s) of interest and adding or removing, in a iterative manner, the node that maximizes or minimizes the quality function score until we reach a given threshold.

3.2. Approach based on proximity

Such an approach was introduced recently in [5]. The overall principal is to find out nodes that are enough similar to the “ego" and then gather them as a community. To this end, proximities values are calculated for all nodes with respect to the “ego". Thus, nodes are ranked based on their proximities values in order to figure out whether there exists skew. In other words, if several nodes are almost similar to the “ ego ”, then, they form a stage on the curve of proximities. Rather, nodes that are strictly different follow the stage from afar since they have low values. Once this classification done, nodes of the stage are considered as the community of the “ ego ”. Nonetheless, the proximity curve usually decreases steadily with a non-scale law fashion, which means that a node may belong to several communities of different sizes or does not belong to any one.

Even though there exists plethora of ego-community detection solutions, we can mention a set of drawbacks or limitations as follows :

1) the detection approach considers only the “ego" and its alters ;

2) the topological structure are the only features they use ignoring their intensity for example ;

3) the links direction are not taken into account while computing metrics.

In this paper, we envision a solution facing these limitations.

It should be noted that there are also some supervised classification methods derived from machine learning but often used in community detection. Let’s take for example the k-nearest neighbors algorithms which classify the target elements (nodes) according to their distance from the nodes constituting the learning sample. Thus, these algorithms detect flat communities as they classify the nodes of network in different communities. However, in this paper, we are interested in ego-community detection algorithms.

4. Affinity degree

In a previous paper [8], we have proposed an algorithm for detecting ego-community detection based only on direct neighborhood. This algorithm seeks to optimize a quality function taking into account the weight and orientation of links. To this end, we have also defined a metric that called “affinity degree” allowing to select the most connected node at each iteration of optimization process of the quality function. In this section, we illustrate the weakness of the affinity degree we proposed in [8]. Then, we explain how to correct it.

In fact, the affinity degree $\mathcal{A}_v(\mathcal{C}_u)$ captures how connected is a node v regarding to the community $\mathcal{C}_u$. We evaluate the affinity degree of a given node according to 2 aspects :

– Separation of node from the rest of network : does the node have more links within the community than outside ? Does it communicate more with the nodes of the community or with the nodes located elsewhere ?

– Level of connectivity within the community : how much the node is connected to the nodes of the community ? Does it communicate further with them ?

To handle the first aspect, we propose the criterion $Separation_v(\mathcal{C}_u)$ which allows to measure the separation of a node v from a community C_u :

$$Separation_v(\mathcal{C}_u) = \frac{|N_v \cap \mathcal{C}_u|}{d(v)} \times \frac{d_w^{\mathcal{C}_u}(v)}{d_w(v)} \quad [2]$$

Where :

– $|N_v \cap \mathcal{C}_u|$ is the number of common neighbors between v and $\mathcal{C}_u$;

– $d(u)$ represents the centrality degree of u ;

– $d_w^{\mathcal{C}_u}(u)$ is the sum of weights of all links sharing between the node v and $\mathcal{C}_u$;

– $d_w(u)$ is the weighted degree of u.

Note that $Separation_v(\mathcal{C}_u)$ value's varies between 0 and 1. If $Separation_v(\mathcal{C}_u) = 0$, thus any v's neighbor is in $\mathcal{C}_u$ while $Separation_v(\mathcal{C}_u) = 1$ refers that all neighbors of v are in $\mathcal{C}_u$.

Although $Separation_v(\mathcal{C}_u)$ can classify nodes according to their level of separation from the community, $Separation_v(\mathcal{C}_u)$ does not distinguish between a node having only one neighbor in the community and another one having as many. In fact, this weakness is manifested in the case where the nodes do not have links outside the community, thus, $Separation_v(\mathcal{C}_u)$ considers that all these nodes are at the same level of relevance, which is not necessarily the case. Therefore, we found it necessary to add another criterion allowing to get an idea about the connectivity degree of nodes within the community.

To evaluate the connectivity degree of a given node v to a community C_u, we propose the following formula :

$$Connectivity_v(\mathcal{C}_u) = \frac{1}{d(v) \times d_w(v)} \quad [3]$$

The intuitive idea behind is that the more a node is very connected to C_u (having many neighbors inside and communicating further with them), the more the score of this criterion will be close to 0.

In the worst case, we can find a node having only one link in the community whose the weight equal to 1, which means that $Connectivity_v(\mathcal{C}_u) = 1$. In the best case, the value of $Connectivity_v(\mathcal{C}_u) \simeq 0$.

Now, we define our degree of affinity as follows :

$$\mathcal{A}_v(\mathcal{C}_u) = Separation_v(\mathcal{C}_u) \times (1 - Connectivity_v(\mathcal{C}_u)) \quad [4]$$

In formula 4, we calculate $1-Connectivity_v(\mathcal{C}_u)$ so that the bounds of $Separation_v(\mathcal{C}_u)$ and $Connectivity_v(\mathcal{C}_u)$ are coherent.

This way of selection has the advantage to define a selection order according to the affinity degree values. The β nodes with the smallest value are selected first. The intuition behind this is that nodes whose affinity degree is small are more likely not to be part of the community. However, this approach is time consuming since we need to estimate the affinity degree of each block of nodes before deciding whether it will be removed or not.

Moreover, the complexity is calculated based on the formula 4. Actually, the first part is computed with a complexity of $4n$, the second part is computed with a complexity of $2n$, which leads to an overall selection cost of $\frac{6n}{\beta}$.

5. Ego-community detection

In this section, we describe the procedure that follows our algorithm to detect k-step ego-centered communities.

5.1. k-neighborhood extraction

To extract the k-neighborhood of a given node u, we use the BFS[1] Algorithm [2]. Our implementation of BFS is to extract neighbors from the ego by step of neighborhood. Otherwise, we extract, first, the set of nodes N_1 that are directly connected to the ego. Next, we extract the set of nodes N_2 that are linked to N_1. Then, we extract the set of nodes N_3 that are linked to N_2. The process of nodes extraction continues until we reach the set of nodes N_k that are linked to the ego through a path length k. We use a vector I for saving the extracted nodes. At each iteration, we save the obtained nodes in I at the condition that they are not already present.

Let's take as example figure 1, the table 1 illustrates the process of 3-neighborhood extraction of node Ahmed. As illustrated on row 7 of table 1, at the second iteration, the neighbors of Sene are Hassan and Maya. Furthermore, Youssef's neighbors are Maya and Ali. But, since Maya is already added to the vector I. Our algorithm adds this time only Ali. The same case occurs between Jamal and Ndong since they have Dia as a common neighbor. Note that the final result is the union of results of all iterations after deleting the repeated neighbors.

First iteration	
Result : Boly, Sene, Youssef, Jamal, Ndong, Sarr	
Second iteration	
Node	*Neighbors*
Boly	Karim, Ismail, Khaled
Sene	Hassan, Maya
Youssef	Maya, Ali
Jamal	Remy, Dia
Ndong	Dia, Nabila
Sarr	Najwa
Result of second iteration : Karim, Ismail, Khaled, Hassan, Maya, Ali, Remy, Dia, Nabila, Najwa	
Third iteration	
Node	*Neighbors*
Khaled	Nabil
Remy	Chris
Result of third iteration : Nabil, Chris	

Table 1 – Illustration of the 3-neighborhood extraction process using BFS algorithm.

1. Breadth-First Search.

The time complexity of BFS can be expressed as $O(n + m)$ such as n is the number of nodes in k-neighborhood and m the number of links.

5.2. Sorting and selecting nodes

After extracting and saving the k-neighborhood of the ego in a vector, we use quick-Sort algorithm's [6] to sort the vector in ascending order according to the affinity degree of nodes. The time complexity of sorting and selecting nodes is $O(n + n\log(n))$. The overall complexity of our algorithm to detect an ego-community from k-neighborhood is $\frac{n}{\alpha}O(2n + n\log(n))$. Such as $\frac{n}{\alpha}$ is the number of iterations done throughout the optimization process of the quality function.

5.3. The Algorithm

Our algorithm initializes the ego community with the k-neighborhood. Then, it selects from the vector I a block of the most useless nodes, namely, those whose affinity degree is the smallest. Next, if the withdrawing of the nodes block from the community reduces more the quality score, they will be removed, otherwise, we leave them in the community and we move on to other ones. This procedure is repeated until the quality score reaches a given threshold.

The block size α is a parameter of our algorithm. In this paper, we are not interested in the impact of block size on the resulting communities.

6. Experimentation

In this section, we aim at assessing the efficiency of the proposed solutions and their feasibility through experimentation. To this end, we implement our algorithms with igraph R package [4] and use an adolescent friendship network [1] containing 2539 nodes and 12969 links. See Fig. 2. It's a directed and weighted network that was created from a study that includes a 90-minute home interview and aims to improve adolescent health in the United States. Each student was asked to list his 5 best female and his 5 male friends. A node represents a student and a link between two students shows that the left student has chosen the right one as a friend. The higher the weight value, the more students interact between them. We mention the existence of 6 big ego-neighborhoods. However, for a purpose of experiment, we focus only on a single ego-neighborhood, namely, the one whose the ego-node label is " 1 " because of its predominance (it's the lowest neighborhood in Fig. 2, colored in purple). Note that in our experiments, we set $k = 2$ that means we build community with a 2-neighborhood at most.

In the remainder of this section, we compare the ego-community algorithm with a larger neighborhood versus ego-community with only alters [8] to prove that the detected ego-communities by the first one are more relevant. Then, we show an example of detected ego-communities of each algorithm to illustrate the difference.

6.1. Algorithms effectiveness

In this part, we aim at evaluating the effectiveness of the previous algorithm we proposed in [8] and that one we proposed in this paper. To this end, we compare them regarding to the 2 criteria (ego-community isolation and separation). Fig. 3a and Fig. 3b show the quality function scores of the 2 algorithms.

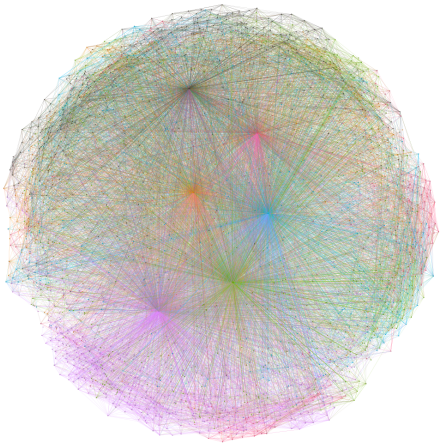


Figure 2 – Visualization of adolescent health network. For readability reasons, we did not display neither the nodes labels nor the links directions.

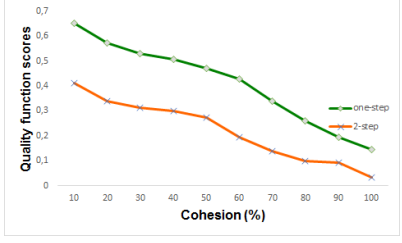
On Fig. 3a, we notice that the more the cohesion increases, the more the quality score decreases, this is explained by the fact that the closer we get to the total cohesion (complete sub-graph), the more the majority of nodes becomes relevant in terms of topological connections. However, we notice that the quality of k-neighborhood algorithm is always better than that of one-step ; we interpret the space difference between the 2 curves on Fig. 3a as follows : as the 2-neighborhood ego-community includes more nodes than 1-neighborhood ego-community, the quality of the first one will always be better by increasing the cohesion.

By analogy, we run another set of experiments in order to evaluate the quality function regarding to the separation of the ego-community.

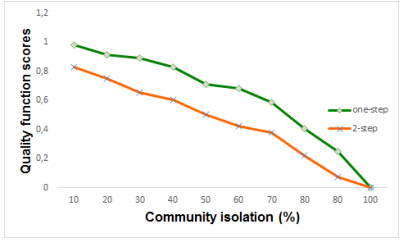
In this respect, we vary the community isolation (its disconnection with the rest of the network) from 10% (*i.e.*, higher communication of the group with the rest of the network) to 100% (total isolation). The community isolation is calculated using the following formula :

$$separation\_rate = \frac{\frac{\mathcal{W}_C^{in}}{|E_C|}}{\frac{\mathcal{W}_C^{in}}{|E_C|} + \frac{\mathcal{W}_C^{out}}{|E_C|}}$$

[h]



(a) Quality functions of 1-neighborhood algorithm vs 2-neighborhood algorithm regarding to the cohesion.



(b) Quality functions of 1-neighborhood algorithm vs 2-neighborhood algorithm regarding to separation of the rest of network.

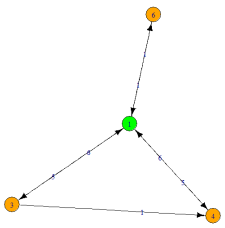
Figure 3 – Quality functions of 1-neighborhood algorithm vs 2-neighborhood algorithm.

Precisely, we divide the intensity of the communication within the group by the intensity of all communication involving one member of the group.

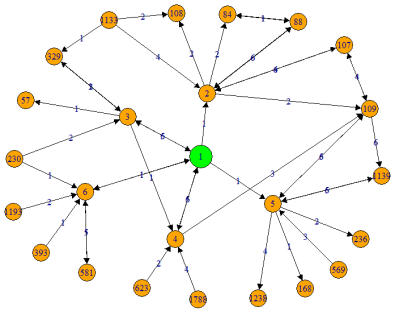
We observe on Fig. 3b three things : firstly, more the communication intensity increases, more the quality scores become small. Secondly, the quality of k-step algorithm is always better than that given by one-step algorithm. Thirdly, we also notice that where the community isolation is maximum, the two algorithms tend towards the same quality, namely 0, because if the isolation reaches 100%, it means that the value of outgoing link weights is zero, thus, the quality score will also be 0.

6.2. Algorithms impact on community structure

To show the pertinence of our k-neighborhood algorithm on the community structure of node whose label is " 1 " of the network depicted on Fig. 2 and extract both the 1-neighborhood and 2-neighborhood ego-communities using algorithm cited in [8] and the one we presented in this paper (i.e. k-neighborhood algorithm).



(a) 1-neighborhood ego-community of ego-node whose label is " 1 ", detected by the algorithm cited in [8].



(b) 2-neighborhood ego-community of ego-node whose label is " 1 ", detected by the k-neighborhood algorithm.

Figure 4 – Detected ego-communities of node " 1 " by one-step and 2-steps algorithms.

As each student lists his 5 best male/female friends, the one-step neighborhood of each node is composed of 6 nodes, including the ego-node, and 15 links, maximally. The minimization process of quality function carried out by the algorithm, cited in [8], led to the rejection of nodes labeled by "2" and "5". Therefore, the one-step ego-community extracted from the neighborhood of ego-node "1" consists only of 3 ego's neighbors, see Fig. 4a. This calls into question the weakness of one-step algorithm [8], because, often, the one-step neighborhood may not be rich of information, that is why, it is necessary to be able to go back towards the top of the hierarchy in order to understand the topological and semantic connections of the node of interest.

On Fig. 4b, we show the ego-community detected by k-neighborhood algorithm. Then, it is clear that with the possibility of broadening the neighborhood depth, we can have an idea about friends of friends of a person, which allows us to, either make a suggestion of friends, or predict a new behavior that the person inspired by its neighbors. Note that for the 2-neighborhood ego-community to be readable, we set the quality threshold ε to 0.4, since if ε is very small, by having a lot of nodes and links, the detected ego-community may be illegible.

To conclude, we proposed the k-neighborhood ego-community detection algorithm in order to cover the weakness of the algorithm cited in [8] which is limited to the direct neighborhood. Experiments have shown that the results provided by k-neighborhood algorithm are more relevant than those of one-step, in terms of internal cohesion and separation from the rest of the network.

7. Conclusion

In this article, we proposed a k-neighborhood ego-community detection algorithm in order to overcome the weakness of the algorithm cited in [8] which is limited to direct one-step neighborhood.

This work can be extended in several ways. For instance, the principle of functioning of the propositions is easily parallelized since the creation of each ego-centered community is made independently of the others. Thus, it is possible to assign a thread to each community building in order to improve the performance of the algorithm. Finally, it is also possible to take into account the detection of multi-level hierarchical communities by recursively applying the algorithm to each community detected.

Références

[1] Adolescent health network dataset – KONECT, September 2016. Available at http://konect.uni-koblenz.de/networks/moreno_health.

[2] Alan Bundy and Lincoln Wallen. Breadth-first search. In *Catalogue of Artificial Intelligence Tools*, pages 13–13. Springer, 1984.

[3] Jean Creusefond, Thomas Largillier, and Sylvain Peyronnet. On the evaluation potential of quality functions in community detection for different contexts. In *International Conference and School on Network Science*, pages 111–125. Springer, 2016.

[4] Gábor Csárdi, Tamás Nepusz, and Edoardo M Airoldi. Statistical network analysis with igraph. 2016.

[5] Maximilien Danisch, Jean-Loup Guillaume, and Bénédicte Le Grand. Multi-ego-centered communities in practice. *Social network analysis and mining*, 4(1) :1–10, 2014.

[6] Charles AR Hoare. Quicksort. *The Computer Journal*, 5(1) :10–16, 1962.

[7] Alexandre Hollocou, Thomas Bonald, and Marc Lelarge. Multiple local community detection. In *Performance Evaluation*, 2017.

[8] Ahmed Ould Mohamed Moctar and Idrissa Sarr. Ego-centered community detection in directed and weighted networks. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, pages 1201–1208, New York, NY, USA, 2017. ACM.

[9] Ju Xiang, Tao Hu, Yan Zhang, Ke Hu, Jian-Ming Li, Xiao-Ke Xu, Cui-Cui Liu, and Shi Chen. Local modularity for community detection in complex networks. *Physica A : Statistical Mechanics and its Applications*, 443 :451–459, 2016.

An improved version of lambda architecture

Miguel Landry Foko Sindjoung*, Alain Bertrand Bomgni*, Elie Tagne Fute*,**, Justin Chendjou*

*Department of Mathematics and Computer Science
University of Dschang
Dschang
PO Box 67 Dschang, Cameroon
**Department of Computer Engineering
University of Buea
Buea
PO Box 63 Buea, Cameroon
miguelfoko@gmail.com, alain.bomgni@gmail.com, eliefute@yahoo.fr, chendjoujustin@gmail.com

RÉSUMÉ. La quantité de données produites de façon journalière ne cesse de s'accroitre de nos jours, notamment grâce à l'avènement des objets connectés à internet. Ces objets produisent des quantités de données très importantes (on parle alors de Big Data) et souvent très sensibles qu'il est nécessaire d'analyser en temps réel et/ou sur une période donnée, ceci notament en vue des prises de décisions. C'est dans le but de faciliter l'analyse de données sur ces deux plans que l'architecture Lambda a vue le jour. Cette architecture décrit différentes couches qui peuvent être combinées afin de traiter les données du Big Data. Dans ce document, nous présentons une version améliorée de l'architecture Lambda. Les résultats de notre implémentation montrent une bonne adéquation entre les outils que nous utilisons et le modèle proposé, ce qui fournit des résultats assez encourageants.

ABSTRACT. The amount of data produced on a daily basis is steadily increasing today, especially with the advent of Internet of Things (IoT). These Things produce very large amounts of data (Big Data) and often very sensitive that it is necessary to analyze in real time and/or over a given period, especially for decision making. It is for the purpose of facilitate the analysis of data on both these plans that the Lambda architecture has emerged. This architecture describes different layers that can be combined to process Big Data. In this paper, we present an improved version of the Lambda architecture. The results of our implementation show a good match between the tools we use and the proposed model, which provides quite encouraging results

MOTS-CLÉS : Internet des Objets, Big Data, Architecture Lambda, Traitement Temps Réel, Traitement par lot.

KEYWORDS : Internet of Things, Big Data, Lambda Architecture, Real Time Processing, Batch Processing.

1. Introduction

In recent years there has been a growing production of data due to the advent of the internet of things (IoT) [1]. IoT refers to a set of things (usually sensors) that can produce or capture data and transfer them to the internet network for immediate or subsequent processing. The amount of data produced by things that constitute the IoT is often very huge and evolves exponentially over time. The data produced are so diverse that traditional processing tools and databases are unable to manage them, it is the Big Data. Big Data refers to sets of data that have become so large that they go beyond intuition and the human capacities of analysis and even those of classical processing tools[2, 3]. Given the importance that these data often have (medical application, military, environmental, enterprise, ...), it is important to find mechanisms that allow their treatment in such a way as to reap the full benefit they provide. Some data needs to be processed in real time to immediate decision-making (for example, patient data), while others need to be studied in the long term (eg statistics produced by a company during a given period). It is in order to meet these two constraints that the Lambda architecture has been proposed, that is this architecture is intended to solve the problems of big data in real time and over time. The implementation of the Lambda architecture therefore requires a special knowledge of the appropriate tools for Big Data problem solving. Based on a study of the tools that can intervene in the processing of big data, we propose an improvement of the Lambda architecture which makes it possible to optimize the processing time of Big Data.

2. Related work

This section aims to present a state of the art on the processing of big data. We start by giving the challenges and the characteristics of big data (section 2.1), then, we present the proposed solutions processing of big data in the literature (section 2.2).

2.1. Big Data and its challenges

In this subsection, we present the characteristics of big data and the challenges it faces.

2.1.1. The characteristics of big data

The issue of big data is a hot topic today due to the amount of data that is produced daily. The plurality of data sources, their volume as well as the type of big data information makes it almost impossible to process these data using conventional data processing means. Indeed, IBM's data specialists present the characteristics of big data in a four-dimensional coordinate system [2] whose axes are volume, velocity, variety and veracity. Figure 1 illustrates these coordinates.

– Volume : it is the amount of data available for processing.

– Velocity : helps to measure the speed of generation, the processing and the aggregation of data.

– Variety : refers to different types of generated data (audio, image, videos, ...). These data can be structured or not.

– Veracity : measured or collected data from practical processes must be detected in real time (before any possibility of corruption or manipulation by an external actor).

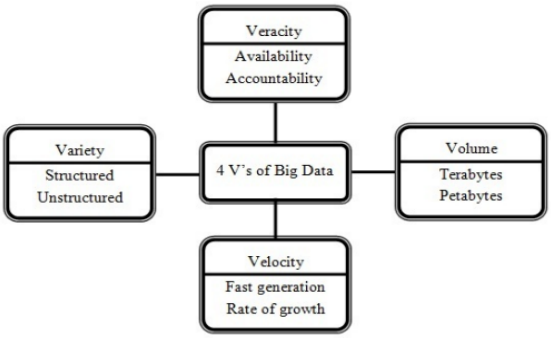


Figure 1 – Big Data's characteristics [1]

2.1.2. Big Data challenges

From the characteristics mentioned in section 2.1.1, several challenges are to be taken up. Indeed, the development of social networks, the multi media, e-commerce, IoT and cloud computing have exploded considerably the volume of data produced [1]. In addition, the need to analyze in real time the data generated by their platforms for the companies renders the traditional processing systems unusable. In this context, new challenges and research problems are encountered [4, 5], among them we have :

– Data management and storage : Because big data uses very large volumes of data that grow exponentially, today's data management systems can not meet the demand because of their limited storage capacity. Moreover, the existing algorithms are not always able to process big data, this because of the heterogeneity of these data. It is therefore interesting to study the possibility of using NoSQL for data backup.

– Data transmission and curation : The amount of data is huge, it is necessary to have an important bandwidth for the transmission of the latter, especially when we know that most of the time they are data from the IoT.

– Data analysis and processing : Response time is an important factor when using big data, especially because the applications that generate these data are mostly very sensitive. In addition, to have real time responses, the processing that is performed on the data must be able to handle very large volumes of data that are inherited. The need to have architectures and tools capable of doing such treatments arises for this purpose.

– Confidentiality and data security : military, medical and many other applications generate confidential data and must be treated with maximum security so that there is no information leak. The majority of data management policies are mostly efficient on static data, but in the context of big data, data varies on a daily basis. Confidentiality and security in the processing of big data is therefore a major objective.

2.2. Current data processing solutions

Data analytics are essential to plan and create decision support systems for optimising the underlying infrastructure. This involves not only processing of the online data, in search for certain events, but also the historical data sources which may be needed to find data patterns which influence decisions. Cloud providers are paramount for the availability and durability to their resources but present various challenges. For instance,

for availability, data is often replicated across multiple servers in different geographical locations, sometimes in untrustworthy locations [6].

Bruns [7] discussed how the current Twitter APIs were extended for third party researchers to deploy their own data analysis on twitter feeds in order to enhance business practices. However unique solutions that allow multiple users of varying backgrounds to write and deploy optimised data processing applications is still needed.

IoT and cloud computing are source of very large volumes of diverse data. Some of the data they produce needs to be analyzed as they arrive (real time processing) while others need to be studied carefully over long period (batch processing). Given the applications from which these data come most often (environment surveillance, monitoring of patients, military applications, online sales companies, ...), it is imperative to find mechanisms that allow not only a real time analysis but also batch processing. It is in the context of performing a data analysis on the two previous plans that $N\ Marz$ et $al.$ [8] proposed the lambda architecture.

The Lambda architecture is a software design pattern that combines both real-time processing and batch processing of big data in a single framework [9]. The figure 2 presents the basic architecture of this design pattern.

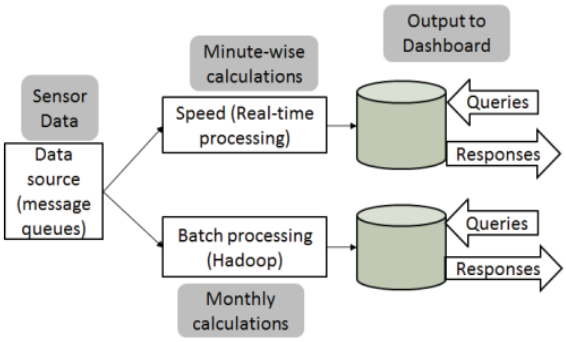


Figure 2 – Basic lambda architecture for real time and batch processing [9]

It caters as three layers (1) Batch processing for pre-computing large amounts of data sets (2) Speed or real time computing to minimize latency by doing real time calculations as the data arrives and (3) a layer to respond to queries, interfacing to query and provide the results of the calculations.

3. An improved version of Lambda architecture

In this section, we present an improved version of the basic Lambda architecture. Indeed, the idea of our solution comes from the fact that the basic architecture does not integrate data ingestion layer, moreover the architecture as presented does not show in a clear way how the old data are obtained for batch processing (it seems that both layers process data in real time). The architecture that we propose is presented in the figure 3.

When data is generated, it is intercepted and ingested by a data-ingestion tool (1). Once the data has been ingested, it is directly made available to a real-time processing tool (2) and at the same time kept in a distributed database for subsequent batch processes (3). During real-time processing, data is regularly processed (4) and the results forming

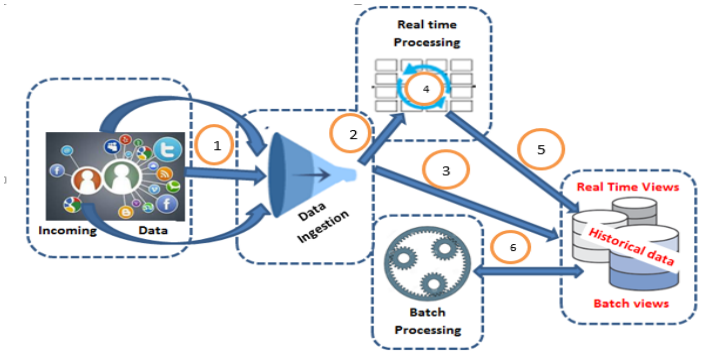


Figure 3 – Our improved lambda architecture

real-time views are stored in a distributed database (5). At considerable time intervals (monthly for example), batch processes are started on the historical data (6) in order to obtain results which will constitute batch views. The batch and real-time views that make up the service layer are therefore merged to answer different user requests.

4. Simulations and results

We present the tools used for our implementation in this section, after which we present our obtained results.

4.1. Big Data Tools used for the simulations

Figure 4 presents the new architecture with tools we used for the implementation. We

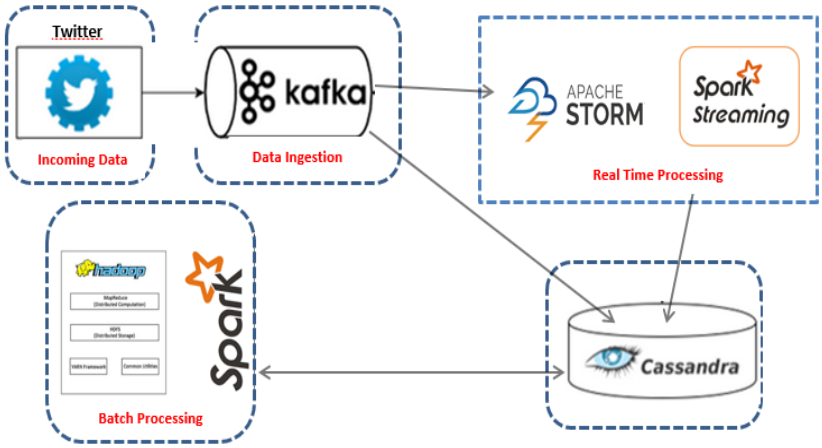


Figure 4 – Our implementation

listen a twitter account as incomming data source in our system, our goal is to count the number of tweets (messages) that arrive in the said account for a given period.

– We use Apache Kafka as data ingestion tool. Kafka[10] is a distributed messaging system that receives and distributes large volumes of data with low latency. It operates according to the producer/consumer model where the data is considered as topics. This means that the producer publishes the topics while the consumer consumes them. The communication between the producer and the consumer is via the HTTP protocol.

– At the batch processing level, we use both Apache Hadoop and Spark, then we make a comparison of results obtained by these two tools. the Hadoop framework[11] with its HDFS, MapReduce and Yarn components enables batch processing. HDFS is a distributed file system that replicates and stores data in cluster machines. MapReduce is a framework for processing and analyzing large volumes of data and Yarn is a framework that aims to separate resource management from the programming model. Although Hadoop is adapted to handle large volumes of data in the context of big data, there are situations where we need the data to remain a little more in memory, in this case, we can think of use of Apache Spark [12] which is a framework to manage large volumes of data just like Hadoop, but with lower latency. It is also important to note that Spark is compatible with the data backup tools used by Hadoop.

– In real time layer, we also make two implementations : one with Apache Storm and another with Spark Streaming and we compare the results. Apache Storm [13] is a popular open source distributed system for processing real-time data. One of the disadvantages of Storm is that it is not able to dynamically optimize between the nodes of the Storm cluster, but that is part of future work in the field. Spark Streaming [14], an extension of Apache Spark is also a distributed system allowing the processing of data in real time. It has a different philosophy than storm. Indeed, in streaming, the received data is stored for a specific time in memory then processed, and returned in Spark RDD (Resilient Distributed DataSet). The disadvantage here is the size of the data to be stored in memory, if it is too short, it can generate multiple RDDs. In addition, in the majority of cases, the data is received through the network, so to ensure the fault tolerance of the data received, Streaming replicates the data through the active nodes.

– Finally, we use Apache Cassandra as distributed database. Apache Cassandra[14] is a distributed storage system for managing very large amounts of structured data spread across the cluster. It provides a highly available, scalable, fault-tolerant, consistent service and is a column-oriented database.

4.2. The obtained results

We make our simulations in a laptop core i3, 4 CPU, 2.4 Ghz ; 8 Go of RAM with Ubuntu 14.04, 64 bits as Operating System. We have in our environment a single node in the Hadoop cluster on which we have a NameNode (Master) and a Datanode (Slave). We also have a single supervisor and a single Nimbus Storm where we have our Storm topology constitute by a Spout and three Bolts. The first bolt makes the split operations on tuples. The second makes the filter operations and the last one makes agregation operations. Our datacenter is a Cassandra cluster constitutes by a node.

Figure 5 shows that Storm processes data faster than Spark. Indeed, Storm processes the set of tweets received (220,000) in 1215 seconds (181 Tweets/second) while Spark processes the same tweets in 2475 seconds (90 Tweets/second). This means that Storm's processing speed is twice that of Spark. This allows us to deduce that Storm is better suited for the real-time processing of big data.

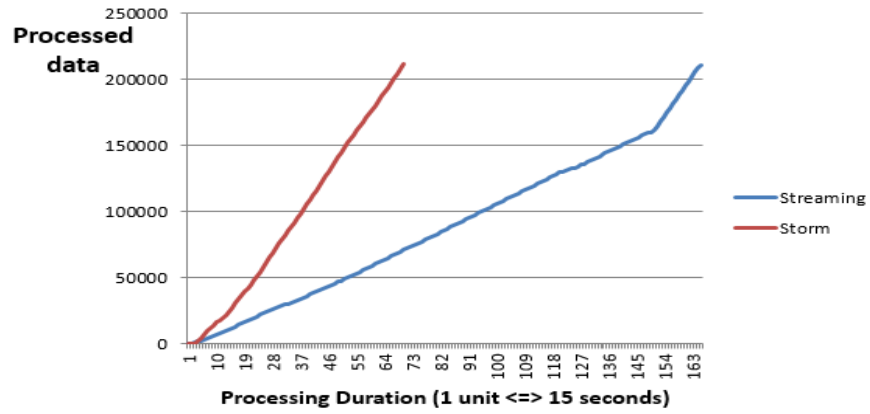


Figure 5 – Performance comparison between Apache Storm and Spark Streaming

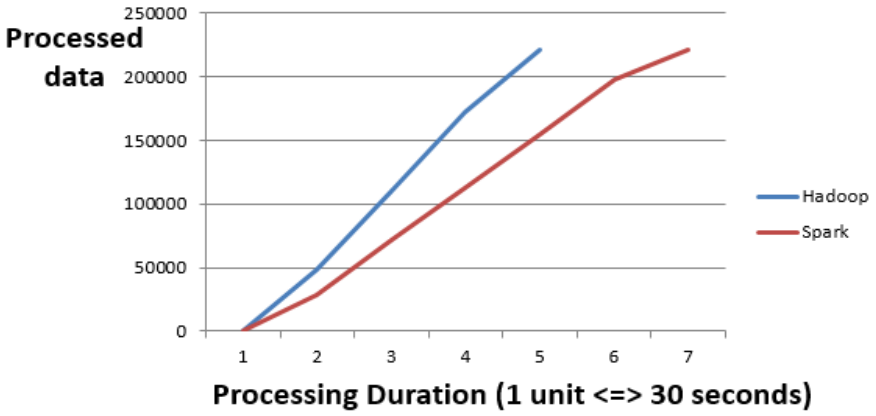


Figure 6 – Performance comparison between Apache Hadoop and Spark

Figure 6 itself presents a comparative curve representing the processed data per unit of time between Hadoop and Spark. On this curve, we notice that Hadoop uses 150 seconds to process the 220,000 tweets present in Cassandra (a speed of 1460 Tweets/s), while Spark takes 210 seconds to process the same amount of data. This allows us to say that Hadoop is faster in batch processing than Spark.

The previous results allow us to conclude that for the implementation of the improved version of the Lambda architecture we present, it is recommended to use Apache Storm as real-time processing tool, and to use Hadoop as batch processing tool.

5. Conclusion and open issues

The lambda architecture [8] is a design pattern that combines real-time processing and batch processing for analyzing big data. Its basic presentation did not include some important aspects for its concrete implementation. In this paper, we are involved in making

a modification on this architecture. Our contribution thus facilitates its implementation. Using tweets from a twitter account as a source of data, we developed an implementation of the new version of lambda architecture, after which we made a comparison between the tools that are used at the real-time and batch layers. The results of our implementation shows that in the implementation of lambda architecture, if we want to have low latency, it is better to use Storm as real-time processing tools and Hadoop as batch processing tools.

Although the results of our implementation are pretty satisfactory, it would be interesting to see the behavior of our implementation when the incoming data is of several varieties and more than the one we used, that is, what will happen if we have 100 000 000 of tweets that arrives per second ? Will the results we obtained be the same ? It might also be interesting to make a comparative study between different data ingestion tools in order to see the real impact of data ingestion in the architecture we proposed. Another perspective would be to ensure that during the data processing by our architecture the security and fault tolerance aspects are taken into account because in the current state this is not the case.

6. Bibliographie

[1] D. P. ACHARJYA, « A survey on big data analytics : challenges, open research issues and tools », *International Journal of Advanced Computer Science and Applications*, vol. 7, n° 2, 2016.

[1] GEORGIOS SKOURLETOPOULOS, CONSTANDINOS X. MAVROMOUSTAKIS, GEORGE MASTORAKIS, JORDI MONGAY BATALLA, CIPRIAN DOBRE, SPYROS PANAGIOTAKIS, EVANGELOS PALLIS, « Big data and cloud computing : A survey of the state-of-the-art and research challenges », *Advances in mobile cloud computing and big data in the 5G Era*, 2017.

[2] SHEN YIN, OKYAY KAYNAK, « Big data for modern Industry : Challenges and Trends », *Proceedings of the IEEE*, vol. 103, n° 2, 2015.

[3] H. HU, Y. WEN, T. -S. CHUA, X. LI, « Towards scalable systems for big data analytics : a technology tutorial », *IEEE Access*, vol. 2, page 652-687, 2014.

[4] H. DEMIRKAN, D. DELEN, « Leveraging the capabilities of service-oriented decision support systems : putting analytics and big data in cloud », *Support Sys*, vol. 55 page 412-421, 2013.

[5] C. L. PHILIP CHEN, CHUN-YANG ZHANG, « Data-intensive applications, challenges, techniques and technologies : a survey on big data », *Information System*, vol. 275 page 314-347, 2014.

[6] M. DIKAIAKOS, G. PALLIS, D. KATSAROS, P. MEHRA, A. VAKALI, « Cloud Computing : Distributed Internet computing for IT and Scientific Research », *IEEE Internet Computing*, 2009.

[7] A. BRUNS, Y. LIANG, L. EUGENE, « Tools and methods for capturing Twitter data during natural disasters. », *First Monday, [S.I]*, 2012.

[8] N. MARZ, J. WARREN, « Big data : principles and the best practices of scalable realtime data systems », *Manning Publications*, 2013.

[9] MARIAM KIRAN, PETER MURPHY, INDER MONGA, JON DUGAN, SARTAJ SINGH BAVEJA, « Lambda architecture for cost-effective batch and speed big data processing », *2015 IEEE International Conference on Big Data (Big Data)*, page 2785-2792, 2015.

[10] , « Apache kafka, Available online », *http://kafka.apache.org/*, , accessed on 21 March 2018.

[11] SHVACHKO K, Kuang H, Radia S, Chansler R., « The hadoop distributedfile system. In : Mass storage systems and technologies (MSST) », *2010 IEEE 26th Symposium on Incline Villiage, Nevada, USA*, page 1-10, May 2010, http://dx.doi.org/10.1109/MSST.2010.5496972.

[12] , « Apache spark, Available online », *https://spark.apache.org/*, , accessed on 21 March 2018.

[13] TOSHNIWAL A., Taneja S., Shukla A., Ramasamy K., Patel JM., Kulkarni S., Jackson J., Gade K., Fu M., Donham J., « Storm@ twitter », *In : Proceedings of the 2014 ACM SIGMOD international conference on management of data. Snowbird, Utah, USA : ACM*, page 147-56, 2014.

[14] MANUEL DIAZ, Christian Martin, Bartolomé Rubio, « State-of-the-art, challenges, and open issues in the integration of internet of things and cloud computing », *Journal of Network and Computer Applications*, page 99-117, 2016.

A parallel pattern-growth algorithm

Parallel pattern-growth

Kenmogne Edith Belise*,a — Nkambou Roger** — Tadmon Calvin* — Engelbert Mephu Nguifo***

* Department of Mathematics and Computer Science, LIFA
University of Dschang, Cameroon
ebkenmogne@gmail.com

** Computer Science Department, Knowledge Management laboratory
University of Québec at Montréal, Canada

*** Computer Science Institute - LIMOS - UMR CNRS 6158
Complexe scientifique des cézeaux, 1 rue de la chebarde, 63178 Aubière cedex, France

RSUM. La recherche des motifs séquentiels est un problème important en fouille de données largement abordée par la communauté de fouille de données, avec un très grand champ d'applications. La recherche des motifs séquentiels vise à extraire un ensemble d'attributs d'un nombre important d'objets collectés dans une base de données. De ce fait, les algorithmes d'extraction des motifs séquentiels sont bien connus pour la consommation à la fois du temps et de la mémoire pour de grandes bases de données. De plus, de nombreuses applications sont critiques en termes de temps d'excécution et impliquent d'énormes volumes de données. De telles applications exigent une puissance d'extraction que les algorithmes séquentiels ne peuvent fournir. Ainsi, il est clairement important d'étudier des algorithmes parallèles. Le travail présenté dans ce papier est orienté vers la conception d'une version parallèle de *prefixSuffixSpan* pour les architectures multi-coeurs en utilisant la méthode de parallélisation PCAM. Nous avons testé notre algorithme parallèle en utilisant plusieurs ensembles de données réelles. Nos expériences ont montré des performances interessantes en termes de vitesse et d"accélération pour presque tous les cas.

ABSTRACT. Sequential pattern mining is an important data mining problem widely addressed by the data mining community, with a very large field of applications. The sequence pattern mining aims at extracting a set of attributes, shared across time among a large number of objects in a given database.Thereby, sequential pattern mining algorithms are well known to be both time and memory consuming for large databases. Moreover many applications are time-critical and involve huge volumes of data. Such applications demand more mining power than serial algorithms can provide. Thus, it is clearly important to study parallel sequential-pattern mining algorithms that take advantage of the computation. The work presented in this paper is directed towards the design of a parallel version of *prefixSuffixSpan* for multi-core architectures using the PCAM parallelization method. We have tested our algorithm using several real-life data sets. Our experiments showed good speedups and accelerations for almost all the cases.

MOTS-CLS : croissance-de-motifs, algorithme parallèle, découverte des motifs séquentiels, vitesse, acceleration

KEYWORDS : pattern-growth, parallel algorithm, sequential pattern discovery, speedup, acceleration

1. Introduction

Sequential pattern mining is a challenging problem since the mining may have to generate or examine a combinatorially explosive number of intermediate subsequences. Thereby, sequential pattern mining algorithms are well known to be both time and memory consuming for large databases. To make sequential pattern mining practical for large data sets, the mining process must be efficient, scalable, and have a short response time. Moreover, since sequential pattern mining requires iterative scans of the sequence dataset with numerous data comparison and analysis operations, it is computationally intensive. Furthermore, many applications are time-critical and involve huge volumes of data. Such applications demand more mining power than serial algorithms can provide. Thus, it is clearly important to study parallel sequential-pattern mining algorithms that take advantage of the computation. Although a significant amount of research results have been reported on serial implementations [18, 9, 8, 6, 13, 22, 3, 14, 16, 17, 7] of sequential pattern mining, there is still much room for improvement in its parallel implementation [20, 21, 15].

The best algorithms for both frequent itemset mining problem and sequential pattern mining problem are based on pattern-growth, a divide-and-conquer algorithm that projects and partitions databases based on the currently identified frequent patterns and grow such patterns to longer ones using the projected databases. We have proven in paper [12, 11] that our sequential pattern-growth algorithm, baptised *prefixSuffixSpan*, outperforms the best previously known sequential pattern-growth algorithm, called *PrefixSpan*. In this paper, we design a parallel version of *prefixSuffixSpan* for multi-core architectures.

The sequel of this paper is organized as follows. Section 2 presents the PCAM parallelization method. Section 3 presents new results. Sub-section 3.1 studies the parallelization of *prefixSuffixSpan*. Sub-section 3.2 designs a multi-core version of the *prefixSuffixSpan* algorithm. Sub-section 3.3 is devoted to the implementation of the multi-core version of *prefixSuffixSpan* and performance analysis. The experimental results show that our parallel algorithm usually achieve interesting speedups. Concluding remarks are stated in section 4.

2. The PCAM parallelization method

In this section, we present the PCAM parallelization method [4]. PCAM stands for Partitioning, Communication, Agglomeration and Mapping. This method organizes the design of a parallel algorithm from a sequential algorithm into four steps. The starting step dealts with the partitioning of the overall computations into tasks. The second step dealts with communications among tasks. The third step studies possible agglomerations of tasks in order to obtain bigger tasks. The fourth step dealts with the mapping, also called allocation, of tasks onto available processors.

The partitioning [19, 1, 2] decomposes the overall computations into either *fine-grain*, *medium-grain* (also called *coarse-grain*) or *large-grain* tasks, depending on the granularity, i.e. size in term of computations, of tasks. A *fine-grain* task [1] consists of a constant number basic operations. A *medium-grain* task [5] consists of a linear number of basic operations. In many sequential algorithms, it is on the form of a depth-one loop whose body computes a constant number of basic operations. A *large-grain* task [5] consists of a large number of basic operations. In many sequential algorithms, it is on the form

of a loop of depth greater than one whose body computes a constant number of basic operations.

The study of communications involves the identification of data to be transferred between tasks as well as the definition of related data structures and of reliable communication protocols for data exchanges beetween tasks. A classic challenging problem is to design communication protocols that optimize communication costs [2]. A non-adequate communication protocol may significantly slows down the execution of the corresponding parallel algorithm. Because of this, the commnication protocol should fit with the allocation of tasks to processors.

The study of agglomerations leads to a medium-grain decomposition from a fine-grain decomposition and to a large-grain decomposition from a medium-grain decomposition. Although agglomerations of large-grain tasks lead to bigger tasks, the granularity of the new decomposition obtained remains unchanged. By gathering tasks, the number of data transfers between them are reduced. Thus, agglomerations contribute significantly to the optimization of communication costs [5]

The mapping consists in assigning tasks obtained from agglomerations to processors so as to minimize communications costs and the sum of idles times of all the procesors used in the parallel algorithm [1, 2, 5]

3. New results

3.1. Parallelizing prefixSuffixSpan

3.1.1. Partitioning prefixSuffixSpan and studying communications therein

In this section, we design a multi-core version of prefixSuffixSpan. This is done following the PCAM parallelization method [4]. At the first glance, prefixSuffixSpan can be decomposed into *projection tasks*. The unique level-one *projection task* takes as input the global dataset and a pattern-growth direction [10], mines frequent items and generates one level-one projected dataset per frequent item. Each non-empty level-one dataset leads to a level-two projection task which takes as input a frequent item, a level-one projected dataset and a pattern-growth direction, and generates length-two sequential patterns and one level-two projected dataset per length-two pattern generated. Each non-empty level-two dataset, in turn, leads to a level-three projection task which takes as input a length-two sequential pattern of the form $\alpha.\alpha\prime$, a level-two projected dataset and a pattern-growth direction, and generates length-three sequential patterns by making grow either prefix α or suffix $\alpha\prime$ and one level-three projected dataset per length-three pattern generated.

More generaly, by considering that the global dataset is of level zero , a level-k projection task takes as input (1) a length-k sequential pattern of the form $\alpha.\alpha\prime$, (2) a level-(k-1) projection dataset, and (3) a pattern-growth direction. If the pattern-growth direction is *left-to-right* (resp. *right-to-left*) it makes grow prefix α (resp. suffix $\alpha\prime$). It generates length-(k+1) sequential patterns and one level-(k+1) projected dataset per generated pattern. In this partitioning, the only task to be executed at the beginning is the level-one task. Because of this, only one thread can work at the beginning while the others threads are waiting for the end of the execution of the level-one task. Thus, this first partitioning is not suitable in minimizing idle times of threads involved in the parallel execution of prefixSuffixSpan. As a consequence, it can be improved.

3.1.2. Partitioning the level-one task and studying communications and synchronization therein

The level-one projection task should be partitioned into a number of parallel smaller tasks, i.e. tasks that could be executed simultaneously, in order to allow all thread to get a task to execute at the beginning. This is done here in eight steps following partitioning techniques developped in [1, 2]. The number of tasks of each step from step 2 to step 7 is equal to the number of threads involved in the parallel execution of prefixSuffixSpan. These steps are described here as follows :

1) *Step 1 :* The global dataset is partitioned into as many partial data sets as there are threads devoted to the parallel execution of prefixSuffixSpan.

2) *Step 2 :* Each thread gets a partial dataset and computes the partial supports of items therein in order to obtain partial supports.

3) *Step 3 :* Partial supports are used to update global supports The update is done by the thread who has computed the partial supports. Partial supports should be stored in a concurrent data structure because of concurrent write operations involving global supports and arisen from many threads. A synchronization barrier is needed here because the next step should begin after the end of this one. It can be done by using a concurrent integer data to count the number of threads who have update the global supports. Such an integer is initialized to zero and incremented after each update of global supports.

4) *Step 4 :* Each thread gets the global supports per item and a partial list of items, then seeks for frequent items in its list of items in order to obtain a partial list of frequent items.

5) *Step 5 :* Partial lists of frequent items are used to update the global list of frequent items. The update is done by the thread who has constructed the partial list. Partial lists should be stored in a concurrent data structure because of concurrent write operations involving the global list and arisen from many threads. A synchronization barrier is needed here because the next step should begin after the end of this one. It can be done by using a concurrent integer data to count the number of threads who have update the global list of frequent items. Such an integer is initialized to zero and incremented after each update of the global list of frequent items.

6) *Step 6 :* Each thread gets the global list of frequent items and computes the left and right weights of the sequences of its dataset received at step 2.

7) *Step 7 :* Partial left (resp. right) weights are used to update the global left (resp. right) weight assuming that it is initialized to zero. The global left and right weights are used to determine the promising pattern-growth direction. The update is done by the thread who has calculated the partial weights. Synchronization issues arisen here are solved as in steps 3 and 5.

8) *Step 8 :* Tasks of this step represent the new level-one projection tasks. Each frequent item leads to such a task.

3.1.3. An improved Partitioning of prefixSuffixSpan

The main weakness of the first partitioning is overcome here by replacing the level-one task with its decomposition into smaller (in term of the amount of computations) tasks. A new partitioning is obtained by replacing level-one task with its decomposition. This leads to an improved partitioning of *prefixSuffixSpan*. It reduces the idle times of threads compared to the previous partitioning.

3.1.4. Issues related to the improved partitioning

Agglomerations : We use an integer value called *depth* which indicates the projection-task level from which agglomerations should be constructed. If the depth value is d, agglomerations are constructed only from the projection tasks of level greater than $d-1$. An agglomeration is obtained by gathering a level-d task with all its descendents. As a consequence, once a thread retrieves a level-d projection task from the pool of projection tasks, it executes that task with all its descendents. The resulting partitioning is a mixture of medium-grain and large-grain tasks. Large-grain tasks permit to reduce the synchronization costs arisen from the handling of the pool of projection tasks.

The concurrent pool of projection tasks : A concurrent pool of tasks is used to handle the storage and retrieval of projection tasks. Once a thread generates a projection task of level lower than the value of *depth*, it saves that task in the pool if the pool is not full. Otherwise, it should execute that generated projection task. Idle threads retrieve projection tasks to execute from the pool. This pool reduces idle times of threads by providing tasks to idle threads.

Mapping : The mapping of tasks onto threads is unknown before the beginning of the parallel execution of *prefixSuffixSpan*. Tasks are assigned to threads during the parallel execution. Because of this, the mapping is dynamic. As mentioned above, idle threads retrieve tasks to execute from the concurrent pool of projection tasks. This contributes to load balancing calculations.

Communications : Communications between threads are performed through four concurrent data structures. As each data structure is a critical resource, it can not be used by two threads simultaneously. The costs [5] of the handling of synchronization related to a concurrent data structure increases with the number of threads needing to access that data structure. This may cause a slow down of the acceleration of the parallel algorithm when the number of threads increases.

Termination criterion of the multi-core algorithm : A concurrent array called *busy* is used. Cell $busy[i]$ contains 1 if the thread numbered i has gotten a projection task from the concurrent pool of projection tasks during its last attempt and 0 otherwise. If all the cell of array *busy* contain 0, it means that no thread has a projection task to execute. When this condition is satisfied, the multi-core algorithm ends.

3.2. A multi-core version prefixSuffixSpan

In this section, we translate the results of section 3.1 into a multi-core version of *prefixSuffixSpan*. Here is the list of functions executed by all thread involved in the multi-core execution of *prefixSuffixSpan*.

1) Function THREADTASKFORSUPPORTCOUNT is a translation of steps 2 and 3 into an algorithm. It is executed by a thread to (1) compute the partial supports per item of its partial dataset, (2) update the global supports per item with its partial supports, (3) wait for all the updates of global supports, and (4) get the global supports.

2) Function THREADTASKTOFINDFREQUENTITEM is a translation of steps 4 and 5 into an algorithm. It is executed by a thread to (1) construct its partial list of frequent items from its partial list of items, (2) update the global list of frequent items with its

partial list of frequent items, (3) wait for all the updates of global list of frequent items, and (4) get the global list of frequent items.

3) Function THREADTASKTOGETGROWTHDIRECTION is a translation into an algorithm. It is executed by a thread to (1) compute its partial left and right weights of its partial dataset, (2) update the global left and right weights with its partial left and right weights, (3) wait for all the updates of global left and right weights, (4) get the global weights, and (5) determine the pattern-growth direction from global weights.

4) Function PROJECTIONTASK is a translation of the description of projection tasks into an algorithm. It is used by a thread to execute a projection task.

5) Function MAINTHREADTASK is the starting point of the execution of all thread involved in the multi-core execution of *prefixSuffixSpan*. The others functions are called in this one.

3.3. Implementation and performance analysis

The data sets used here are collected from the webpage (http ://www.philippe-fournier-viger. com/spmf/index.php) of SPMF software. This webpage provides large data sets in SPMF format that are often used in the data mining litterature for evaluating and comparing algorithm performance. All experiments are done on a 32-cores. All the algorithms are implemented in Java and grounded on SPMF software [17]. The experiments consisted of running the multi-core version of *prefixSuffixSpan* on each data set and for a given number of threads ranging from two to thirty two while decreasing the support threshold until algorithms became too long to execute or ran out of memory. We also studied the influence of the depth's value on the algorithm's performance when the number of threads is thirty two. For each execution, we recorded the execution times, the speed up and the accelerations. The speed up of a parallel execution is defined as follows.

$$\text{Speed up for n threads} = \frac{\text{execution time for one thread}}{\text{execution time for n threads}}$$

The acceleartion of a parallel execution is defined as follows.

$$\text{Acceleration for n threads} = \frac{\text{speed up for n threads}}{\text{n}}$$

The speed up is upperly bounded by the number of threads while the acceleration is upperly bounded by 1. In the following, we analyze the performance of our multi-core algorithm per data set. The experimentations show that the speed up may increase (1) as the number of threads increases, (2) as the depth increases, (3) as the support threshold decreases, and (4) as the number of sequential patterns increases. They also show that the speed up may be very sensitive to the change of depth. The acceleration of our parallel algorithm on four real-life data sets is within range $[0.58\ 1]$ for minimum support thresholds and thirty two threads. In [20], a parallel version of the well known *PrefixSpan* algorithm is proposed. The acceleration of that parallel algorithm on five synthetic data sets is within range $[0.25\ 0.5]$ [20] for minimum support thresholds and thirty two processorss. However, a divide-and-conquer property, though minimizing inter-processor communication, causes load balancing problems, which restricts the scalability of parallelization.In [20], synthetic data sets show better speed up than real ones. This is because synthetic data sets have more frequent items and, after the large projected databases are partitioned, the sub-databases derived are of similar size. However, in real data sets, the number of frequent items is small and even when the large tasks are partitioned into smaller subtasks, the size

of the subtasks may still be larger, or even much larger. It is clear that the performance of our parallel algorithm is better compared to the performance of the parallel version of PrefixSpan proposed in [20].

4. Conclusion

In this paper, we have proposed a parallel implementation of the *prefixSuffixSpan* mining algorithm. This parallel version of *prefixSuffixSpan* is obtained in four main steps : (1) partitioning of *prefixSuffixSpan* into tasks following the PCAM parallelization method, (2) studying issues related to the partitioning, namely (2.1) agglomerations, (2.2) the concurrent pool of projection tasks, (2.3) mapping, (2.4) communications and synchronization, and (2.5) the termination criterion, (3) translating tasks into algorithms, and (4) implementing algorithms.

We have tested our algorithm using several real-life data sets. Our experiments showed good speedups and accelerations for almost all the cases. These results outperform the best previous ones [20].

5. Bibliographie

[1] CLÉMENTIN TAYOU DJAMEGNI , PATRICE QUINTON , SANJAY V. RAJOPADHYE , TANGUY RISSET , MAURICE TCHUENTE, « A reindexing based approach towards mapping of DAG with affine schedules onto parallel embedded systems », *J. Parallel Distrib. Comput.*, vol. 69, n° 1, 1–11, 2009.

[2] CLÉMENTIN TAYOU DJAMEGNI , MAURICE TCHUENTE, « A Cost-Optimal Pipeline Algorithm for Permutation Generation in Lexicographic Order », *J. Parallel Distrib. Comput.*, vol. 44, n° 2, 153–159, 1997.

[3] CHIA-YING HSIEH , DON-LIN YANG , JUNGPIN WU, « An Efficient Sequential Pattern Mining Algorithm Based on the 2-Sequence Matrix », *Workshops Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, 583–591, 2008.

[4] IAN T. FOSTER, « Designing and building parallel programs - concepts and tools for parallel software engineering », *Addison-Wesley*, 1995.

[5] JEAN FRANÇOIS DJOUFAK KENGUE , PETKO VALTCHEV , CLÉMENTIN TAYOU DJAMEGNI, « Parallel Computation of Closed Itemsets and Implication Rule Bases », *Parallel and Distributed Processing and Applications, 5th International Symposium, ISPA 2007, Niagara Falls, Canada, August 29-31, 2007, Proceedings*, 359–370, 2007.

[6] JIAN PEI , JIAWEI HAN , BEHZAD MORTAZAVI-ASL , JIANYONG WANG , HELEN PINTO , QIMING CHEN , UMESHWAR DAYAL , MEICHUN HSU, « Mining Sequential Patterns by Pattern-Growth : The PrefixSpan Approach », *IEEE Trans. Knowl. Data Eng.*, vol. 16, n° 11, 1424–1440, 2004.

[7] JIAWEI HAN , MICHELINE KAMBER, « Data Mining : Concepts and Techniques », *Morgan Kaufmann*, 2000.

[8] JIAWEI HAN , JIAN PEI , YIWEN YIN, « Mining Frequent Patterns without Candidate Generation », *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA.*, 1–12, 2000.

[9] KARAM GOUDA , MOSAB HASSAAN , MOHAMMED J. ZAKI, « Prism : An effective approach for frequent sequence mining via prime-block encoding », *J. Comput. Syst. Sci.*, vol. 276, n°

1, 88–102 2010.

[10] KENMOGNE EDITH BELISE, « The Impact of the Pattern-Growth Ordering on the Performances of Pattern Growth-Based Sequential Pattern Mining Algorithms », *Computer and Information Science*, vol. 10, n° 1, 23–33 2017.

[11] KENMOGNE EDITH BELISE , TADMON CALVIN , ROGER NKAMBOU, « A pattern growth-based sequential pattern mining algorithm called prefixSuffixSpan », *EAI Endorsed Trans. Scalable Information Systemsurnal*, vol. 4, n° 12, e4, 2017.

[12] KENMOGNE EDITH BELISE, « Contribution to the sequential and parallel discovery of sequential patterns with an application to the design of e-learning recommenders », *PhD Thesis. The University of Dschang, Faculty of Sciences, Department of Mathematics and Computer Science*, waiting for defense.

[13] LIONEL SAVARY , KARINE ZEITOUNI, « Indexed Bit Map (IBM) for Mining Frequent Sequences », *Knowledge Discovery in Databases : PKDD 2005, 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, Porto, Portugal, October 3-7, 2005, Proceedings*, 659–666, 2005.

[14] MOHAMMED JAVEED ZAKI, « TSPADE : An Efficient Algorithm for Mining Frequent Sequences », *Machine Learning*, vol. 42, n° 1/2, 31–60 2001.

[15] MOHAMMED JAVEED ZAKI, « Parallel Sequence Mining on Shared-Memory Machines, *J. Parallel Distrib. Comput.*, vol. 61, n° 3, 401–426 2001.

[16] NIZAR R. MABROUKEH , CHRISTIE I. EZEIFE, « A taxonomy of sequential pattern mining algorithms », *ACM Comput. Surv.*, vol. 43, n° 1, 3 2010.

[17] PHILIPPE FOURNIER-VIGER , ANTONIO GOMARIZ , TED GUENICHE , AZADEH SOLTANI , CHENG-WEI WU , VINCENT S. TSENG, « SPMF : a Java open-source pattern mining library », *Journal of Machine Learning Research*, vol. 15, n° 1, 3389–3393 2014.

[18] RAKESH AGRAWAL , RAMAKRISHNAN SRIKANT, « Proceedings of the Eleventh International Conference on Data Engineering, March 6-10, 1995, Taipei, Taiwan », *Mining Sequential Patterns*, 3–14, 1995.

[19] SABEUR ARIDHI , LAURENT D'ORAZIO , MONDHER MADDOURI , ENGELBERT MEPHU NGUIFO, « Density-based data partitioning strategy to approximate large-scale subgraph mining », *Inf. Syst.*, vol. 48, 213–223 2015.

[20] SHENGNAN CONG , JIAWEI HAN , JAY HOEFLINGER , DAVID A. PADUA, « A sampling-based framework for parallel data mining », *Proceedings of the ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, PPOPP 2005, June 15-17, 2005, Chicago, IL, USA*, 255–265, 2005.

[21] VALERIE GURALNIK , GEORGE KARYPIS, « Parallel tree-projection-based sequence mining algorithms », *Parallel Computing*, vol. 30, n° 4, 443–472 2004.

[22] ZHENGLU YANG , YITONG WANG , MASARU KITSUREGAWA, « LAPIN : Effective Sequential Pattern Mining Algorithms by Last Position Induction for Dense Databases », *Advances in Databases : Concepts, Systems and Applications, 12th International Conference on Database Systems for Advanced Applications, DASFAA 2007, Bangkok, Thailand, April 9-12, 2007, Proceedings*, 1020–1023 2007.

6. Annex

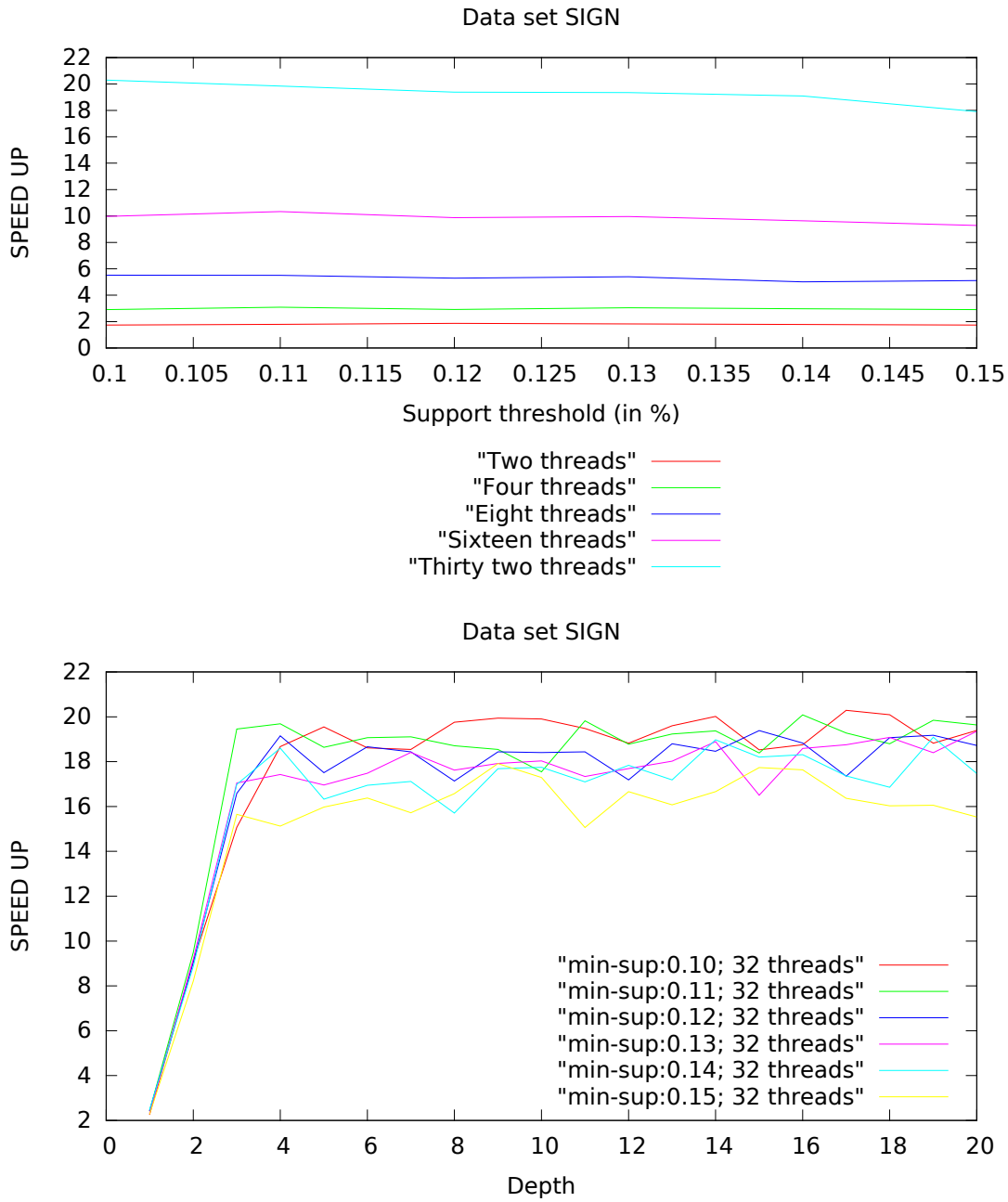


Figure 1. *Performances of the multi-core version of prefixSuffixSpan on the real-life data set SIGN. The speed up increases (1) as the number of threads increases, (2) as the depth increases in general up to* 4*, (3) slightly as the support threshold decreases. In the first part of the figure, the speed up is quite stable and the acceleration for the minimum support threshold is within range* $[0.63\ 0.86]$*.*

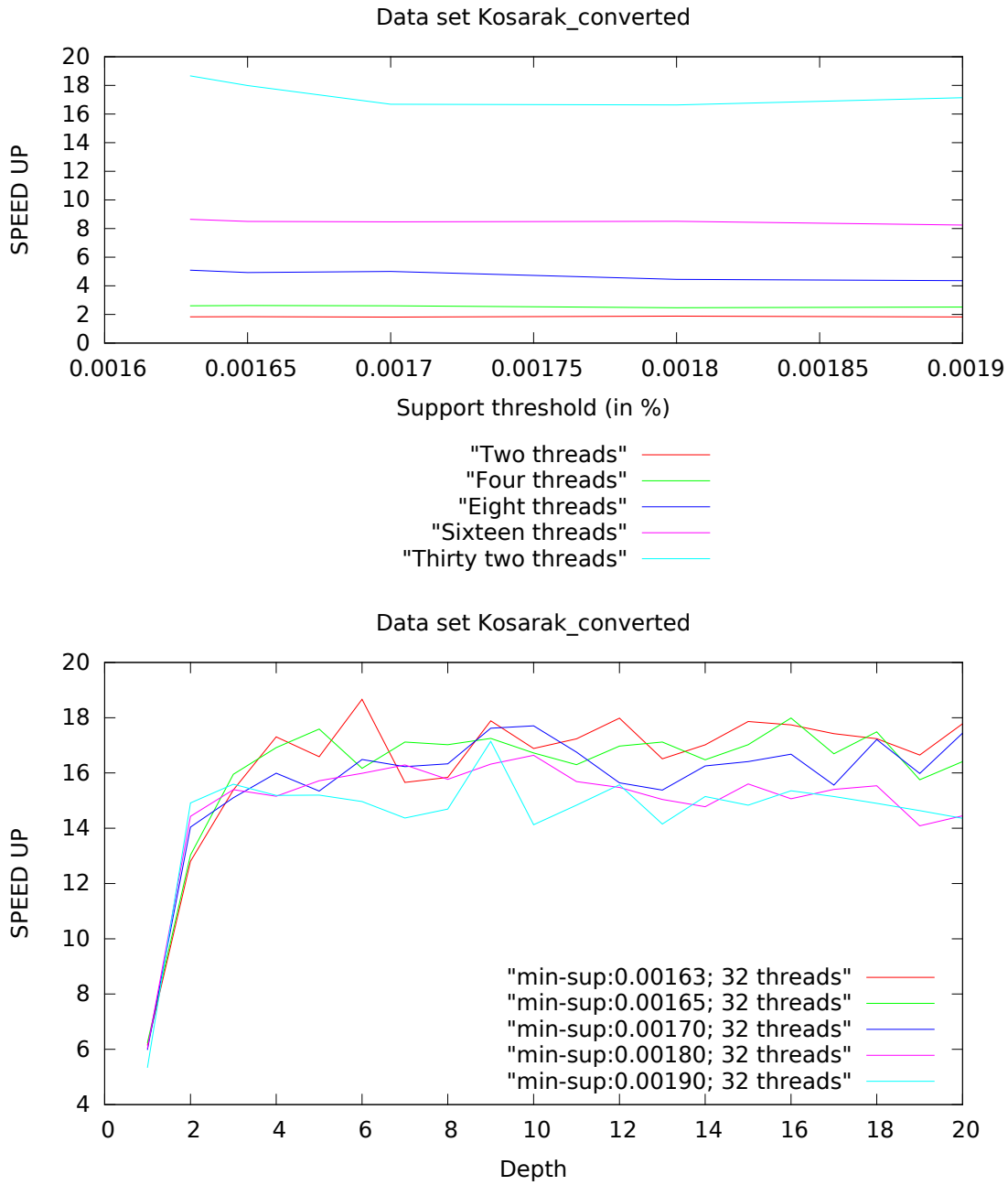


Figure 2. *Performances of the multi-core version of prefixSuffixSpan on the real-life data set Kosarak_converted. The speed up increases (1) as the number of threads increases, (2) as the depth increases in general, (3) as the support threshold decreases in general for thirty two threads. In the first part of the figure, the speed up is quite stable and the acceleration for the minimum support threshold is within range* $[0.58\ 0.91]$.

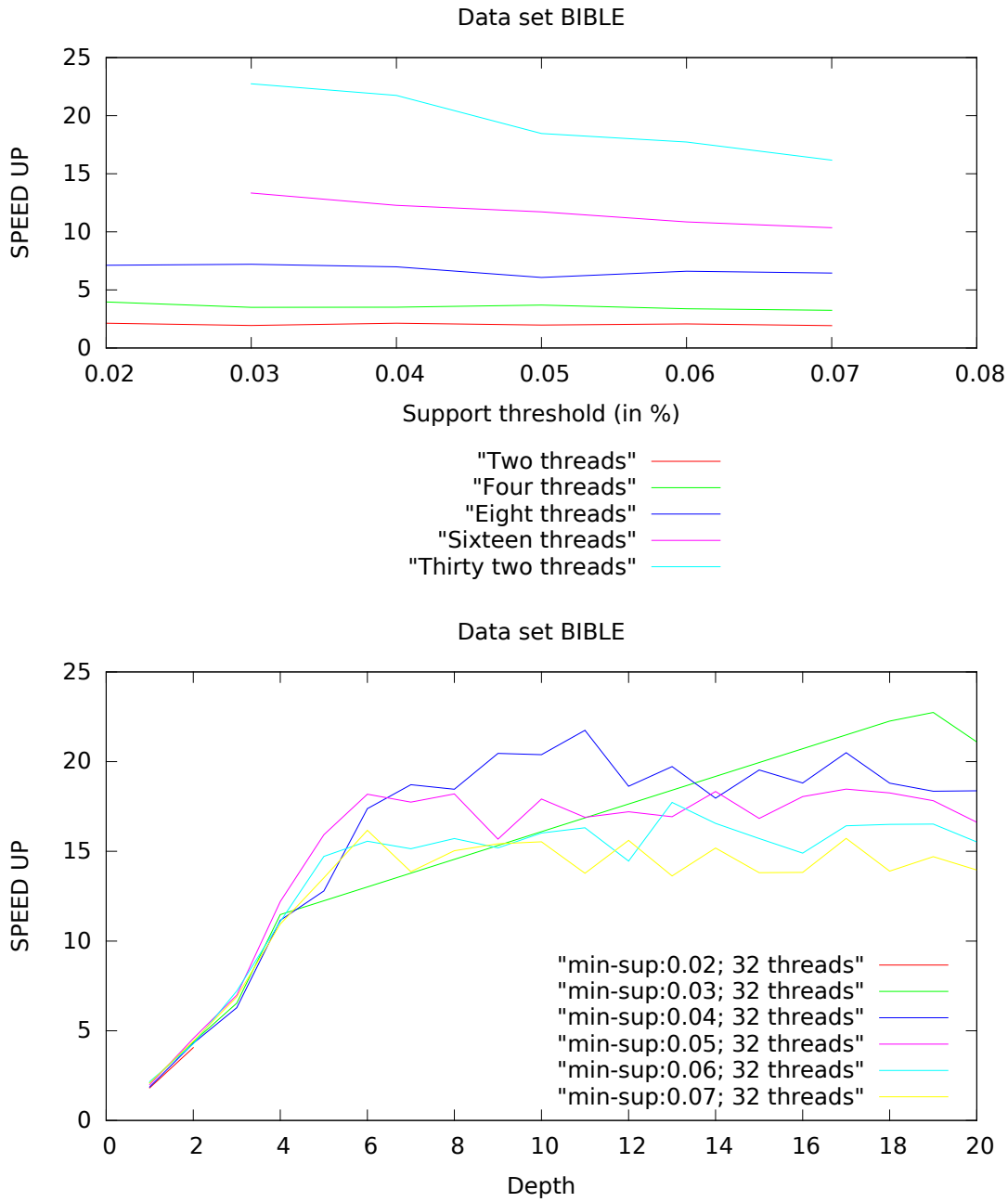


Figure 3. *Performances of the multi-core version of prefixSuffixSpan on the real-life data set BIBLE. The speed up increases (1) as the number of threads increases, (2) as the depth increases in general, (3) slightly as the support threshold decreases. In the first part of the figure, the speed up is relatively stable and the acceleration for the minimum support threshold is within range* $[0.71\ 1]$.

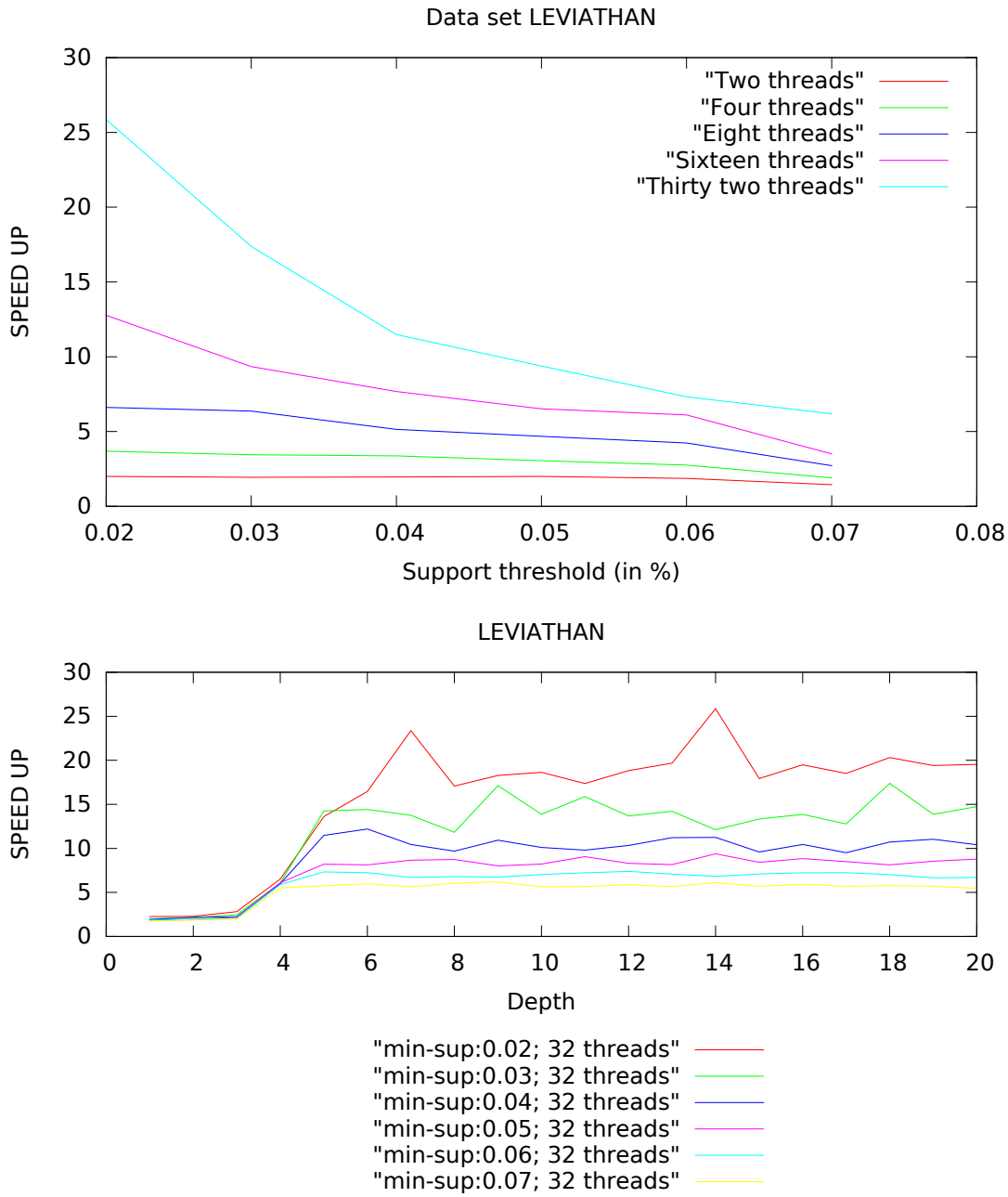


Figure 4. *Performances of the multi-core version of prefixSuffixSpan on the real-life data set LEVIATHAN. The speed up increases (1) as the number of threads increases, (2) as the depth increases in general up to 6, (3) as the support threshold decreases. In the first part of the figure, the speed up decreases significantly as the support threshold decreases when the number of threads is sixteen or twenty two, and the acceleration for the minimum support threshold is within range* $[0.80\ 1]$.

Communities in directed networks

Towards a hybrid model of semantic communities detection

Gamgne Domgue Félicité, Tsopze Norbert, Ndoundam René, Arnaud S. R. M. Ahouandjinou

IRD UMI 209 UMMISCO,
University of Yaounde I
Yaounde,Cameroon
felice.gamgne@gmail.com, fgamgne@uy1.uninet.cm
tsopze@uy1.uninet.cm
ndoundam@gmail.com
arnaud.ahouandjinou@univ-littoral.fr, ahou.arn@gmail.com
Laboratoire LISIC, Université du Littoral de la Cote d'Opale(ULCO),62228 Calais, France,

ABSTRACT. Community detection in directed network is of vital importance to find cohesive sub-groups. Many existing graph clustering methods mainly focus on the relational structure and vertex properties, but ignore edge directionality during the clustering task, in case of directed graphs. In this paper we propose a hybrid semantic similarity which includes node attribute informations along with the network structure and link semantic. Then by application of a partitioning clustering technique, we evaluate its performance and results on a built textual based dataset with ground truth. We argue that, depending on the kind of data we have and the type of results we want, the choice of the clustering method is important and we present some concrete examples for underlining this.

RÉSUMÉ. La détection des clusters orientés constitue davantage un challenge dans l'analyse des réseaux. Plusieurs approches de clustering s'attardent uniquement sur la structure et les attributs des noeuds, mais ignorent la sémantique portée par les liens dans le cas des graphes orientés. Dans cet article nous proposons une mesure de similarité hybride qui combine les informations structurelles, les attributs et l'orientation des liens. Par application de cette mesure à un algorithme de clustering , nous évaluons les performances de cette nouvelle approche sur un jeu de données que nous avons construit avec vérité de terrain. Selon le type de données exploité et le type de résultats escomptés, nous montrons que le choix de la méthode de classification est important via quelques illustrations.

KEYWORDS : Directed attributed network, Graph clustering, Link semantic, Social network

MOTS-CLÉS : Réseau orienté attribué, Clustering, Sémantique des liens, Réseau social

1. Introduction

Cluster extraction is one of the main tasks of descriptive modelisation in datamining area. Like this, most of graph partitioning methods, useful for strongly connected community detection [7], focus on relational structure, but ignore node properties or attributes. More the recent approaches tended to find cohesive subgroups by combining node attributes with link informations in graph. These informations only concerned the structure data like frequent link-pattern(neighbourhood and leadership). Nevertheless, combining these different data types leads to the problem of semantic classification, because of the "inconsistent" similarity measures omitting the link *semantic* (meaning edge's directionality). A new challenge in community detection consists on meaningful cluster extraction based on three parameters : structure, node attributes and link semantic. In this paper, we propose an hybrid technique dealing with the *semantic based topological* structure of the graph, and we show that with textual attributes joined to vertices, it is possible to extract semantic clusters. We perform our experiments through the construction of an attributed directed network with ground truth, Normalized Mutual Information (NMI) and Density measures are used for evaluations. The work of incorporating structural *semantic* and attribute data has not yet been throughout studied in the context of large social graphs. This is the motivation of our work for which key contributions are summarized next : studying of the relationship between semantic similarity of species in a food web network and showing that the type of data determine the result, thus a textual attribute strengthens the semantic topology and helps to discover more relevant communities.

The document is organized as follows. The Section 2 presents related works based on graphs partitioning methods that take into account both features and structure relationship. The formal description of the idea is presented in Section 3, then some hybrid approaches based on both links and attribute information are suggested in Section 4. An experimental study describing the constructed dataset and the expected results according to the technique are presented in the section 5. After that experiment description, an evaluation on different semi-hybrid and hybrid models are shown in the Section 6, and the Section 7 concludes the study.

2. Related works

The well-known graph clustering techniques use the relationships between vertices to partition the graph into several densely connected components, but do not use the properties of the nodes. The problem is to combine both graph data and attribute data simultaneously in order to detect clusters that are densely connected and similar in the attribute space. Few recent studies have addressed the problem of clustering in attributed networks. Next, we present a classification of the existing methods of clustering in attributed graph based on their methodological principles.

Edge weighting based approaches : In order to integrate the attribute or structure information in the clustering process, these methods define a node attribute similarity that will be used to weight the existing edges. In literature, some relevant approaches have been proposed [1]. The first approach of the following section is based on this idea.

Pattern-based approaches : These methods focus on the structure or relational property of the graph, based on kernels information Li et al. [2]. In the same way, Gamgne et al. [8] extracted kernels through the neighbourhood overlap. The relationship information

is based on either the structural equivalence i.e. two vertices belong to the same cluster if they own the same neighbours or leadership i.e. vertices are connected to the same leader. They defined a *kernel degree* measure which denotes the similarity of nodes in their roles of leader (high in-degree) or follower (low in-degree) as studied by Gamgne et al. [9]. Its limit is that it does not deal with node attributes.

Quality function optimization based approaches : This family of approaches extend the well-know graph based clustering methods to consider both attribute information and topological structure. Authors in [6] proposed an extension of the Louvain algorithm with a modification of modularity by including an attribute similarity metric. [5] propose the **I-Louvain** algorithm which uses the inertia based modularity combined with the Newman's modularity.

Unified distance based approaches : They consist in transforming the topological information of the network into a similarity or a distance function between vertices. Zhou et al. [4] exploit the attributes in order to extend the original graph to an augmented one. A graph partitioning is then carried out on this new augmented graph. A neighborhood random walk model is used to measure the node closeness on the augmented graph. Then, they proposed a **SA-Cluster** algorithm that make use of a random walk distance measure and K-Medoids approach for the measurement of a node's closeness.

All of these methods have the limit that their topological property does not deal with link semantic, meaning edge directionality in directed networks. Yet the majority of real-life networks are represented as directed graphs, and link direction helps in improving partition quality.

We present in the Section 4, methods handling both topological and node attributes and that are easy to use, while the next section shows how formally a generic clustering approach could be implemented.

3. Problem Statement

An attributed graph is denoted as $G = (V, E, W)$, where V is the set of nodes, E is set of edges, and W is the set of attributes associated to the nodes in V for describing their features. Each vertex v_i is described by a real attribute vector $d_i = (w_1(v_i), ..., w_j(v_i), ..., w_m(v_i))$ where $w_j(v_i)$ is the attribute value of vertex v_i on attribute w_j. Into such network, clustering of attributed graph should take into account both structure network and attribute information by achieving a good balance between the following two properties : *(i)* vertices within one cluster are closed to each other in terms of "structure", meaning that vertices are arranged according to a semantic pattern, while vertices between clusters are not patterned; *(ii)* vertices within one cluster are more similar by their attributes than vertices from different clusters that could have quite different attribute values. In this work, we consider that the partitioning process focuses both on *semantic based topology* and node attributes. In others words, the structure concept includes not only link density, but also link semantic. The approach consists in dividing the set of nodes V into a partition of k clusters C_i, such that :

1) $C_i \cap C_j \neq \Phi \; \forall i \neq j$ and $\cup_i C_i = |V|$, where Φ is an empty set,

2) The semantic similarity takes into account three criteria : the link density, the node attribute and the link direction,

3) Vertices within clusters are semantically connected, while the vertices in different clusters are sparsely connected.

Likewise, we assume that an information network like a food web network can be represented by an attributed directed graph. Then, species relationship corresponds to a network in which each vertex represents a species and is described by a vector $d_i = (w_{i1}, w_{i2})$ where w_{i1} is the discrete attribute according to the diet mode (0 for "$carnivorous$" and 1 for "$herbivorous$") and w_{i2} the textual attribute denoting mode of reproduction (either " $oviparous$" or "$viviparous$") ; an edge from node a to node b means that species a is consumed by species b (*"Prey-Predator"* relationship). Thus, partitioning this kind of graph leads to integrate both ($density$ and $semantic$) topological and ($discrete$ or $textual$) attribute knowledge.

4. Clustering Graph models

Approaches for graph clustering described in this section separately handle both relational information and vertex attributes, and differ by their manner of combining relational data and attributes.

4.1. Attribute and Relational based clustering methods

Attribute based clustering method first exploits attributes by graph enrichment through a node attribute similarity (NAS) function [1,4,6]. According to the **SA-Cluster** method [4], the unified random walk distance is applied to an augmented graph. On the other hand, cosine distance between vertices v_i and v_j could be used, as defined as $SimA(v_i, v_j)$ in **SAC1** method [6].

In the relational based clustering model, structural properties are considered first through either a neighbourhood similarity. Li in [2] proposed a hierarchical clustering by filtering process of cores (kernels) based on structural information, then merging them by their attributes similarity. The core filtering is based on a frequent itemsets process through a similarity we labelled here $simS(v_i, v_j)$; it could be based on geodesic distance [7]. Formally, $simS(v_i, v_j) = \frac{1}{1+disS(v_i,v_j)}$. See Sect.4.2 below.

4.2. Semi Hybrid clustering

Semi-hybrid techniques combine simultaneously structural and attribute similarities through a weighted function as in Eq.1. **W-Cluster** and Combe's Model [3] are typical instances of this technique.

$$disG(v_i, v_j) = \alpha disT(v_i, v_j) + \beta disS(v_i, v_j) \tag{1}$$

$disT$ and $disS$ denote euclidean distance for attribute data and geodesic distance for structure data respectively. A straightforward way to integrate link semantic is to combine relational, attribute and semantic similarities by adding another factor to the Eq.1 as described below.

4.3. Proposed Hybrid Clustering Model

To avoid confusion to that semi-hybrid method (not taking into account link direction), we add semantic property based on edge directionality named $simR(v_i, v_j)$ [8] and we call *semantic clusters* the groups detected from a directed attributed graph partitioning hybrid model. The proposed approach combines simultaneously 3 information data through a Node Attribute and Edge Directionality Similarity ($NAEDS$) as defined in Eq.2. Then,

we have applied $NAEDS$ in Louvain's method to find answer of the following question: *Whether semantic communities be detected by dealing with direction of the edges?*

$$simG(v_i, v_j) = \alpha simT(v_i, v_j) + \beta simS(v_i, v_j) + \gamma simR(v_i, v_j) \quad (2)$$

The equation Eq.2 computes a global Similarity $simG(v_i, v_j)$ between two vertices v_i and v_j by the linear combination of 3 measures respectively corresponding to each type of information. $simT(v_i, v_j)$ is the attribute based similarity. It is an arithmetic average between discrete attribute based similarity $simADiscr(v_i, v_j)$ (determined by counting the number of attribute values nodes have in common) and textual attribute based similarity $simA(v_i, v_j) = \frac{1}{1+\sqrt{\Sigma_d(w_i^d - w_j^d)^2}}$ based on the euclidean distance. $simS(v_i, v_j)$ corresponds to the relational based similarity (see Sect.4.1).
And $simR(v_i, v_j) = \frac{|\Delta_{ij}|}{|\Delta_j|} * \frac{|\Gamma_j^{in} \cap \Gamma_i^{in}|}{|\Gamma_j^{in} \cup \Gamma_i^{in}| - \theta}$ as defined by Gamgne et al. [8], represents edge directionality based similarity which focuses on triad density and neighbourhood of vertices. Then the global similarity measure is used as pairwise similarity measure in the Louvain's method to partition the graph into clusters. The objective is to evaluate the scalability of the method based on this global similarity by extracting semantic clusters. α, β and γ are weighting factors that enable to give more importance to the structural, attribute or semantic similarity. $\gamma = 1 - \alpha - \beta$ and $\alpha, \beta, \gamma \neq 0$.

5. Experimental Study

In this section, we performed extensive experiments to evaluate the performance of the linear combination-based approach on real-world network datasets. All experiments were done on a $2.3GHz$ $Intel$ $Pentium$ IV PC with $6GB$ main memory, running Windows 8. Python and R package were used for implementations.

5.1. Experimental Datasets and evaluation measures

To our knowledge, there is no referenced benchmark with relational and attributes information handling link semantic (edge directionality). We construct a small ground truth dataset, a food web network, in order to compare each vertex to its real cluster. So, two datasets for experiments are used :

Food web : A typical illustration dataset as shown in Fig.1 is case of food web network where a vertex represents a species and edge the relationship between prey and predator.

Political Blogs Dataset: A directed network of hyperlinks between weblogs on US politics. This dataset contains $1,490$ weblogs with $19,090$ hyperlinks between these webblogs. Each blog in the dataset has an attribute describing its political leaning as either *liberal* of *conservative*.

We use two measures of Density and Normalized mutual information (NMI) to evaluate the quality of clusters generated by different methods.

5.2. Assumptions on food web illustration

Here we enumerate partitioning scenario and present expected results. We consider 5 subsets of vertices A, B, C, D, E describing species diet mode and by their reproduction mode, to be real semantic cluster of the hybrid clustering. The Table 1. shows the described illustration network according to each property :

Table 1: Number of species by nutrition sector and mode of reproduction

Diet Mode		Mode of reproduction	Number
A	Carnivorous	Viviparous	8
B	Carnivorous	Oviparous	3
C	Herbivorous	Viviparous	7
D	Herbivorous	Oviparous	4
E	Vegetables	Asexual or sexual	3
Total			25

– Semi attribute semantic (Textual) : 3 clusters in which species are grouped by their mode of reproduction. The ground truth partition is formally defined as $P_a = \{A \cup C, B \cup D, E\}$.

– Semi Relational-semantic (Neighbourhood) : 3 clusters in which species are grouped by their diet mode. The ground truth partition is formally defined as $P_r = \{A \cup B, C \cup D, E\}$.

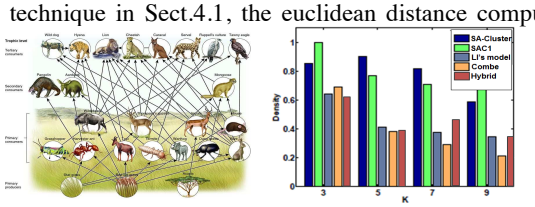
– Semantic : 5 clusters (species categories) : If we want to identify species by their both diet mode and mode of reproduction characteristics, then attributes(textual information), relational and directionality properties should be used. Like this, the resulting partition is $P_s = \{A, B, C, D, E\}$.

6. Model evaluations and results

6.1. Evaluation on illustration dataset

Given that this study focuses on directed attributed graphs which have not yet been investigated in detail, the evaluation consists in checking these assumptions described in Sect.5.2, by evaluating stated models of Sect.4 (M_a, M_r, SH_{ar}). We compare these 3 models (M) and (SH) with the hybrid model (H_s). The synthesis of results is shown in Table.2, according to the Normalized Mutual Information (NMI) measure [1]. Then clusters issued from the ground truth clustering transcripts the following partitions : the group of species by their diet mode (P_r), by their mode of reproduction (P_a), and by the both simultaneously (P_s).

– **Clustering according to textual attributes : M_a Model.** In this approach corresponding to the technique in Sect.4.1, the euclidean distance computed on the tex-



(a) Food web illustration network with Diet sectors

(b) Density comparison on Poltitical Blogs

Figure 1: Example of datasets and results

Table 2: Results : NMI

Models	P_r	P_a	P_s
M_r	**0.753**	0.350	0.323
M_a	0.741	**0.842**	0.625
SH_{ar}	$[0.028 - 0.291]$	$[0.205 - \mathbf{0.441}]$	$[0.085 - 0.397]$
H_s	$[0.098 - 0.217]$	$[0.110 - 0.185]$	$[0.558 - \mathbf{0.895}]$

tual attributes helps to weight each edge ; then an unsupervised method is applied to the resulting graph. The method performs well when the ground truth partition is $P_a = \{A \cup C, B \cup D, E\}$ by a higher NMI value (0.842) than considering the partitions P_r or P_s.

– **Clustering according to relations :** M_r **Model.** This method firstly exploits relations and secondly, with attributes handling, it detects communities so that the nodes in the same community are densely connected as well as homogeneous [2]. The NMI value for the ground truth partition $P_r = \{A \cup B, C \cup D, E\}$ is higher (0.753) than its value for the ground truth partition P_a and P_s. This result demonstrates that a technique based on successively relations then attributes, performs well in case of detecting two clusters of species with a densely internal connectivity, corresponding to diet mode.

– **Semi-hybrid attributed based clustering :** SH_{ar} **Model.** As far as this method is concerned, it deals with both types of information simultaneously as studied by Largeron [3] through a weighted distance function. In experiments, the NMI value fluctuates as a function of the weighting factors α and β. It changes its value according to the weighting factor α. NMI is in the interval $[0.028 - 0.291]$ for P_r ground truth and $[0.205 - 0.441]$ for P_a when α values are respectively 0.5 and 0.75. $\beta = 1 - \alpha$. SH_{ar} Model performs the best for the ground truth P_a, meaning that textual attributes describe better the vertices similarity, but produces weak outcomes as proved by [3] for the overall results.

– **Hybrid attributed based clustering :** H_s **Model.** The objective of this hybrid based experiment consists in 2 ways. First it shows that the consideration of the textual attributes improves better the cluster semantics through the highest NMI values as presented in bold in the Table.2. Second it shows that combining simultaneously the three types of information which are link semantic, relational and attribute properties respectively, leads to the highest NMI for that expected partition $P_s = \{A, B, C, D, E\}$. Like this, it detects the five classifying species clusters by their diet and reproduction mode simultaneously with a NMI value of 0.895 when the weighting factors α and β both equal 0.33; NMI value decreases to 0.558 when the weighting factors α and β equal 0.5 and 0.40 respectively, meaning that the negligence of the third factor relating to link semantic property affects the result.

6.2. Evaluation on Polblogs dataset

The Table 3 presents NMI for P_s partition, with $\alpha = \beta = \gamma = 0.33$, while the figure 1b compares Density for each model through the number of cluster. These results strengthen the interpretation according to that high density does not inevitably denote good separation of communities.

Table 3: Results : $Density$

Models	SAC1	SA-Cluster	Li's model	Combe's model	Hybrid model
NMI	0.153	0.350	0.323	0.675	**0.878**

7. Conclusion and future works

This work focused on the presentation of a hybrid clustering approach based on a proposed similarity. This measure takes into account 3 properties : semantic, relational and attributes. As presented below, we obtained different results according to the clustering technique and to the kind of data in the directed attributed food web graph we built.

An illustration on a food web network helped to underline the choice of each method relating to the kind of information (textual or numeric). The experiments show that on the one hand, the consideration of textual documents as attributes in the partitioning process leads to expected results based on the determination of species by their reproduction and nutrition modes simultaneously, and on the other hand, the properties strengthens the cluster semantic as computed through the NMI highest value. Nevertheless it has been difficult to integrate simultaneously two textual attributes relating to both reproduction mode and nutrition mode. For this reason, the second one has been processed as a numeric. Although this method is simple, it is hard to set/tune the parameters as well as interpret the weighted similarity function. Future works intend to apply large real-world networks and study weighting factors distribution.

8. References

[1] K. STEINHAEUSER, N. V. CHAWLA, "Identifying and evaluating community structure in complex networks", *Pattern Recognition Letters*, (2009).

[2] H. LI , Z. NIE , W. C. LEE, "Scalable Community Discovery on Textual Data with Relations", *ACM conference on Information and knowledge management*,pp. 1203-1212, (2008).

[3] D. COMBE, C. LARGERON , M. GERY, E. EGYED-ZSIGMOND "Détection de communautés dans des réseaux scientifiques à partir de données relationnelles et textuelles.", *MARAMI*, (2012).

[4] Y. ZHOU, H. CHENG , Y. JEFFREY XU "Graph Clustering Based on Structural/Attribute Similarities", *Adv. Intell. Data Anal.*, pp. 181-192 (2009).

[5] D. COMBE, C. LARGERON, M. GERY, E. EGYED-ZSIGMOND "I-louvain: An attributed graph clustering method.", *Adv. Intell. Data Anal. XIV*, pp. 181?192. Springer (2015).

[6] T. DANG, E. VIENNET "Community detection based on structural and attribute similarities.", *In: International Conference on Digital Society (ICDS)*, pp. 7-12 (2012).

[7] NEWMAN, M.E., GIRVAN M. " Detecting community structure in networks." *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 38(2), pp. 321-330, 2004.

[8] F. GAMGNE , N. TSOPZE, R. NDOUNDAM, " Novel method to find directed community structures based on triads cardinality." *Proceedings of CARI'16.*, vol. 2016, pp. 8-15, (2016).

[9] F. GAMGNE , N. TSOPZE, " Communautés et rôles dans les réseaux sociaux." *Actes du CARI'14*, pp. 157-164, (2014).

Generic heuristic for the `mnk-games`

Abdel-Hafiz ABDOULAYE*, Vinasetan Ratheil HOUNDJI*, Eugène C. EZIN*, Gael AGLIN**

* Institut de Formation et de Recherche en Informatique (IFRI)
Université d'Abomey-Calavi (UAC)
Abomey-Calavi
Bénin
** Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM)
Université catholique de Louvain (UCL)
Louvain la Neuve
Belgique

RÉSUMÉ. Les jeux en situation d'adversité sont très étudiés en intelligence artificielle. Parmi ces jeux, nous distinguons les jeux `mnk`. Un jeu `mnk` est un jeu dans lequel deux joueurs placent, chacun à son tour, une pièce de leurs couleurs respectives sur un plateau $m \times n$. Le vainqueur est le joueur qui obtient en premier un alignement de k pièces de sa couleur sur une ligne soit horizontalement, verticalement ou diagonalement. Pour la résolution de ce type de jeu, des algorithmes de recherche basés sur le parcours d'arbre comme alpha-beta sont utilisés avec des heuristiques spécifiques. Nous proposons une heuristique générique permettant d'évaluer les noeuds des arbres pour les jeux `mnk`. Nous utilisons une méthode d'apprentissage automatique (en particulier l'algorithme Q-learning) pour adapter les différents paramètres. Les résultats des tests montrent qu'en moyenne notre approche est meilleure que certaines heuristiques connues.

ABSTRACT. Adversarial games are very studied in artificial intelligence. Among these games, there are the `mnk-games`. An `mnk-game` is a board game in which two players take turns in placing a piece of their color on an $m \times n$ board. The winner is the player who first gets k pieces of his own color in a row; horizontally, vertically, or diagonally. For the resolution of this type of games, search algorithms based on tree search like alpha-beta are coupled with specific heuristics. In this paper, we propose a generic heuristic to evaluate the moves for `mnk-games`. Then we use a machine learning algorithm (in particular the Q-learning algorithm) to fit the different parameters of the heuristic to each `mnk-game`. This allows us to determine better parameters for the heuristic. For the tests, we associate it with alpha-beta. The experimental results show that our approach is better than some known heuristics.

MOTS-CLÉS : jeux mnk, heuristique générique, apprentissage automatique, Q-learning, alpha-beta.

KEYWORDS : mnk-games, generic heuristic, machine learning, Q-learning, alpha-beta.

1. Introduction

Artificial intelligence is a branch of computer science that aims to understand and build intelligent entities. It is involved in a variety of areas including games. A game is a good testing field for artificial intelligence. For determistic, turn-taking and zero-sum games, some methods based on the tree search as alpha-beta [6, 3] have been proposed to play them more easily and faster. Most of these methods use an evaluation function to improve the final result. Actually the search space of games can be very large and then it is difficult to explore the whole tree in a reasonable time. In this case, the usage of a heuristic represents an alternative. In adversarial search, a heuristic is a function applied on nodes, that evaluates the state of the game by estimating the players gain in order to choose the most promising move.

In this paper we focus on a particular game category, the `mnk-games`. An `mnk-game` [7, 12] is a board game in which two players take turns in placing a piece of their color on an $m \times n$ board. The winner is the player who first gets k pieces of his own color in a row ; horizontally, vertically, or diagonally. We propose a generic heuristic based essentially on the notion of threat (see Section 2). This heuristic makes an evaluation of hits given the parameters associated with the threats. The parameters have static values favoring the choice of the best move at a moment of the game. Firstly, the parameters of the proposed generic heuristic have been set experimentally but it does not always guarantee that they are good for each game. Therefore we use a machine learning algorithm to improve the quality of the parameters of the generic heuristic. We use and experiment an approach which, due to machine learning and specially reinforcement learning, permit to determine the parameters of the generic heuristic mentioned above in order to have better parameters.

This paper is organized as follows : Section 2 gives some theoretical notions used in the paper ; Section 3 explains our generic heuristic and the machine learning method which we use to improve the quality of the different parameters of the heuristic ; Section 4 presents some experimental results ; and Section 5 concludes and provides some perspectives.

2. Background

In this section we define the notion of threat, present some heuristics that use threats, and briefly explain the reinforcement learning.

2.1. Threat

In `mnk-games`, the threat is a very important notion. It represents a configuration of aligned pieces in a certain way that can assure to its player a certain winning trend. It may be advantageous or not because threats of player are always compensated with the threats of second player for the evaluation of the position. Several works use the notion of threat (see for example [1, 4, 14]).

To make this concept more understandable, we will take a `mnk-game` whose purpose is to align 5 pieces. We give the different possible threats in an environment in which the combination of five (05) aligned pieces is winning : the *four*, the *three*, the *two* and the *ones*. A *four* is an alignment of four pieces of one player either horizontally, vertically or diagonally. There are several types of *four* categorized into three categories. Below we present six configurations that give the *four* :

– Type 1 : four consecutive aligned pieces whose extremities are free ;

– Type 2 : four consecutively aligned pieces whose location at one extremity is free while the second is occupied ;

– Type 3 : four consecutive aligned pieces whose extremities are occupied ;

– Type 4 : four pieces aligned with a jump location and whose extremities locations are free ;

– Type 5 : four pieces aligned with a jump location and whose location at one extremity is free while the second is occupied ;

– Type 6 : four pieces aligned with a location jump and whose extremities locations are occupied.

These are the three categories : the *four open* (type 1), the *four half-open* (type 2, 4, 5, 6) and the *four closed* (type 3).

2.2. Heuristic of Shevchenko

Shevchenko [8] does an analysis of the combinations of pieces present on the game board on the lines as well as on the columns and the diagonals. The analysis concerns only the player's pieces. In this sense, Shevchenko only takes into account the threats of one player on the board. Moreover, given k, the number of pieces to align before winning, interresting threats are those of size of k, $k-1$, and $k-2$ with no distinction between half-open, open and closed types. Parameters are associated with the threats according to their size and remain unchanged until the end of the game. These settings are : 100 for the k size threat, 10 for the $k-1$ threat, and 1 for the $k-2$ threat. Shevchenko used his heuristic for Gomoku game.

2.3. Heuristic of Chua Hock Chuan

The application of the Chua Hock Chuan heuristic [5] requires to find the alignments of player's and opponent's pieces on the lines, the columns and the diagonals. It therefore considers the threats of the player and the opponent present on the board but only those of size k, $k-1$ and $k-2$. There is no distinction between half-open and open types. The parameters associated with the player's and opponent's threats are fixed until the end of the game. We have : threat size k (100 for the player and -100 for the opponent) ; threat size $k-1$: 10 for the player and -10 for the opponent ; threat size $k-2$: 1 for the player and -1 for the opponent. Chua Hock Chuan proposed this heuristic for Tic Tac Toe game.

2.4. Reinforcement learning

The reinforcement learning problem is a kind of direct framework of the problem of interaction learning to achieve a goal. The learner or decision maker is called the agent that interacts with its environment. The agent selects the actions and the environment responds and presents new situations to the agent. The environment gives rise to rewards, special numerical values that the agent tries to maximize. A complete specification of an environment defines a task, an instance of the reinforcement learning problem.

Formally, the basis of the reinforcement learning model is : a set of states S of the agent in the environment ; a set of actions A that the agent can perform ; and a set of reward scalar values R that the agent can obtain.

At each step t of the algorithm, the agent perceives its state $s_t \in S$ and the set of possible actions $A(s_t)$. It chooses an action $a \in A(s_t)$ and receives from the environment a new state s_{t+1} and a reward r_{t+1}. Based on these interactions, the reinforcement learning algorithm must allow the agent to develop a $\Pi : S \rightarrow A$ policy that allows him to maximize the amount of rewards. Thus the reinforcement learning method is particularly suited to problems that require a compromise between the quest for short-term rewards and long-term rewards.

If we had to identify a central and new idea to reinforce learning, it would certainly be learning by Temporal Difference (TD) [10, 11, 2]. TD learning is a machine learning method based on

prediction. TD methods use experience to solve the prediction problem. Given some experience following a Π policy, they update their v estimate of v_Π (value obtained by following the Π policy) for non-terminal states s_t occurring in this experiment. A policy is a rule that the agent follows for the choice of actions, given the state in which he is. At the moment $t+1$, they immediately form a target and make a useful update using the observed reward R_{t+1} and the estimate $V(S_{t+1})$. The equation (1) gives the update formula.

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)] \tag{1}$$

One of the most important advances in reinforcement learning has been the development of an off-policy TD control algorithm called Q-learning [11, 9, 13]. Q-learning is used to find an optimal action selection policy. It works by learning an action-value function that ultimately gives the expected utility of taking a given action in a given state and following the optimal policy thereafter. When such an action value function is learned, the optimal policy can be constructed by simply selecting the action with the highest value in each state. The algorithm has an update formula which calculates the quantity of a state-action combination :

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma max Q(S_{t+1}, a) - Q(S_t, A_t)] \tag{2}$$

Before learning begins, Q returns a fixed (arbitrary) value chosen. Whenever the agent selects an action, observes a reward and a new state (that may depend on both the previous state and the selected action) then Q is updated.

3. Generic heuristic and determination of the different parameters

This section specifies the basic elements for the generic heuristic's definition, gives the formula and describes the method that we propose for parameters determination using reinforcement learning.

3.1. Generic heuristic

For a game in which the number of pieces to line up to win is k, the big threat that needs to be created and that always leads to a win is the $k-1$ open type and the only combination that can come up in this configuration is the $k-2$ open type threat. On the other hand, the half-open threat of $k-2$ is not interesting in itself. Also, the $k-1$ half-open threat is very close to victory, it is less important than a $k-2$ open-car type. In the latter case, we can open $k-1$ half-open type as it can also lead to the $k-1$ open threat , which is very interesting. In general, the game becomes decisive when on threats of size $k-2$ and $k-1$. Threats smaller than $k-2$ are not interesting. These are threats that are not co-affected by the same weighting as those that are adverse or not. A large weighting is a given to the threats against the player's threats to prevent the opponent from taking an irreversible advantage. On the other hand, threats classified as uninvolved (less than $k-2$ and the half-open threat of $k-2$) are co-assigned in the same way as players.

Below we present the formula of the proposed generic heuristic based on the threats defined above. The different values of the parameters of the heuristic were first fixed regarding the priority of threats and experimentations.

$$A = \begin{cases} \sum_{i=1}^{k-3} (a_{2i-1}p_{i,1} + a_{2i}p_{i,2}) + a_{2(k-2)-1}p_{k-2,1} + 100p_{k-2,2} + 80p_{k-1,1} + 250p_{k-1,2} + 1000000p_k & \text{if } k > 3 \\ a_1 p_{k-2,1} + 100p_{k-2,2} + 80p_{k-1,1} + 250p_{k-1,2} + 1000000p_k & \text{if } k = 3 \end{cases}$$

$$B = \begin{cases} \sum_{i=1}^{k-3} (a_{2i-1}q_{i,1} + a_{2i}q_{i,2}) + a_{2(k-2)-1}q_{k-2,1} + 1300q_{k-2,2} + 2000q_{k-1,1} + 5020q_{k-1,2} + 1000000q_k & \text{if } k > 3 \\ a_1 q_{k-2,1} + 1300q_{k-2,2} + 2000q_{k-1,1} + 5020q_{k-1,2} + 1000000q_k & \text{if } k = 3 \end{cases}$$

$$f = A - B$$

in which A is the evaluation of the player's threats on the board ; B is the evaluation of the opponent's threats on the board ; a_i is the coefficient of the lower threat index i ; $p_{i,1}$ is the player's number of half-open threats of size i ; $p_{i,2}$ is the player's number of open threats of size i ; p_i is the player's number of threats without hole of size i ; $q_{i,1}$ is the opponent's number of half-open threats of size i ; $q_{i,2}$ is the opponent's number of open threats of size i : q_i is the pponent's number of threats without hole of size i : n is the number of alignment leading to victory.

3.2. Proposed method for automatic determination of parameters

The generic heuristic is based on the notion of threat. The parameters of the heuristic are assigned according to the importance of the threats and are the same for each game. The best option is to be able to determine the parameters adapted to the situations encountered during the resolution of the game and to the rules of displacement. In this setion we propose a way to automatically update the parameters of the heuristic with the Q-learning method.

Q-learning uses a quality evaluation function Q. This allows to have a table of values (Q values) that helps in the choice of an action when we are in a given state. The value of the parameters greatly influences the evaluation of a position and the choice of the move to play. Therefore, we consider that using Q-learning, we must evaluate the quality of the parameters of the heuristic. The Q values are the parameters to be determined. Before the learning begins, the Q function returns a fixed value chosen by the programmer. We recall that the heuristic used fixed parameters determined experimentally and which proved their worth. These are the parameters that we use as initial Q values.

The algorithm allows to update a value by time step, a parameter in our case. It is necessary to find the parameter to update for a given step. We get a new step when a move is made. The alpha-beta algorithm returns the best estimated move to go to the next step. We look for all the threats on the board after a move and identify the most important threat. The parameter associated with the latter is the update. Also we associate a reward to each type of threat since the chosen action leads to a certain configuration of threats. In general, it is null except for the goal state (state of the board where the player has the "n" required pawns aligned). Identifying the most important threat allows you, at this stage, to know which reward to use for the update.

The actual update is done using the formula (2) and requires finding the maximum value Q in the next state. It will be necessary to determine the types of threat that the legal movements could create in order to take the maximum of their parameters. If we have an S state that, after an A action, leads to an S' state, we simulate the possible moves from the S' state, collect the most important threats likely to be created after each move and identify the most important of them. Its parameter is the maximum value Q sought.

The evaluation of a position takes into account the threats of the player and the opponent. As a result, we also update the threat parameters of the opponent to avoid any bad evaluation of the heuristic. This update is done following the same principles that we described for the player. The parameters are the initial Q values and we identify the largest enemy threat created by an action. The Q function is used for the corresponding parameter and requires to find the maximum value of the parameters for the most important enemy threats obtained after the simulation from the S' state.

We make two updates at each step : one on the parameter associated with the most important threat of the player and the other on the parameter associated with the most important threat of the opponent.

In addition, the learning rate and the reduction factor of the update formula have a large impact on the learning process. Since the game play environment is entirely deterministic, we chose a learning rate of 1. The reduction factor was chosen to meet the basic requirements of heuristics. A value too close to 1 would make the threat coefficients too close to each other. This leads to a bad evaluation of the positions and thus makes the heuristic less efficient. A value close to 0 ensures to a certain extent a favorable difference between the parameters. After many tests, we selected $\gamma = 0.1$. The update formula (2) becomes :

$$Q(S, A) = R + \gamma maxQ(S', a) \tag{3}$$

We use Algorithm 1 to update the parameters of the heuristic.

Algorithm 1: Parameters Update Algorithm

INITIALIZE THE TABLE OF VALUES WITH THE EXISTING FIXED PARAMETERS
repeat
  Initialize S with the current state of the board
  repeat for each step of the episode
    Choose A `// action from alpha-beta algorithm using generic heuristics`
    Execute the action A
    Find the parameter P_P associated with the player's most important threat
    Find the parameter P_O associated with the opponent's most important threat
    Receive the reward R_P corresponding to the player's most important threat
    Receive the reward R_O corresponding to the opponent's most important threat
    Observe the following state S'
    $P_P(S, A) \leftarrow R_P + \gamma maxP_P(S', a)$ `//` $maxP_P(S', a)$ `is the maximum value of parameters for the most important player's threats that can be created from the new state` S'
    $P_O(S, A) \leftarrow R_O + \gamma maxP_O(S', a)$ `//` $maxP_O(S', a)$ `is the maximum value of parameters for the most important opponent's threats that can be created from the new state` S'
    $S \leftarrow S'$
  until *S terminal*
until *the end of the episode*

At each stage of the episode, the alpha-beta algorithm uses the heuristic with the new parameters for actions selection.

4. Experimental results

For the experiments we consider an agent who uses the alpha-beta algorithm and the generic heuristic (heuristic with the parameters automatically updated with the Q-learning) to a depth limit of 4. We conducted numerous tests to determine the time that would allow the proposed solution to provide a set of stable parameters. After analyzing the results, we make the intelligent agent play against itself for 5 episodes because from this level, the parameters seem not to change anymore and each of them seems to have been updated at least once.

For Gomoku, we check the performance by playing against a player who uses alpha-beta and Shevchenko's heuristic [8]. For Tic tac toe, we compare the improved heuristic and Chua Hock Chuan's heuristic [5].

We conducted several tests grouped into categories. For each game considered, we distinguish three (03) categories based on the number of pieces to be aligned to win the game : category 1 with $k = 3$; category 2 with $k = 4$; and category 3 with $k = 5$. For each category, we vary the size of the board with $m = n$ and the tests are done for according to the player who starts the game. The maximum size is 11.

4.1. Gomoku : Improved heuristic vs Shevchenko's heuristic

4.1.1. Category 1 : $k = 3$

See Table 1 and Table 2.

With $k = 3$ the first player always wins the game no matter the size of the board with one exception because for a size of 7 our approach wins the party as a second player. It defended herself and created advantageous situations. We conclude that for $k = 3$, the first player has an advantage that leads him to victory. For better analysis, we change k.

4.1.2. Category 2 : $k = 4$

See Table 3 and Table 4.

With $k = 4$, our approach always wins the game it plays first or not for a size bigger than k. It is very effective at this level.

4.1.3. Category 3 : $k = 5$

See Table 5 and Table 6.

With $k = 5$, our heuristic always wins the game when it plays first or not for a size over k. It is more efficient than Shevchenko's heuristic even if the game ends with a draw when it plays second for $k = 5$.

According to those three analysis, we conclude that the player using the improved heuristic won the most games regardless of the number of pieces to be aligned and the size of the board. Our heuristic has proven its efficiency against Shevchenko's heuristic.

4.2. Tic tac toe : Generic heuristic vs Chua Hock Chuan's heuristic

4.2.1. Category 1 : $k = 3$

See Table 7 and Table 8.

As second player, the generic heuristic is totally out of date and is not effective as first player.

4.2.2. Category 2 : $k = 4$

See Table 9 and Table 10.

Analysis 5 : With $k = 4$, our approach always wins the game no matter it plays first or not for a size bigger or equal to 6. It concedes no defeat. This number of pieces to align is favorable.

4.2.3. Category 3 : $k = 5$

See Table 11 and Table 12.

Our approach always wins the game no matter it plays first or not for a size bigger than or equal to 8. It concedes no defeat. Just like $k = 4$, $k = 5$ is good for our approach.

The Chua Hock Chuan's heuristic [5] showed the effectiveness of the latter for $k = 3$ but our approach was successful for $k > 3$.

5. Conclusion and perspectives

In this work, we have determined a generic evaluation function for `mnk-games` to solve all games in this category with the same solution. Then we have proposed a method for determining the parameters of the generic heuristic using machine learning to obtain better parameters for each game considered. It is based on the functioning of Q-learning, which is an "off-policy" TD control algorithm. We combined the improved heuristic, Shevchenko's heuristic and Chua Hock Chaun's heuristic with the alpha-beta algorithm to make a comparison on Gomoku and Tic tac toe. The results of the tests showed that at different levels the improved heuristic is on average more efficient.

As future works we would like to intensively compare our generic heuristic on a large variety of `mnk-games` wrt other heuristics. A good perspective of this work is to determine some elements to improve the approach proposed by focusing on the eligibility traces associated with TD methods [4]. We can also automatically learn the winning strategies used in games played for each `mnk-game`.

6. Bibliographie

[1] L.V. ALLIS, H.J. VAN DEN HERIK, M.P.H. HUNTJENS, Go-moku solved by new search techniques, Computational Intelligence, 12(1), p.7-23., 1996.

[2] ANDREW G. BARTO, Temporal difference learning, *www.scholarpedia.org/article/Temporal_difference_learning*, 2007.

[3] TRISTAN CAZENAVE, Des Optimisations de l'Alpha-Beta, Laboratoire d'Intelligence Artificielle, Département Informatique, Université Paris 8, 2011.

[4] TRISTAN CAZENAVE, A Generalized Threats Search Algorithm, International Conference on Computers and Games. Springer, Berlin, Heidelberg, 2002.

[5] CHUA HOCK CHUAN, Java Games, *http ://www3.ntu.edu.sg/home/ehchua/programming/java /JavaGame_TicTacToe_AI.html*, 2017.

[6] SAMUEL H. FULLER, JOHN G. GASCHNIG, *Analysis of the alpha-beta pruning algorithm*, Department of Computer Science, Carnegie-Mellon University, 1973.

[7] H.J VAN DEN HERIK, J.W.H.M UITERWIJK, J.V RIJSWIJCK, Games solved : Now and in the future, *Artificial Intelligence*, Vol. 134, p. 277-311, 2002.

[8] MYKOLA SHEVCHENKO, GOMOKU & Minimax-alphabeta search, *https ://github.com/nshevchenko/GomokuAlphabeta/blob/master/gomoku-minimax-alphabeta.pdf, 2016.*

[9] OLIVIER SIGAUD, OLIVIER BUFFET, *Markov Decision Processes in Artificial Intelligence*, The MIT Press Cambridge, Massachusetts, 2010.

[10] RICHARD S. SUTTON, *Learning to Predict by the Method of Temporal Differences*, Machine Learning, vol.3, p.9-44, 1988.

[11] RICHARD S. SUTTON, ANDREW G. BARTO, Reinforcement Learning : An Introduction, *The MIT Press Cambridge*, Massachusetts, 2012.

[12] J.W.H.M UITERWIJK, H.J VAN DER HERIK, *The advantage of the initiative*, Information Sciences, p. 43-58, 2000.

[13] CHRISTOPHER WATKINS, PETER DAYAN, *Q-learning*, Machine Learning, p. 279-292, 1992.

[14] I-CHEN WU, DEI-YEN HUANG, A New Family of k-in-a-row Games, Advances in Computer Games, 2005.

7. Appendix

7.1. Gomoku experimental results

7.1.1. Category 1 : $k = 3$

Tableau 1. *Gomoku : Generic heuristic vs Shevchenko's heuristic for $k = 3$, test n°1*

Players	First Player	Taille $\in$ [3,11]
Alpha-beta + generic heuristic	Yes	Win
Alpha-beta + heuristic of Shevchenko	No	Loss

Tableau 2. *Gomoku : Generic heuristic vs Shevchenko's heuristic for $k = 3$, test n°2*

Players	First player	Taille $\in \{3,4,5,6,8,9,10,11\}$	Taille = 7
Alpha-beta + generic heuristic	No	Loss	Win
Alpha-beta + heuristic of Shevchenko	Yes	Win	Loss

7.1.2. Category 2 : $k = 4$

Tableau 3. *Gomoku : Generic heuristic vs Shevchenko's heuristic for $k = 4$, test n°3*

Players	First player	Taille = 4	Taille $\in$ [5,11]
Alpha-beta + generic heuristic	Yes	Draw	Win
Alpha-beta + heuristic of Shevchenko	No	Draw	Loss

Tableau 4. *Gomoku : Generic heuristic vs Shevchenko's heuristic for $k = 4$, test n°4*

Players	First player	Taille = 4	Taille $\in$ [5,11]
Alpha-beta + generic heuristic	No	Draw	Win
Alpha-beta + heuristic of Shevchenko	Yes	Draw	Loss

7.1.3. Category 3 : $k = 5$

Tableau 5. *Gomoku : Generic heuristic vs Shevchenko's heuristic for $k = 5$, test n°5*

Players	First player	Taille $\in$ [5,11]
Alpha-beta + generic heuristic	Yes	Win
Alpha-beta + heuristic of Shevchenko	No	Loss

Tableau 6. *Gomoku : Generic heuristic vs Shevchenko's heuristic for $k = 5$, test n°6*

Players	First player	Taille = 5	Taille $\in$ [6,11]
Alpha-beta + Generic heuristic	No	Draw	Win
Alpha-beta + heuristic of Shevchenko	Yes	Draw	Loss

7.2. Tic tac toe experimental results

7.2.1. Category 1 : $k = 3$

Tableau 7. *Tic tac toe : Generic heuristic vs Chua Hock Chuan's heuristic for $k = 3$, test n°7*

Players	First player	Size = 3	Size = $\{4,6,7\}$	Size = $\{5,8,9,10,11\}$
Alpha-beta + generic heuristic	Yes	Draw	Loss	Win
Alpha-beta + heuristic of Chua Hock Chuan	No	Draw	Win	Loss

Tableau 8. *Tic tac toe : Generic heuristic vs Chua Hock Chuan's heuristic for $k = 3$, test n°8*

Players	First player	Size = 3	Size $\in$ [4,11]
Alpha-beta + generic heuristic	No	Draw	Loss
Alpha-beta + heuristic of Chua Hock Chuan	Yes	Draw	Win

7.2.2. Category 2 : $k = 4$

Tableau 9. *Tic tac toe : Generic heuristic vs Chua Hock Chuan's heuristic for $k = 4$, test n°9*

Players	First player	Size = $\{4,5\}$	Size $\in$ [6,11]
Alpha-beta + generic heuristic	Yes	Draw	Win
Alpha-beta + heuristic of Chua Hock Chuan	No	Draw	Loss

Tableau 10. *Tic tac toe : Generic heuristic vs Chua Hock Chuan's heuristic for $k = 4$, test n°10*

Players	First player	Size = $\{4,5\}$	Size $\in$ [6,11]
Alpha-beta + generic heuristic	No	Draw	Win
Alpha-beta + heuristic of Chua Hock Chuan	Yes	Draw	Loss

7.2.3. Category 3 : $k = 4$

Tableau 11. *Tic tac toe : Generic heuristic vs Chua Hock Chuan's heuristic for $k = 5$, test n˚11*

Players	First player	Taille $\in$ [5,7]	Taille $\in$ [8,11]
Alpha-beta + generic heuristic	Yes	Draw	Win
Alpha-beta + heuristique de Chua Hock Chuan	No	Draw	Loss

Tableau 12. *Tic tac toe : Generic heuristic vs Chua Hock Chuan's heuristic for $k = 5$, test n˚12*

Players	First player	Taille = $\{5,6\}$	Taille $\in$ [7,11]
Alpha-beta + generic heuristic	No	Draw	Win
Alpha-beta + heuristic of Chua Hock Chuan	Yes	Draw	Loss

Scaling the ConceptCloud Browser to Large Semi-Structured Data Sets

Joshua Berndt, Bernd Fischer, Arina Britz

CSIR Center for AI Research
Stellenbosch University
South Africa

RÉSUMÉ. Les ensembles de données semi-structurés, tels que les révisions de produits ou les données de journaux d'événements, deviennent simultanément plus largement utilisés et en même temps de plus en plus volumineux. Cet article décrit ConceptCloud, un navigateur interactif flexible pour les ensembles de données semi-structurés, mettant l'accent sur les modifications architecturales à une architecture basée sur serveur apportées pour accommoder des ensembles de données en constante croissance. ConceptCloud utilise une combinaison d'une visualisation intuitive du nuage de tags avec un treillis des concepts sous-jacent pour fournir une structure formelle pour la navigation dans un ensemble de données sans connaissance préalable de la structure des données ou compromettre l'évolutivité.

ABSTRACT. Semi-structured data sets such as product reviews or event log data are simultaneously becoming more widely used and growing ever larger. This paper describes ConceptCloud, a flexible interactive browser for semi-structured datasets, with a focus on the recent trend of implementing server-based architectures to accommodate ever growing datasets. ConceptCloud makes use of an intuitive tag cloud visualization viewer in combination with an underlying concept lattice to provide a formal structure for navigation through datasets without prior knowledge of the structure of the data or compromising scalability. This is achieved by implementing architectural changes to increase the system's resource efficiency

MOTS-CLÉS : architecture client-serveur, données semi-structurés, nuage de tags, treillis de concepts

KEYWORDS : client-server architecture, semi-structured data, tag cloud, concept lattice

1. Introduction

ConceptCloud [5] is a visualisation and exploration tool for semi-structured data sets, such as software revision control meta-data or product reviews. It uses a concept lattice [2] built from the data set as underlying navigation structure but presents the data itself in form of a tag cloud that concisely summarizes the users current selection in one view and allows further navigation through tag selection and deselection, without constricting the user to pre-defined search paths.

ConceptCloud began as an interactive browser based tool for Git and SVN repositories [1, 4] and has been extended to accept further semi-structured data sets in XML and JSON files. However, its application to large data sets (e.g., the ACM Digital Library, see [6]) have shown the limitations of its original client-based architecture. We have thus re-designed and re-implemented the system to use a new server-based architecture, which also necessitated some user interface changes. In this paper we describe the new architecture and interface and show that it yields 10x performance improvements. Specifically, we present the formal background for the ConceptCloud System, limitations of the original implementation, changes made, and preliminary experimental results over a wine review data set.

2. Formal Concept Analysis

Formal Concept Analysis (FCA) is a theory of data analysis that uses lattice-theoretic methods to investigate abstract relations between objects and their attributes. In FCA, information is represented as a binary cross table, or *context*, where the rows denote objects, eg. products and the columns attributes eg. price or ratings. ConceptCloud uses the concept lattice derived from semi-structured data as its navigation structure. An incidence relation $\mathcal{I}$ indicates which objects in the table have which attributes.

Definition 1 *A formal context is a triple $(\mathcal{O}, \mathcal{A}, \mathcal{I})$ where $\mathcal{O}$ and $\mathcal{A}$ are sets of objects and attributes, respectively, and $\mathcal{I} \subseteq \mathcal{O} \times \mathcal{A}$ is an arbitrary incidence relation.*

Definition 2 *Let $(\mathcal{O}, \mathcal{A}, \mathcal{I})$ be a context, $O \subseteq \mathcal{O}$, and $A \subseteq \mathcal{A}$. The common attributes of O are defined by $a(O) = \{a \in \mathcal{A} | \forall o \in O : (o, a) \in \mathcal{I}\}$, the common objects of A are denoted by $\omega(A) = \{o \in \mathcal{O} | \forall a \in A : (o, a) \in \mathcal{I}\}$.*

Formal concepts are pairs of objects and attributes $\langle O, A \rangle$, where $O \subseteq \mathcal{O}$ and $A \subseteq \mathcal{A}$ such that O is the set of all objects that have all attributes from A and A is the set of attributes that are common to all objects in O.

Definition 3 *Let $\mathcal{C}$ be a context. $c = \langle O, A \rangle$ is called a concept of $\mathcal{C}$ iff $\alpha(O) = A$ and $\omega(A) = O$. $\pi_O(c) := O$ and $\pi_A(c) := A$ are called extent and intent of c, respectively. The set of all concepts of $\mathcal{C}$ is denoted by $B(\mathcal{C})$.*

Concepts are partially ordered by inclusion of extents such that a concepts extent includes the extent of all of its subconcepts ; the intent-part follows by duality.

Definition 4 *Let $\mathcal{C}$ be a context, $c_1 = \langle O_1, A_1 \rangle$, $c_2 = \langle O_2, A_2 \rangle \in B(C)$. c_1 and c_2 are ordered by the subconcept relation, $c_1 \leq c_2$, iff $O_1 \subseteq O_2$ or equivalently, $A_2 \subseteq A_2$. The structure of $B(C)$ and $\leq$ is denoted by $\mathcal{B}(\mathcal{C})$.*

The basic theorem of FCA states that the structure induced by the concepts of a formal context and their ordering is always a complete lattice [2]. Such concept lattices have strong mathematical properties and reveal structural and hierarchical properties of the original data. They can be computed automatically from any given relation between objects and attributes. The greatest lower bound or meet and least upper bound or join can also be expressed by the common attributes and objects.

Theorem 1 *Let $\mathcal{C}$ be a context, then $\mathcal{B}(\mathcal{C})$ is a complete lattice, called the concept lattice of $\mathcal{C}$. Its meet and join operation for any set $\{\langle A_i, B_i\rangle | i \in I\} \subset B(\mathcal{C})$ of concepts are given by :*
$\bigwedge_{i\in I}(O_i, A_i) = (\bigcap_{i\in I} O_i, \alpha\,(\omega(\bigcup i \in I A_i)))$
$\bigvee_{i\in I}(O_i, A_i) = (\omega(\alpha(\bigcup_{i\in I} O_i)), \bigcap_{i\in I} A_i)$

Each attribute and object has a uniquely determined defining concept in the lattice. The defining concepts can be calculated directly from the attribute or object, respectively, and need not be searched in the lattice.

Definition 5 *Let $\mathcal{B}(\mathcal{O}, \mathcal{A}, \mathcal{I})$ be a concept lattice. The* defining concept *of an attribute $a \in \mathcal{A}$ is the greatest concept c such that $a \in \pi_A(c)$ holds. It is denoted by $\mu(a)$. The defining concept of an object $o \in O$ is the smallest concept c such that $o \in \pi_O(c)$ holds. It is denoted by $\sigma(o)$.*

Efficient algorithms exist for the computation of the concept lattices and the meet and join of concepts in the lattice [3]. For a detailed introduction to FCA see [2].

3. ConceptCloud

ConceptCloud [5] is a browser for semi-structured datasets which allows the user to navigate, via tag clouds, through a dataset in what is known as an *explorative search*. This type of exploration requires no predefined knowledge of the domain or dataset. The user iteratively selects an attribute or object tag in a tag cloud, and the ConceptCloud system adjusts the tag cloud to display all other tags attached to objects possessing the selected attribute tag(s). This is achieved by maintaining a focus concept from which a tag cloud is created.

Formally, the *focus concept* $c = \langle O, A\rangle$ is the concept whose extent is the set of objects that share the set of currently selected attributes, F, within the tag cloud, such that $\alpha(\omega(F)) = \pi_A(c) = A$. the new focus concept. concept and A new is the attributes of the new focus concept.

The focus concept can be further refined by iteratively adding elements to F. When an additional attribute is added to F, we update the focus concept by computing the meet, as per Theorem 1, of the current focus concept c and the concept introduced by the additional attribute. In Section 4 we will discuss how this was changed.

The explorative search process corresponds to the process of stepping through a concept lattice, wherein the selection of an attribute moves us to the point in the lattice where all linked objects contain that attribute. As we select further attributes we move further down the lattice. If we deselect an attribute we move back up the lattice and have access to a different set of attributes and objects. This corresponds to the refinement of the focus concept by adding and removing elements from F. This approach was tested in a user study conducted in [6] and found that users were able to answer complex scientometric

questions using ConceptCloud with a mean correctness of 73%, with the users' prior research experience having no statistically significant effect on results. For further detail see [1]. ConceptCloud presents the data in the form of a tag cloud, where the frequency of each tag denotes its importance. Each tag cloud is a word cloud-like window, wherein all of the objects and attributes in the lattice are represented as tags, words whose size denote their importance within this window. More specifically in ConceptCloud, each tag in a tag cloud's size is based on the frequency of its occurrence within the sub selection of the dataset, coloured differently based on its category (namely the type value of the attribute or object). Tag clouds are constructed, to help distinguish the different properties of the data set, by taking the extent of the focus concept $c = \langle O, A \rangle$, then for each $o_i \in O$, we determine its defining concept c_i, see Definition 5. We then collect all the intents of these defining concepts. These are the attributes we display in the tag cloud. Finally we add the objects to the tag cloud so that they may be directly selected or searched within the tag cloud. Our initial focus concept will have no selected attributes, and thus the tag cloud created from it will contain tags representing all attributes and objects. Formally we have (here $\uplus$ denotes multiset union) :

Definition 6 *The tag cloud from a concept c = (O, A) ∈ B(𝒞) is defined as τ (c) =* $O \uplus \biguplus_{o \in O} \pi_A(\sigma(o))$.

By constructing the objects in the tag cloud, we induce subconcepts of the focus concept, from which the tag cloud was derived, and all concepts having a non-bottom meet with that focus concept.

The initial implementation of the ConceptCloud system was geared towards exploration of the metadata of software repositories [1, 5]. As such it was not built with scaling in mind since the metadata within a software repository forms a comparatively small semi-structured dataset. When the use of the application shifted from analysis of these repositories to analysis of other semi-structured datasets[6], it became apparent that some of the design choices initially made were no longer feasible. One such choice was to have each tag cloud display tags representing all attributes and objects without any limit. The lack of a display limit also exists in the displayed representation of the context table, which too will display all attributes and objects. In practice this is increasingly resource intensive for larger datasets.

4. Scalable ConceptCloud Architecture

In order to reduce the resource intensive nature of ConceptCloud, changes to the initial implementation had to be made. A fixed sized subset of all tags was selected to represent the underlying concept lattice. This in turn necessitated a way for the user to interact with tags that may not be displayed in the initial window. The logical choice was to incorporate autocomplete based search functionality, as mentioned in Section 4.2. For this to function correctly a caching structure and separation of the data in memory was required. For this we implemented a postgresql database, which table's are generated based off the structure of the input dataset. The number of tags in a tag cloud was limited to the top 5000 tags, by frequency in the extent of the focus concept for that tag cloud. This limit was imposed to maintain a functional interface and not overwhelm the user. Additionally the attributes to be displayed in the context table representation was configured and limited. The context table representation, known as the *table view* limits the displayed results but allows the user to page through the list of all results. Finally the system creates its concepts on the

fly, meaning that the overhead of creating the full lattice is avoided, allowing for generally responsive tag cloud creation and rendering.

The architectural changes presented are a generalization of the architecture used in [6], a highly specialized version of ConceptCloud used to visualize the DBLP, Computer Science Bibliography. This dataset, at the time of writing, has over 4 million records. The scalable ConceptCloud Architecture presented is not specialized to any specific dataset and may be used for any well formed JSON dataset. It will correctly generate the required caching databases and create the required tag clouds. Additionally the user interface was updated to better function with the new architecture.

4.1. ConceptCloud User Interface

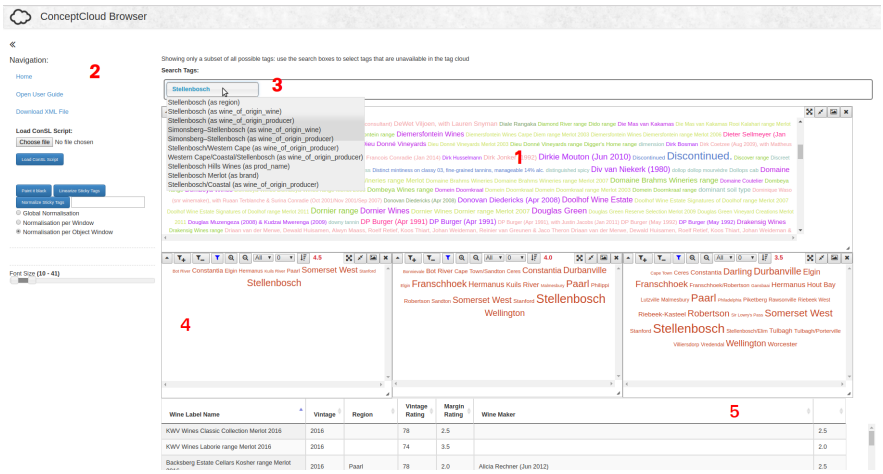


Figure 1. *Navigation User Interface*

The navigation user interface consists of the following components :

– The *main window* (1) wherein the tag clouds are displayed. On initial execution this displays the initial tag cloud viewer with the default focus. The tag cloud within the main window displays the top 5000 most relevant tags. A user selecting a tag in this window causes the focus concept to be recalculated and the viewer to then display the point in the lattice wherein the updated focus concept is relevant.

– The *Navigation Menu* (2) provides the user with various utilities relating to saving the lattice as well as uploading a ConSL script [4] to automate the display of the viewers. The further options allow the user to change the scaling of the tags within the tag cloud viewers. 4.1.

– *Search Functionality* (3) was introduced as only the top 5000 tags are displayed in the main window, the user may wish to interact with tags that are not currently displayed. As the user inputs their search terms, ConceptCloud displays an auto-completed list of terms and their category. This is done by making use of the caching database. Selecting one of these terms updates the focus concept, and as a result, the main window together with any other tag cloud viewer that contains the selected term as its focus concept. This action is identical to if they were to select a displayed tag in the main window.

– *Sticky Tag Cloud Viewers* (4) are sub-windows of the main window that contain each displayed tag cloud, and once created always appear below the main window, they can however be moved from their initial position. Each Sticky Tag Cloud Viewer contains the displayed tags for the sticky concept for that viewer, referred to as the sticky tag. A sticky tag is an object or attribute to which the viewer is fixed. Selecting a new focus concept will adjust these windows to use the union of the sticky and selected focus concept as their focus concept. The sticky tag is displayed next to the Tag Cloud viewer's menus in red. The menus exist to adjust the display of the contained tags. These viewers allow the user to have multiple differentiated views, eg. viewing the ratings of wines across multiple vintages, with a window for each vintage or rating.

– The *Table View* (5) displays the underlying context table for the concept lattice connected to the initial tag cloud viewer. Selecting and deselecting tags will cause this to update to reflect the concept table corresponding to the concept lattice of the focus concept(s). In many datasets there is a multitude of attributes for each object in the context table, often too many to display concisely. The attributes to be displayed in the Table View can be configured to solve this. The results appear in a fixed page size list, further easing resource usage.

4.2. Navigation Architecture

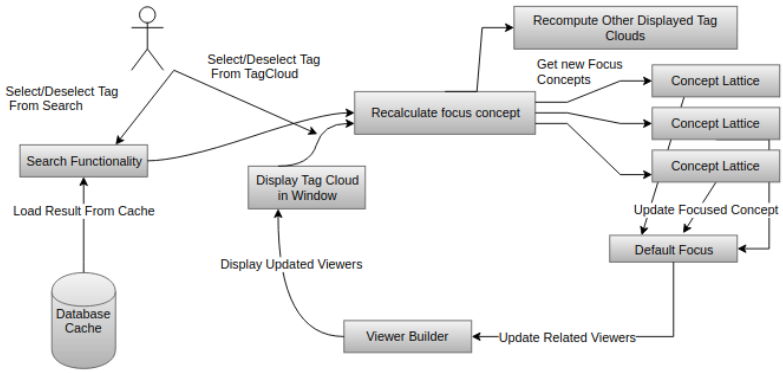


Figure 2. *Navigation Architecture*

An explorative search in ConceptCloud is the process whereby the user selects and deselects tags in a tag cloud allowing them to step through the underlying concept lattice. The user is additionally able to create sub windows, which are additional tag cloud viewers with stickied tags. These sub windows have one attribute set for it and will display the related subsection of the concept lattice. Selecting a new focus concept from a tag in any of the viewer windows, including the initial tag cloud, causes the selected tag union with stickied tag for each window (as the tag represents either an attribute or an object), to become the focus concept of each viewer. The related portion of the lattice is displayed if stickied tag and new focus concept are not disjoint. Otherwise an empty viewer window is displayed. The architecture of the navigation subsystem is outlined in figure 2. Limiting the displayed tags necessitated that we have two ways to interact with the tags. One based directly on the selection of a displayed tag in a tag cloud, and another based on searching for a tag that may or may not be displayed within a tag cloud. In this way we are able

to maintain the original ConceptCloud functionality. The architecture of the navigation subsystem is shown in Figure 2 :

– Select / Deselect Tags from the tag cloud : The user clicks a tag within one of the displayed tag clouds, this then causes all tag cloud windows to take this new selected tag and use it to recalculate their focus concept, causing each tag cloud window/viewer to update their controllers to request the relevant section of the underlying concept lattice should it exist. A new tag cloud is then constructed by the viewer builder and displayed, in the case of multiple concepts, if no intersection in the lattice between these concepts exists, an empty tag cloud is displayed in the corresponding viewer. Deselection of a tag is when the user clicks the highlighted red tag in the tag cloud, this removes it from all the tag cloud window's focus concept and updates the lattice for each window accordingly.These actions correspond directly to the explorative search mentioned in Section 3.

– Select / Deselect Tag From Search : An autocomplete based search which starts autocompletion after three characters, providing the user with the tag name and the category in which they wish to search. As the user enters text into the search bar, the search functionality performs a lookup in the database cache, and provides a list of closest tag name matches, and their categories to the user in the form of a dropdown list. The user may then click a tag from the list. From this point onwards the system acts as if a tag had been selected from a tag cloud as before. All tag cloud windows add the selected tag to their focus concept and all underlying lattices are updated. The viewer builder constructs new viewers with the updated lattice sections, causing all tag clouds to be updated with the relevant data. Deselection works identically as above.

4.3. Experiments

To show the difference in the architectures, we ran a series of experiments with a typical application driven semi-structured dataset. A series of typical user actions were taken, automated and then timed to display the differences in execution times for the different architecture.

The dataset used, contained 16306 wine reviews, where each review has the following attributes : name, varietal, vintage, review year, review, reviewer, points, price, country, location, region, winery, review phrases. Where the final field, review phrases, is a keyphrase extraction of the review field. This dataset was chosen as it succinctly displays the difference in performance between the two architectures.

For the experiments the following actions serve as our experiments ; initial rendering and the creation of new windows for high, medium and low volume tags. All times given are in milliseconds. These are all run on a machine with the following processing specifications, a 6th Generation Intel Core i7-6700HQ (3.5GHz) and 8Gb DDR4 2133Mhz.

For each architecture, the server and client response times are measured. The results, averaged across 20 runs were as follows :

User Action	Old Architecture (ms)	New Architecture (ms)
Initial Rendering	4863	378
Category Change	182	42
New High Volume Tag Render	7822	488
New Medium Volume Tag Render	4904	374
New Low Volume Tag Render	4218	314

The user actions carried out involved the following ; changing the category filter to va-

rietal, creating a new tag cloud with the United States as the high volume tag (count of 5335), creating a new tag cloud with 2005 vintage as the medium volume tag (count of 1782) and creating a new tag cloud with 2001 as a vintage for the low volume tag (count of 190).
We note that for each action the execution time is in each case at least an order of magnitude faster for the server-based architecture when compared to the same operation on the old architecture, on an identical dataset. The large speedup is due to the much lower resource cost of the new architecture, as we are no longer rendering the entire dataset, but only the top 5000 tags and a far smaller table view. Even when rendering less than 5000 tags, the fact that the initial cloud and table view are so resource intensive in the old client-based architecture means creating any additional tag clouds will suffer. The new server based architecture does not have have this issue.

5. Future Work

In this paper we described ConceptCloud, an interactive browser for semi-structured datasets and the changes made to it to enable it to more easily deal with large datasets. We showed that changes made resulted in a large speedup that made using large datasets feasible. For future work there are plans to move the ConceptCloud application from a client server application using a web client, to a client server with a mobile client. Currently we are working on adding support for processing ontological datasets, and using ontologies to enrich the data within the dataset. Finally we are adding support for geolocation data, and specialized interactions such as opening a tag on a map based on it's geolocation.

6. Bibliographie

[1] G J. GREENE AND B. FISCHER, « Interactive Tag Cloud Visualization of Software Version Control repositories », *Software Visualization (VISSOFT), 2015 IEEE 3rd Working Conference*, 2015.

[2] B. GANTER AND R. WILLE, « Formal Concept Analysis - Mathematical Foundations », *Springer, 1999.*

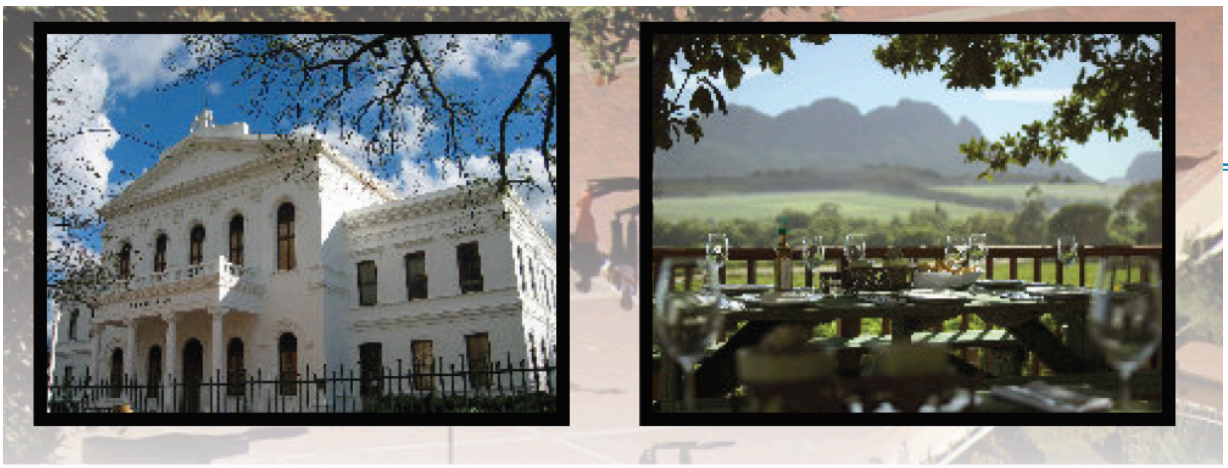
[3] ZAKI, MOHAMMED AND HSIAO, CHING-JUI AND OTHERS « CHARM : An Efficient Algorithm for Closed Association Rule Mining », *1999.*

[4] G J. GREENE, M. ESTERHUIZEN, B. FISCHER « Visualizing and Exploring Software Version Control Repositories using Interactive Tag Clouds over Formal Concept Lattices », *Information & Software Technology volume volume 87, 223–241 Elsevier, 2017.*

[5] G J. GREENE, B FISCHER « Conceptcloud : A Tag Cloud Browser for Software Archives. », *ACM SIGSOFT International Symposium on Foundations of Software Engineering, 22, 759–762 ACM, 2014.*

[6] M. DUNAISKI, G J. GREENE, B. FISCHER « Exploratory Search of Academic Publication and Citation Data using Interactive Tag Cloud Visualizations », *Scientometrics, Volume 110(3), 1539–1571 Elsevier, 2017.*

CARI 2018, Stellenbosch, South Africa



Sponsored by:

AUF Agence Universitaire de la Francophonie

cirad La recherche agronomique pour le développement

Inria

IRD Institut de Recherche pour le Développement

ifip

EDITORS : NABIL GMATI - ERIC BADOUEL - BRUCE WATSON

EDITORS : NABIL GMATI - ERIC BADOUEL - BRUCE WATSON