Fouille de données agricoles par Modèles de Markov cachés

Jean-François Mari, Florence Le Ber, Marc Benoît

Fouille de données agricoles par modèles de Markov cachés

Jean-François Mari[1] Florence Le Ber[1,2] Marc Benoît[3]

[1] UMR 7503 LORIA

[2] INRA LIAB

[3] INRA SAD

UMR 7503 LORIA, B.P. 239, 54506 Vandœuvre-lès-Nancy

jfmari@loria.fr

Résumé

Nous développons des outils de fouille de données spatio-temporelles à partir de modèles de Markov d'ordre supérieur. Ces modèles permettent de représenter des observations temporelles et spatiales comme des successions d'états où les transitions entre états dépendent, suivant l'ordre du modèle, de l'état courant et des n *états précédents. Ils ont été utilisés sur des données* `Ter Uti`*, qui sont des données spatio-temporelles d'utilisation du territoire, afin d'extraire les régularités d'utilisation des terres agricoles. Ce travail a été effectué en lien avec des experts agronomes. Dans ce papier, nous détaillons les modèles utilisés et la démarche mise en œuvre avec les agronomes. Nous présentons aussi des outils de visualisation que nous avons développés pour faciliter l'appropriation par les experts des résultats de la fouille. Finalement nous montrons l'intérêt de notre approche pour la fouille de données spatio-temporelles.*

Mots Clef

Fouille de données spatio-temporelles, successions culturales, données `Ter Uti`, modèles stochastiques, modèles de Markov cachés.

Abstract

We are developing tools for temporal and spatial data mining, using high-order Hidden Markov Models. These models are capable to represent spatial and temporal sequences of states in which the transitions between the states depend on the `n` *previous states according to the order of the Markov chain. They have been applied on spatial and temporal data concerning land use, named* `Ter Uti` *data, in order to find agricultural land use regularities. This paper presents the models we have used and the way we have interacted with the agronomists. Also, we describe some tools that we have developed to help data visualisation. We show that Hidden Markov models are powerful tools for temporal ans spatial data mining.*

Keywords

Temporal data mining, crop rotations, `Ter Uti` data, stochastic models, Hidden Markov Models.

1 Introduction

Dans [14], nous avions défini la *fouille* ou *extraction de connaissances à partir de données* comme une activité consistant à analyser un ensemble de données brutes de façon à en extraire des informations qui peuvent être considérées comme des éléments de connaissances et donc devenir exploitables [16, 10]. Un système de fouille de données s'articule généralement autour de quatre composantes principales : `(1)` les bases de données et leur système de gestion, `(2)` un système à base de connaissances d'aide à la résolution de problème sur le domaine relatif aux données, `(3)` un système d'étude et d'analyse de données symboliques, s'appuyant sur l'induction ou la classification, un système d'analyse de données numériques et de statistiques, `(4)` une interface se chargeant des interactions et de la visualisation des résultats intermédiaires et finaux.

Dans cet article, nous présentons nos travaux en extraction de connaissances à partir de données numériques, temporelles et spatiales. Ces travaux ont été développés sur une application concernant l'étude de successions de cultures, en lien avec des chercheurs agronomes. Nous nous limitons pour le moment aux éléments `(1)`, `(3)` et `(4)` du système idéal décrit ci-dessus.

La base de données contient des données temporelles et spatiales sur l'occupation du territoire, relevées systématiquement à l'échelle nationale. Le système d'analyse que nous proposons est fondé sur des méthodes de classification et d'apprentissage à l'aide de modèles de Markov cachés (HMM comme *Hidden Markov Model*). Nous décrivons les outils de visualisation que nous avons développés pour permettre aux experts – agronomes, en l'occurrence – de manipuler les données et les résultats du système d'analyse.

Cet article est construit de la façon suivante : nous décrivons tout d'abord la problématique agronomique et les données dont nous disposons puis nous introduisons les modèles de Markov cachés avant de décrire notre démarche. La dernière partie est une discussion.

2 Contexte

La recherche de successions se rattache au problème de la recherche d'épisodes. Agrawal [2, 1] propose des algorithmes non numériques d'extraction de séquences. Nous proposons en revanche d'utiliser des algorithmes numériques d'estimation et de classement mis au point dans notre laboratoire [20] pour la reconnaissance de la parole.

Les études récentes en reconnaissance de la parole [21, 13] montrent qu'un phonème peut être précisément modélisé par un HMM de trois états. On définit ainsi trois états du conduit vocal produisant chacun une distribution de sons élémentaires stationnaires. La dynamique de la parole est alors représentée par une chaîne de Markov du second ordre sur l'ensemble des trois états. Le temps est discrétisé et à chaque pas de temps t, le conduit vocal émet un son élémentaire, puis change d'état en fonction des états occupés aux temps $t-1$ et $t-2$. Le geste intelligent qu'est la production de parole est modélisé par deux processus stochastiques : l'un émet des sons en fonction des états, l'autre effectue les changements d'état.

Les HMM s'appuient sur la théorie de probabilités et de l'estimation statistique. Leur point fort est la possibilité d'autoriser un apprentissage non supervisé par convergence depuis une valeur initiale jusqu'à la maximisation d'un critère « objectif » calculé sur un gros ensemble de données. Le modèle obtenu permet une segmentation en zones stationnaires et transitoires. Ces caractéristiques en font des outils appropriés pour dégager des régularités temporelles ou spatiales comme le montrent les travaux en reconnaissance de la parole [17, 23] et comme celà a été démontré dans différents domaines : segmentation d'images [5], génétique [24, 8], robotique [3], fouille de données [7], aide au diagnostic [11].

Nous avons pu réutiliser plus de 95% des programmes initialement écrits pour le traitement de la parole. Notre effort s'est essentiellement porté sur deux points : 1) l'élaboration d'une démarche d'extraction de connaissances avec des experts et à partir des données et 2) la définition d'outils de visualisation donnant aux experts une vue synthétique des données.

3 Les successions de cultures

3.1 Intérêt de l'étude

L'occupation du territoire agricole change d'années en années, en raison de deux faits majeurs, la libération continue de territoires par la disparition d'exploitations agricoles et l'évolution des systèmes de production.

Le choix de l'ordre des successions des cultures est un résultat du métier d'agriculteur et intègre de nombreux facteurs : dates de récolte et d'implantation des cultures, état laissé par une culture après la récolte et organisation de chantiers entre parcelles.

L'étude de ces successions est un enjeu pour l'agronomie car :

– la nature de ces successions et de leur évolution est un indicateur des dynamiques en cours dans l'agriculture étudiée ;

– ces successions contribuent à des effets environnementaux majeurs : pollution des ressources en eau, structuration des paysages.

Le but de l'étude, pour les agronomes, est de recenser les successions dominantes et leurs évolutions à l'échelle de toute une région. En Lorraine, par exemple, ils s'attendent à retrouver les rotations dominantes, soient : colza-blé-orge, colza-blé-blé, colza-blé, ainsi que ces mêmes rotations incluant le maïs à la place du colza. Ils s'attendent également à observer les effets de la Politique Agricole Commune : introduction de la jachère dans les successions, disparition des prairies permanentes. Du point de vue des informaticiens fouilleurs de données, la recherche de successions de cultures dans les bases de données `Ter Uti` est un bon sujet pour tester les méthodes stochastiques. En effet, si les règles de successions adoptées par les agriculteurs ne sont pas connues avec précision, on admet que dans un système en équilibre sur une parcelle, la règle de succession ne dépend que de l'occupation actuelle de la parcelle et de l'occupation de la ou des deux années précédentes. Cette hypothèse nous permet d'utiliser les modèles de Markov d'ordre un et deux pour traiter les données dont nous disposons. Nous adoptons de fait une attitude bayésienne qui consiste à mesurer par une probabilité l'apparition d'un événement issu d'un processus dont on ne maîtrise pas tous les paramètres.

Ce travail de fouille se fait au profit de plusieurs utilisateurs :

– les agronomes qui peuvent quantifier des connaissances qualitatives, infirmer ou confirmer des hypothèses et élaborer des modèles dynamiques d'attribution de terres dans le paysage ainsi que des modèles de propagation de pollution de l'eau ; ces deux phénomènes étant intimement liés aux successions de cultures pratiquées dans une région ;

– les décideurs des ministères de l'agriculture et de l'environnement ou les enseignants des écoles d'agronomie qui cherchent à avoir une vue d'ensemble de l'évolution des phénomènes agraires.

3.2 Les données `Ter Uti`

La connaissance qu'ont les agronomes des successions de cultures est fondée sur des enquêtes de terrain, auprès des agriculteurs et des services techniques. Elle se fonde aussi sur des enquêtes statistiques de grande envergure, comme l'enquête `Ter Uti` menée par les services statistiques du Ministère de l'Agriculture depuis le début des années 1980.

Ce sont des résultats de cette enquête `Ter Uti` qui constituent notre base de données : nous disposons de résultats sur 23756 points situés en Lorraine et dont l'occupation a été relevée tous les ans, en juin, de 1992 à 1999. Ces points sont répartis dans l'espace de la façon suivante : tous les 4 km on relève 36 points situés sur une grille carrée et séparés les uns des autres de 200 m. La représentativité d'un point est proche de 100 hectares [19]. Nous ne connaissons pas la localisation exacte des points dont nous disposons (secret statistique). Les occupations sont réparties en différentes classes (environ 80) qui vont de « marais salants, étangs d'eau saumâtre » à « peupliers épars » en passant par « superficie en herbe à faible productivité potentielle ». Certaines de ces classes ne sont pas ou peu présentes en Lorraine (« glaciers, neiges éternelles » mais aussi « pomme de terre ») aussi avons nous restreint le nombre de classes à 49, par regroupement ou suppression. Parmi ces classes, nous nous intéressons particulièrement aux occupations agricoles dominantes en Lorraine, c'est-à-dire : maïs, blé, orge, colza, prairies temporaires, prairies permanentes, vergers.

4 Définition d'un HMM

Un modèle de Markov caché est défini par la donnée de :

- $\mathbf{S} = \{O, s_1, s_2, \ldots, s_N, F\}$, un ensemble fini comprenant N états, un état initial `0` et un état final `F` ;
- $\mathbf{A}$ la matrice donnant les probabilités de transition entre états ; $\mathbf{A} = (a_{ij})$ pour un modèle d'ordre 1 (HMM1), $\mathbf{A} = (a_{ijk})$ pour un modèle d'ordre 2 (HMM2) ;
- $b_i(.)$ les lois des densités associées aux états s_i.

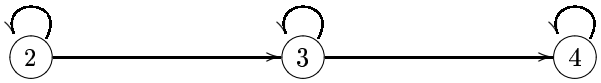
4.1 Différences entre chaînes de Markov et HMM

Une chaîne de Markov possède un ensemble d'états – les cultures d'une parcelle – directement observables. La chaîne de Markov définit un seul processus stochastique. Dans un HMM, l'observation d'une culture n'est pas uniquement associée à un état mais dépend d'une répartition de cultures. Il y a deux processus stochastiques.

– Le premier est caché pour un observateur et est défini sur l'ensemble des états. C'est une chaîne de Markov

d'ordre un ou deux.

– Le deuxième est qualifié de processus visible. Il émet une observation – une culture – à chaque pas de temps en fonction des densités de probabilité définies sur chacun des états. Le premier processus gouverne le second [4]. On considère ainsi que la répartition des cultures dans une région donnée évolue selon un processus de Markov. La répartition à un pas de temps donné ne dépend que de la répartition aux pas précédents suivant l'ordre du modèle. Nous faisons là une analogie avec l'évolution du conduit vocal et la reconnaissance de la parole. L'observation d'une culture correspond à l'observation d'un symbole acoustique élémentaire tel qu'un spectre fréquentiel à court terme calculé sur le signal acoustique. Nous verrons au § 5.2 que l'analogie s'arrête rapidement puisque nous devons traiter le HMM d'une façon complètement différente.



état 2		état 3		état 4	
prairies	0.31	prairies	0.29	blé	0.29
blé	0.22	blé	0.26	prairies	0.27
orge	0.16	colza	0.14	colza	0.17
colza	0.12	orge	0.11	orge	0.12
maïs	0.07	maïs	0.08	maïs	0.06
jachères	0.05	jachères	0.05	vergers	0.02

FIG. 1 – *Modèle 1 : HMM effectuant une segmentation en trois périodes pendant lesquelles les observations sont supposées stationnaires. On remarque la progression du blé et la disparition de la jachère au fil des ans. Les états cachés sont dénotés 2, 3 et 4.*

Nous sommes tout d'abord intéressés par la localisation de segments temporels (périodes d'observation) pendant lesquels la distribution des cultures ne varie pas. Nous nous limiterons donc à des modèles possédant 2 ou 3 états mais autorisant des transitions en boucle sur eux-mêmes comme le montre la figure 1. Ceci revient à étudier le phénomène sur autant de périodes différentes.

Dans un HMM, il n'est plus possible de mesurer la probabilité d'une succession de cultures puisqu'une culture n'apparaît qu'à l'intérieur d'une répartition constituant l'état. Pour pallier ce défaut, dans certains cas, nous introduisons un état uniquement associé aux cultures majoritaires (blé, maïs, orge, ...). Le HMM obtenu possède deux types d'états : les états cachés qui sont des densités d'observations et les états *"de Dirac"* associés à une culture définis par une densité où la probabilité de cette culture vaut 1 et les probabilités de toutes les autres occupations valent 0. Finalement la figure 2 donne la topologie d'un modèle qui sera utilisé par la suite.

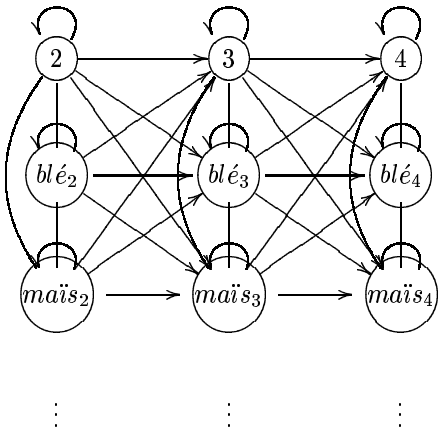


FIG. 2 – *Modèle 2 : Les états notés 2, 3 et 4 sont associés à une distribution de cultures, contrairement aux états dénommés par une occupation. Le nombre de colonnes définit le nombre de périodes d'observation. Les connexions sans flèches représentent des transitions bidirectionnelles.*

4.2 Estimation automatique d'un HMM

Une fois donnés un corpus de données et la topologie du graphe des transitions entre états, différents algorithmes permettent l'apprentissage d'un HMM. Quelque soit l'ordre des modèles, nous utilisons l'algorithme Forward - Backward [22] qui est une variante de l'algorithme EM [15]. L'apprentissage se fait itérativement en partant d'un modèle où toutes les transitions sont équiprobables et où les densités $\mathtt{f}(./\mathtt{s_i})$ sont fixées. L'algorithme Forward - Backward calcule un nouveau modèle plus adapté aux données dans lequel la vraisemblance du corpus a augmenté. Ce nouveau modèle est utilisé dans une nouvelle itération jusqu'à ce que la vraisemblance du corpus atteigne un maximum local. Le résultat est constitué par les nouvelles valeurs des transitions $\mathtt{a_{ijk}}$ et des densités $\mathtt{f}(./\mathtt{s_i})$ (cf. figure 1). Lorsque N est le nombre d'états et T le nombre d'observations, l'algorithme Forward - Backward a une complexité en $\mathtt{N}^3 \times \mathtt{T}$ pour un HMM_2.

Le choix du modèle initial influe sur le résultat final. Pour évaluer l'adéquation du modèle obtenu par convergence, nous utilisons une mesure de distance entre états [27]. Par essais successifs, nous définissons et obtenons un modèle dépendant de plusieurs facteurs : résultats des expériences précédentes, topologie initiale, mode de convergence, critère d'arrêt. Les agronomes n'interviennent pas dans ce travail. Par contre, ils ont un rôle central pour l'interprétation.

5 Quelles mesures sur les successions ?

5.1 Les transitions entre états

Les valeurs des probabilités de transitions entre les états traduisent les durées des périodes temporelles associées. Dans un HMM_1 dont la topologie est donnée par la figure 1, la probabilité $d_j(l)$ que la chaîne stochastique boucle l fois dans l'état j suit une loi géométrique de paramètre a_{jj}.

$$d_j(l) = a_{jj}^{l-1} \times (1 - a_{jj})$$

Dans un HMM_2 de même topologie, où les états successifs sont notés $i = j-1,\ j,\ k = j+1$, la durée de séjour dans l'état j est gouvernée par deux paramètres : la probabilité d'entrer dans l'état j et de n'y rester qu'un pas de temps, et la probabilité d'y séjourner au moins deux pas de temps. La distribution de durée de séjour est représentée par une loi géométrique comme le montrent les équations suivantes :

$$\begin{aligned} d_j(0) &= 0 \\ d_j(1) &= a_{ijk}, \\ d_j(l) &= (1 - a_{ijk}) \cdot a_{jjj}^{l-2} \cdot (1 - a_{jjj}), \quad l \geq 2 \end{aligned}$$

Cette formule se généralise aux modèles de topologie quelconque. Un HMM_2 est capable de modéliser plus précisément la durée des phénomènes stationnaires représenté par un état caché. Cette propriété, que les modèles d'ordre un n'ont pas, est fondamentale lorsqu'on s'intéresse à l'évolution d'un processus stochastique à horizon fini ($T \leq 10$).

Dans le modèle donné dans la figure 2, la proba-

bilité de la succession maïs-maïs-maïs est donnée par plusieurs probabilités de transition suivant les colonnes d'états traversés. Par ordre chronologique on a les successions d'états :

$$maïs_2 - maïs_2 - maïs_2$$
$$maïs_2 - maïs_2 - maïs_3$$
$$maïs_2 - maïs_3 - maïs_3$$
$$maïs_2 - maïs_3 - maïs_4$$
$$\ldots$$

A l'aide des valeurs des probabilités de transition, les experts peuvent étudier l'évolution d'une succession culturale au fil des ans.

La lecture de la matrice **A** qui donne la valeur de la probabilité d'une succession de trois états pour un HMM_2 n'est pas aisée ; aussi préférons nous afficher la variation en fonction du temps de la probabilité *a posteriori* que le système passe de l'état s_i à l'état s_j entre les instants $t-1$ et t. L'algorithme Forward - Backward donne une estimation de $\texttt{Prob}(\texttt{q}_{\texttt{t}-1} = \texttt{s}_\texttt{i}, \texttt{q}_\texttt{t} = \texttt{s}_\texttt{j} / \texttt{observations})$ qui représente l'évolution des probabilités de transitions pendant la période d'étude.

Nous avons créé des outils de visualisation permettant l'affichage de cette probabilité (cf. fig. 3). La largeur et la couleur du trait renseignent sur la valeur de la probabilité de transition. Les états cachés sont représentés par le symbole « ? » mais la distribution de cultures associées peut apparaître en surimpression. L'expert peut choisir de visualiser certains états et fixer le seuil d'affichage des transitions.

5.2 Les répartitions de cultures dans les états

Les HMM effectuent un appariement élastique entre l'ensemble des états et une suite d'observations. Le classement d'une observation temporelle dans un état dépend de l'intégralité de la suite d'observations et de la topologie du modèle. Dans un modèle gauche – droite dans lequel la chaîne de Markov ne peut re-visiter un état qu'elle a quitté, les états identifient des périodes temporelles. Dans le modèle de la figure 2, les états cachés captent toutes les observations qui ne s'alignent pas sur les états *"de Dirac"*. Les états cachés ont, en quelque sorte, un rôle d'états de réserve et permettent de faire une segmentation progressive des données. Les cultures qui apparaissent comme étant les plus fréquentes dans un état ou qui intéressent les agronomes peuvent être alors placées dans un nouvel état prévu à cet effet dans une expérience suivante. C'est ce que nous avons fait pour passer du modèle 1 de la figure 1 au modèle 2 de la figure 2. Cette façon de construire un modèle est originale et constitue la principale différence avec la reconnaissance de la parole où on s'attache peu à trouver un sens aux états.

6 Démarche expérimentale

Nous avons développé une démarche d'*extraction de connaissances à partir de bases de données* (ECBD) sur les points `Ter Uti` en relation directe avec les experts du domaine (agronomes) à l'aide d'outils de visualisation variés, définis en fonction des besoins. Il s'agit d'une démarche d'extraction de connaissances à l'aide d'un environnement intégré [26, 9]. Elle se décline en différentes étapes successives et entremêlées qui ont permis

de mettre à jour les connaissances contenues dans les données et leurs interprétations possibles.

6.1 Étude des suites de cultures

A partir du premier modèle – modèle 1 – donné dans la figure 1, la première étape a consisté, avec les agronomes, à déterminer les occupations à examiner prioritairement dans l'ensemble des occupations de la base `Ter Uti` ; c'est-à-dire à la fois les occupations les plus fréquentes et les plus instables — *a priori* : blé, orge, maïs, jachère, colza et prairies. Nous nous sommes aidés des tables données figure 1.

Ceci nous a permis de construire le modèle 2, semblable à celui de la figure 2, dans lequel on étudie les transitions entre les états associés au blé, orge, maïs, jachère, colza et prairie. Les états 2, 3 et 4 jouent le rôle d'état de réserve.

Les valeurs de $\texttt{Prob}(q_{t-1} = s_i,\ q_t = s_j \ /\ 0)$ sont facilement lisibles et interprétables telles qu'elles sont présentées sur la figure 3. Les experts ont reconnu les successions de cultures majoritaires représentées par des lignes brisées dans l'espace des trajectoires d'états, à savoir : colza-blé-orge, colza-blé, et la monoculture blé-blé. Ils observent une augmentation de la part de ces successions, au détriment de la variabilité préexistante : ils ne s'attendaient pas à cette « simplification du monde ». En revanche, ils peuvent expliquer en 1993 le passage de ces successions dans l'état de réserve (noté «?» sur la figure 3) : cet état représente majoritairement les terres en jachère et on retrouve donc là l'effet de la Politique Agricole Commune en matière de gel des terres. Lors de ce passage en jachère les experts observent également que l'orge disparaît au profit du blé et du colza. Ils expliquent ce choix général par le fait que le blé et le colza sont économiquement plus intéressants que l'orge.

6.2 Étude des suites de successions

Le modèle a été mis en œuvre sur différents jeux de données (représentant différentes petites régions) et a donné des résultats similaires à ceux de la figure 3. À partir de ces résultats, les experts se sont intéressés à l'évolution et aux transitions entre successions avec pour hypothèse, issue de l'interprétation de la figure 3, une simplification des types de successions.

Nous avons alors tenté de déterminer les successions majoritaires en fixant leur taille à 3. Cette valeur n'est pas arbitraire mais correspond à une connaissance du domaine : dans la presque totalité des cas, les successions s'organisent en rotation avec des têtes de rotation qui reviennent au plus tous les 3 ans. Tous les triplets possibles de la base sont considérés (6 triplets pour 8 années, donc 81915 pour 23756 points, mais seulement 1109 triplets différents). Chaque triplet constitue une observation de trois occupations successives faites pendant la période 1992 – 1997. Les résultats de la classification ont été montrés sous forme de simples tableaux aux experts qui y ont trouvé plusieurs intérêts :

- repérer les successions dans les triplets, c'est-à-dire regrouper les différentes permutations : on vérifie ainsi que (colza blé orge), (blé orge colza) et (orge colza blé) ont à peu près la même représentation dans chacun des états ; de même pour (blé colza blé) et (colza blé colza) ;
- repérer et évaluer les successions majoritaires : on vérifie que deux successions (colza blé orge) et (colza blé) représentent une grosse partie des terres cultivées (environ 28%) ;

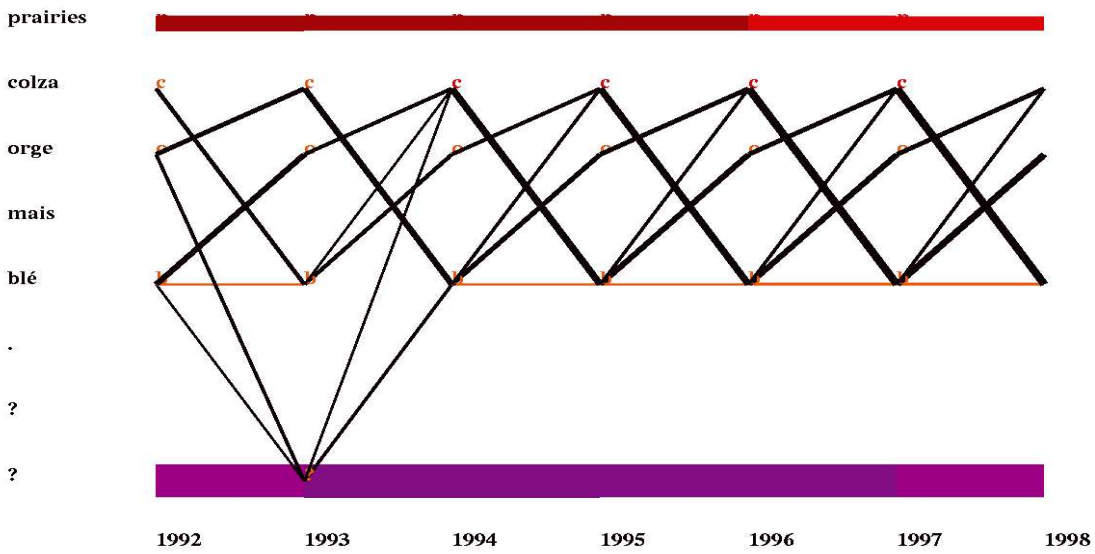


FIG. 3 – *Résultats du modèle 1 pour une partie des données Lorraine (Argonne Meusienne) entre 1992 et 1998 : les probabilités de transitions d'un état à l'autre sont représentées en fonction du temps. L'épaisseur du trait est proportionnelle à la valeur de la probabilité. La ligne dénotée « ? » correspond à un état de réserve affiché. Les prairies sont largement dominantes par rapport aux différentes cultures. On observe une progression des transitions blé-orge-colza.*

– étudier la progression, l'apparition ou la disparition des différentes successions dans la période considérée.

À l'issue de cette analyse, les experts ont défini les successions à étudier davantage, qu'ils ont réparties en grandes classes (colza + 2 céréales, colza + 1 céréale, maïs + 2 céréales, maïs + 1 céréale, monocultures). Nous avons alors, comme lors de la première étape, construit un nouveau modèle – appelé modèle 3 – dans lequel des états n'émettent que ces triplets et leurs permutations circulaires (cf. figure 5). Les résultats de ce modèle peuvent être présentés de la même façon que sur la figure 3 : les trajectoires entre états mises en évidence dans la figure 3 deviennent des lignes droites dans la figure 4. L'observation de cette représentation conduit les experts à confirmer leur interprétation, à savoir une simplification du monde, avec croissance de la part des successions principales (colza + 2 céréales , colza blé) et augmentation de la monoculture (blé sur blé et maïs sur maïs). La figure 4 montre une partie de ces conclusions. On voit nettement la progression de la succession colza + 2 céréales.

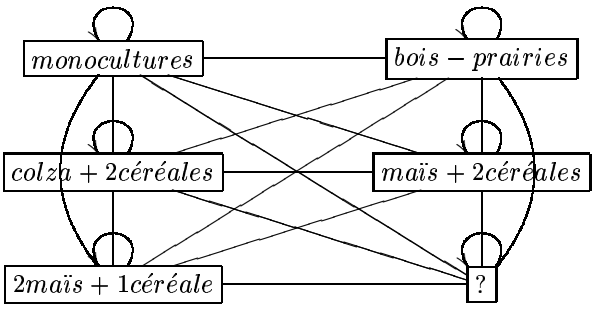


FIG. 5 – *Modèle 3 : Topologie d'un modèle de triplets. Un état particularise une succession de trois cultures définie a priori par l'expert. Toutes les transitions sont bidirectionnelles.*

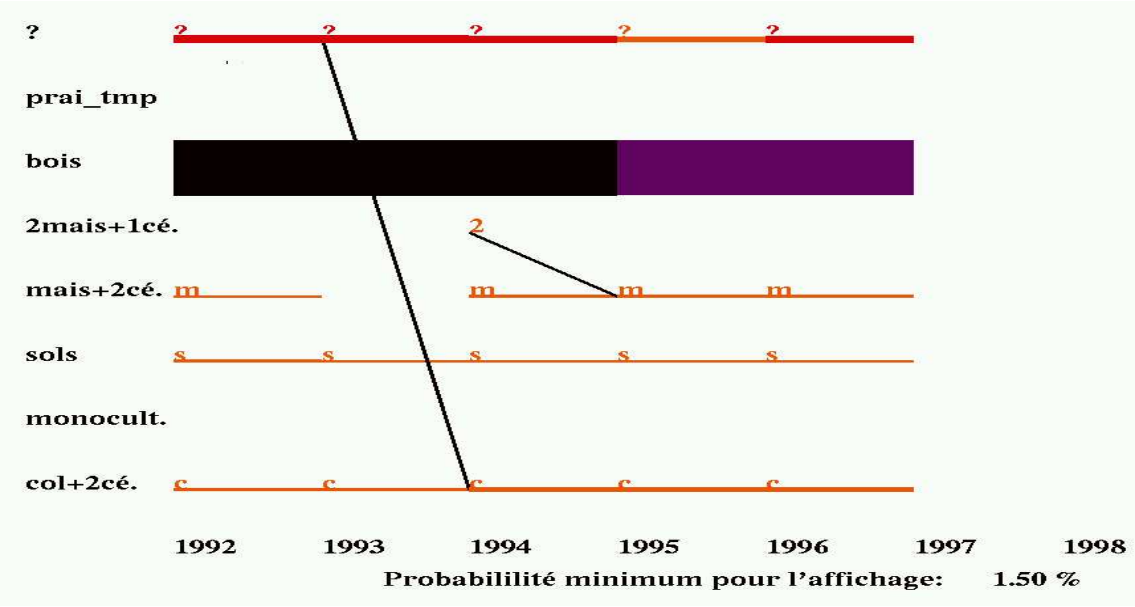


FIG. 4 – *Résultats obtenus à partir du modèle 3 pour une partie des données lorraines (Argonne Meusienne) entre 92 et 99. Une observation est une succession de trois cultures. Les transitions entre successions sont indiquées par les lignes obliques, les variations de l'importance d'une même succession sont représentées par l'épaisseur du trait horizontal. On observe une progression des successions colza + 2 céréales.*

6.3 Segmentation spatiale

Jusqu'à présent, nous n'avions classé les points qu'en fonction de leurs prédécesseurs et successeurs temporels. Le processus de classification n'utilisait pas l'information spatiale du point `Ter Uti`. En introduisant une relation d'ordre sur les points du plan, on passe d'un problème de segmentation 2D à un problème de segmentation 1D. Les HMM permettent alors d'effectuer une segmentation en régions géographiques homogènes du point de vue des densités spatiales de cultures.

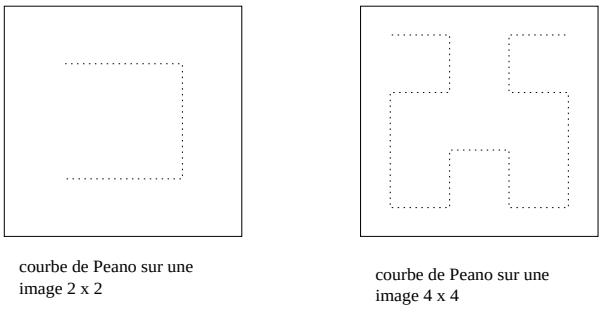


FIG. 6 – *Définition de la courbe fractale de Peano.*

L'utilisation de la courbe de Peano qui parcourt tous les points du plan en respectant la notion de voisinage spatial [5] nous a permis de réaliser une segmentation spatiale des données : un point n'est plus classé en fonction de ses précédents temporels, mais en fonction de ses précédents spatiaux selon la courbe de Peano. Avec un changement minime dans le programme d'apprentissage, nous avons traité l'ensemble des données, chaque année individuellement et toutes les années ensemble. Nous avons traité les données brutes (un point = une culture) et les données des successions (un point = une succession de cultures sur trois années). Ce sont ces der-

nières qui ont fourni les résultats les plus stables.

FIG. 7 – *Image satellitaire de la Lorraine. À l'aide de cette carte, on peut effectuer une localisation des régions homogènes de la figure 8.*

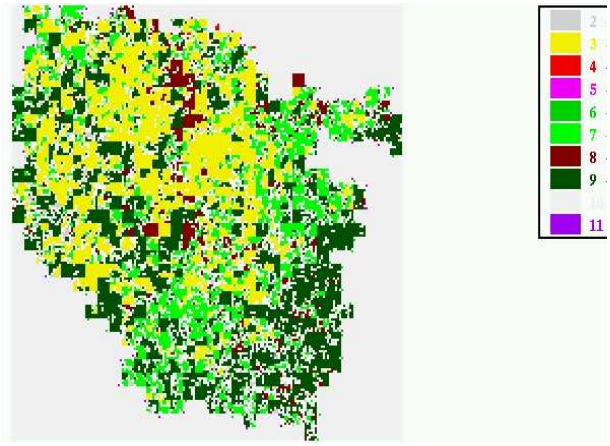


FIG. 8 – *Les successions culturales de la Lorraine. Classification obtenue par un HMM_2 à 10 états après 10 itérations.*

Un modèle où tous les états sont des distributions uniformes sert de modèle initial pour l'algorithme d'apprentissage Forward-Backward. L'appariement entre les sites `Ter Uti` (occupation du sol + localisation spatiale) et les états du modèle définissent des régions homogènes. Ce processus de classification est non supervisé ; l'utilisateur définissant seulement le nombre initial d'états correspondant aux nombres maximum de régions à découvrir.

Nous obtenons des cartes où se distinguent cinq états principaux, les autres états provenant d'une division abusive par l'algorithme :

un état – état 8 – à majorité de bâti (30 %), forêt, sols nus, zones humides qui suit le cours des grandes vallées ;

un état – état 9 – à majorité de forêt (98 %) qui englobe le massif vosgien et les forêts de Meuse ;

un état – état 2 – à majorité de prairies (30 %), forêt (20 %) et cultures fourragères (6 %) caractéristique des régions d'élevage ;

un état – état 3 – à majorité de cultures (30 % pour blé, colza, et orge) et prairie (10 %) définissant les régions céréalières ;

un état – état 7 – à majorité de prairies (68 %) et prés vergers (5 %) au fond des vallées vosgiennes et en pieds de côtes.

On retrouve globalement la localisation des occupations affichées sur la carte de la figure 7 telles qu'elles sont trouvées sur une image satellitaire à une échelle 4 fois plus fine. La localisation de ces états est interprétée par les experts en fonction de la géologie et des caractéristiques techniques des exploitations. Ainsi on retrouve les régions géologiques des Vosges gréseuses, des Vosges granitiques, du plateau lorrain (calcaire), des vallées de la Moselle et de la Meuse, des vallées vosgiennes, des plaines argileuses. On trouve aussi la distinction, sur le même terrain géologique (argiles de Keuper), entre les petites exploitations laitières du Chatenois et les grandes structures mixtes du Saulnois. Cette der-

nière distinction, relevée par un expert, méritera d'être confortée par une étude sur les types d'exploitations agricoles en fonction des petites régions agricoles. L'interaction avec les experts débouche ainsi sur la spécification d'une nouvelle classification dans laquelle nous devons croiser deux sources différentes de données.

6.4 Validation des modèles

Les aller-retour entre experts agronomes et informaticiens se sont traduits par l'élaboration de plusieurs modèles pour valider les hypothèses des experts. Nous nous sommes aidés de la durée des états comme révélatrice de l'instabilité des successions ainsi que du contenu de l'état de réserve qui capte toutes les exceptions et permet d'effectuer une segmentation progressive des données. À partir d'un modèle simple – le modèle 1 – donné dans la figure 1, nous avons élaboré les modèles 2 et 3 pour mesurer plus précisément des choix de successions. La figure 4 montre un système quasiment en équilibre. Les successions dominantes ont été trouvées, elle sont représentées par les états faiblement connectés entre-eux.

La carte issue d'une segmentation 2D est un document porteur d'une information de synthèse appréciée des experts qui la comparent spontanément à d'autres cartes et ainsi la valident en partie. Confrontés à ces résultats, les experts remettent à jour, confirment ou infirment leurs connaissances.

La validation passe aussi par la confrontation des successions trouvées sur des zones géographiques (régions agricoles) avec les informations en provenance de sources : typologie d'exploitation agricole, enquêtes, cartes de différentes données pedo-morphologiques (relief et sol). Notre prochaine étape sera d'utiliser des HMM pour obtenir une classification spatiale et temporelle et l'expliquer en la reliant à des caractéristiques locales connues.

Conclusions et perspectives

Nous avons extrait et quantifié les successions de cultures à l'aide de modèles de Markov cachés d'ordre deux. Les segments temporels et les zones spatiales sont décrits par des répartitions de cultures et montrent les évolutions quantitatives à la fois des cultures et des successions. Grâce à la construction de plusieurs modèles stochastiques et leur estimation à partir d'un gros corpus de points `Ter Uti`, nous avons trouvé un ensemble quasiment stables de successions. Des experts agronomes ont pu expliquer l'évolution et la localisation par plusieurs arguments politico-économiques. Ce travail de fouille de données a été accompli grâce à l'utilisation de logiciels de spécification de modèles et de visualisation de résultats étudiés pour la circonstance avec les agronomes.

Ces résultats montrent que les HMM sont des outils d'extraction de régularités temporelles et spatiales prometteurs et qu'ils possèdent leur place dans un environnement de fouille de données spatio-temporelles. Une suite logique de ce travail consistera à croiser la classification spatiale et la classification temporelle en effectuant une classification spatio-temporelle qui fera émerger des classes ayant une cohérence à la fois spatiale et temporelle.

Enfin, il faut souligner que, pour les agronomes, les successions culturales sont des objets de recherche centraux [25] révélateurs du métier d'agriculteur, mais qui ont donné lieu à peu de publications. L'intérêt de notre

étude est donc double : d'une part apporter aux agronomes des informations générales qui valident et complètent les informations connues sur le terrain ; d'autre part leur donner les moyens d'avoir « en routine » des informations sur les successions culturales, à partir de bases de données statistiques. Ainsi, nos outils ont pu être utilisés par un étudiant [12] sur les données du bassin de la Seine. Finalement ces connaissances pourront entrer dans les modèles actuellement en cours de développement [18, 6] qui permettront d'effectuer des simulations prospectives sur l'occupation agricole du sol et ses effets en matière d'environnement.

Remerciements

Nous remercions le service régional des statistiques agricoles de la DRAF Lorraine pour l'accès aux données `Ter Uti`.

Références

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen et A.I. Verkano. Fast discovery of association rules. In U. MI Fayyad, éditeur, *Advances in Knowledge Discovery and Data Mining*, pages 307 – 328. AAAI Press, 1996.

[2] R. Agrawal et R. Srikant. Mining sequential pattern. In *Eleventh Int. Conf. on Data Engineering (ICDE'95)*, pages 3 – 14, 1995.

[3] O. Aycard, F. Charpillet, D. Fohr et J.-F. Mari. Place Learning and Recognition Using Hidden Markov Models. In *Proceedings IEEE-RSJ on International Conference on Intelligent Robots and Systems*, pages 1741 – 1746, Grenoble, France, Septembre 1997.

[4] J. K. Baker. Stochastic Modeling for Automatic Speech Understanding. In D.R. Reddy, éditeur, *Speech Recognition*, pages 521 – 542. Academic Press, New York, New-York, 1974.

[5] B. Benmiloud et W. Pieczynski. Estimation des paramètres dans les chaînes de Markov cachés et segmentation d'images. *Traitement du signal*, 12(5):433 – 454, 95.

[6] M. Benoît et F. Papy. Pratiques agricoles et qualité de l'eau sur le territoire alimentant un captage. In *L'eau dans l'espace rural*, pages 323–338. INRA, 1997.

[7] D. J. Berndt. Finding Patterns in Time Series . In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth et R. Uthurusamy, éditeurs, *Advances in Knowledge Discovery and Data Mining*, pages 229 – 248. AAAI Press / The MIT Press, 1996.

[8] L. Bize, F. Muri, F. Samson, F. Rodolphe, S. Dusko Ehrlich, B. Prum et P. Bessières. Searching Gene Transfers on Bacillus Subtilis Using Hidden Markov Models. In *RECOMB'99*, 1999.

[9] R.J. Brachman et T. Anand. The Process of Knowledge Discovery in Databases. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth et R. Uthurusamy, éditeurs, *Advances in Knowledge Discovery and Data Mining"*, pages 37–57, Menlo Park, California, 1996. AAAI Press / MIT Press.

[10] R.J. Brachman, P.G. Selfridge, L.G. Terveen, B. Altman, A. Borgida, F. Halper, T. Kirk, A. Lazar, D.L. McGuinness et et L.A. Resnick. Integrated support for data archaeology. *Internatio-*

nal Journal of Intelligent and Cooperative Information, 1993.

[11] L. Bréhélin, O. Gascuel et G. Caraux. Apprentissage de séquences de vecteurs booléens à l'aide de Modèles de Markov Cachés avec Patterns. Application au test de circuits intégrés. In *Conférence d'apprentissage*, pages 25 – 35, 1999.

[12] M. Caty. Évolution des pratiques agricoles et liens avec l'évolution de la qualité de l'eau dans le bassin de la Seine. Mémoire de fin d'étude ENGEES, 1999. INRA SAD Mirecourt.

[13] C. Cerisara, J.-P. Haton, J.-F. Mari et D. Fohr. A Recombination Model for Multi-band Speech Recognition. In *Proceedings IEEE-ICASSP*, Seattle, USA, 1998.

[14] M.-P. Chouvet, F. Le Ber, J. Lieber, L. Mangelinck, A. Napoli et A. Simon. Analyse des besoins en représentation et raisonnement dans une représentation à objets – L'exemple de Y3. In *LMO'96*, pages 150–169, Leysin, Suisse, octobre 1996. EPFL.

[15] A.P. Dempster, N.M. Laird et D.B. Rubin. Maximum-Likelihood From Incomplete Data Via The EM Algorithm. *Journal of Royal Statistic Society, Ser. B (methodological)*, 39:1 – 38, 1977.

[16] W.J. Frawley, G. Piatetsky-Shapiro et C.J. Matheus. Knowledge discovery in databases: An overview. *The AI Magazine*, 1992.

[17] F. Jelinek. Continuous Speech Recognition by Statistical Methods. *IEEE Trans. on Acoutics, Speech and Signal Processing*, 64(4):532 – 556, April 1976.

[18] F. Le Ber et M. Benoît. Modelling the spatial organisation of land use in a farming territory. Example of a village in the "Plateau Lorrain". *Agronomie: Agriculture and Environment*, 18:101–113, 1998.

[19] M. Ledoux et S. Thomas. De la photographie aérienne à la production de blé. *Agreste, la statistique agricole*, 5, juillet 1992.

[20] J.-F. Mari. Reconnaissance de mots enchaînés à l'aide de modèles markoviens discrets. In *Actes Congrès AFCET Reconnaissance des Formes et Intelligence Artificielle*, pages 859 –867, Grenoble, Novembre 1985.

[21] J. F. Mari. *Perception de signaux complexes et interaction homme-machine.* Habilitation à diriger des recherches, Université Henri Poincaré - Nancy 1, 1996.

[22] J.-F. Mari, J.-P. Haton et A. Kriouile. Automatic Word Recognition Based on Second-Order Hidden Markov Models. *IEEE Transactions on Speech and Audio Processing*, 5:22 – 25, Janvier 1997.

[23] J.-F. Mari et A. Napoli. Modèles stochastiques pour la classification de signaux temporels. In *Actes des cinquièmes rencontres de la société francophone de classification*, pages 51 – 54, Lyon, France, Septembre 1997.

[24] F. Mury. *Comparaison d'algorithmes d'identification de chaînes de Markov cachées et application à la détection de régions homogènes dans les séquences d'ADN.* Thèse de doctorat, Université René Descartes, Paris V, 1997.

[25] M. Sébillotte. *Encyclopedia Universalis*, chapitre Les systèmes de culture. 1988.

[26] M. Stefik. *Introduction to Knowledge Systems.* Morgan Kaufmann Publishers, Inc., 1995. San Francisco, California.

[27] J. T. Tou et R. Gonzales. *Pattern Recognition Principles.* Addison-Wesley, 1974.