Classification de successions culturales par modèles de Markov

Jean-François Mari, Florence Le Ber, Marc Benoît

Classification de successions culturales par modèles de Markov

Jean-François Mari (1), Florence Le Ber (2,1) et Marc Benoît (3)

(1) UMR 7503 LORIA, B.P. 239, 54506 Vandœuvre-lès-Nancy
(2) INRA LIAB, Forêt d'Amance, 54280 Champenoux
(3) INRA SAD, Domaine du Joly, 88400 Mirecourt

Résumé

Nous nous proposons d'étudier les successions culturales pratiquées en Lorraine, afin d'intégrer cette connaissance dans un modèle d'organisation spatiale de territoires agricoles en cours de développement à l'INRA. Pour réaliser cette étude, nous utilisons des données `Ter Uti` qui constituent un relevé de l'utilisation du territoire depuis une vingtaine d'années. Ces données sont traitées avec des algorithmes d'apprentissage développés au LORIA pour la reconnaissance de la parole. Ces algorithmes s'appuient sur les modèles de Markov d'ordre supérieur qui permettent de représenter des observations temporelles 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. Nous montrons dans cet article quel est l'intérêt de cette approche sur une étude sur les bases de données lorraines.

Mots-clés occupation du sol, successions culturales, données `Ter Uti`, modèles stochastiques, modèles de Markov.

Introduction

La libération continue de territoires par la disparition d'exploitations agricoles et l'évolution des systèmes de production induit des changements de l'occupation du sol. Ces changements peuvent être responsables de problèmes tels que : pollution, érosion, évolution du paysage. La compréhension de l'organisation des territoires agricoles et le suivi de leur dynamique sont des points-clés pour prévoir ces problèmes et raisonner la mise en œuvre par les agriculteurs des changements nécessaires à leur résolution.

Un des points d'entrée à la compréhension de l'évolution du territoire agricole est la connaissance des successions de cultures pratiquées par les agriculteurs. Le choix des successions est en effet lié à différents critères qui évoluent au cours du temps : impossibilités techniques, contraintes imposées par la Politique Agricole Commune (par exemple, la mise en jachère d'une partie des terres labourées), circonstances particulières dues à la météorologie de l'année, etc.

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 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, la prochaine culture d'une parcelle ne dépend que de l'occupation actuelle de la parcelle et de l'occupation de la ou des deux années précédentes. Nous verrons plus loin que cette hypothèse nous permet d'utiliser les modèles de Markov d'ordre un et deux pour traiter les données dont nous disposons.

Les modèles de Markov cachés (HMM comme Hidden Markov Model) permettent un alignement élastique entre un nombre quelconque d'observations temporelles et un ensemble d'états définis *a priori*. Ils sont utilisés depuis 20 ans en reconnaissance de la parole [Jelinek, 1976] et commencent à être utilisés en génétique pour la recherche de séquences de gènes dans les molécules d'ADN ainsi qu'en robotique et plus récemment en fouille de données temporelles [Berndt, 1996; Mari et Napoli, 1997].

Ces modèles sont issus du domaine des probabilités et des statistiques qui les ont dotés d'algorithmes d'apprentissage automatique à partir de gros corpus de données. Alors que les informaticiens qui spécifient des systèmes fondés sur une représentation des connaissances essaient de constituer des bases de règles acquises auprès d'un expert, l'apprentissage d'un HMM est réalisé 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 qu'un expert peut expliquer. Nous adoptons de fait une attitude bayésienne qui consiste à mesurer par une probabilité l'apparition d'un évènement issu d'un processus dont on ne maîtrise pas tous les paramètres.

Dans cet article, nous présentons une approche de classification de données temporelles avec des chaînes de Markov pour la reconnaissance et l'étude des successions de cultures. Nous avons utilisé pour cela un lot de données `Ter Uti` sur la Lorraine. Nous allons décrire successivement la problématique agronomique, les modèles de Markov, les données dont nous disposons et la méthode que nous avons employée, avant de donner quelques résultats sur des exemples particulièrement pertinents pour la Lorraine.

1 Matériel et méthode

1.1 Les données `Ter Uti`

Nous disposons de données `Ter Uti` qui représentent 16383 points situés en Lorraine et dont l'occupation a été relevée de 1992 à 1998. 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. 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 la catégorie « 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 d'hiver, orge d'hiver, colza, prairies temporaires, prairies permanentes, vergers.

1.2 Les modèles de Markov cachés d'ordre deux

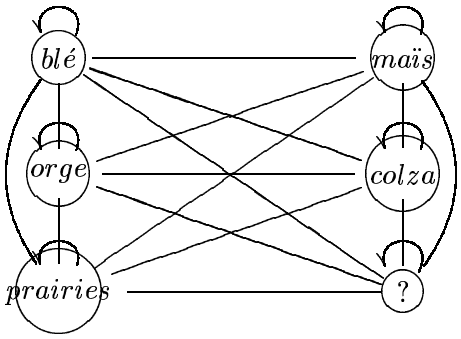
Un modèle de Markov caché d'ordre 2 (HMM_2) [Mari *et al.*, 1997] est défini par la donnée de deux processus stochastiques. Le premier est une chaîne de Markov d'ordre 2 définie par :

- $\mathbf{S} = \{\mathtt{s}_1, \mathtt{s}_2, \ldots, \mathtt{s}_N\}$, un ensemble fini de N états ; $N = 6$ dans l'exemple de la figure 1 ;
- $\mathbf{A} = (\mathtt{a}_{ijk})$ la matrice donnant les probabilités de transition entre états $s_i \rightarrow s_j \rightarrow s_k$;
- π le vecteur de probabilités initiales des états.

On notera $q_1, q_2, \ldots q_T$ la chaîne de Markov de durée T.

Le second processus stochastique, défini sur un intervalle de temps $[\mathtt{1}, \mathtt{T}]$ – [1992, 1998] dans notre étude – est une suite de variables aléatoires $\mathtt{O}_1, \ldots \mathtt{O}_T$ dont les réalisations sont les occupations d'un point `Ter Uti`. Lorsque la chaîne de Markov $q_1, q_2, \ldots q_T$ se trouve dans l'état s_i à l'instant t ($\mathtt{q}_t = \mathtt{s}_i$) on observe une culture ($\mathtt{O}_t = \mathtt{c}_l$) provenant d'une distribution de cultures dépendant de l'état s_i. On représente chacune des N distributions par une loi discrète sur l'ensemble des 49 occupations des sols comme le montre la table 1(b). On définit par $\mathtt{b}_i(\mathtt{l})$ la probabilité $\mathtt{Prob}(\mathtt{c}_l/\mathtt{s}_i)$ de la culture $\mathtt{c}_l$ – parmi les 49 de notre étude – sur l'état $\mathtt{s}_i$ – parmi les $\mathtt{N}$ du modèle. Certaines lois ne privilégient qu'une seule culture. Dans ce cas la probabilité de cette culture est 1 alors que toutes les autres probabilités d'occupation sont nulles. C'est le cas des états appelés *blé*, *maïs* , *orge*, *colza* et *prairies* représentés dans la figure 1. On peut alors identifier l'état de la chaîne à cette culture.

L'ensemble de ces paramètres $\lambda = \{S, A, \pi, b_i(.)\}$ constitue le modèle de Markov caché λ d'ordre 2.



(a) topologie

0	sols à couvert. boisée	0.59
1	sols artifi. non bâtis	0.12
2	autres sols non bâtis	0.08
3	sols bâtis	0.04
4	jachères	0.04
5	potagers	0.03
⋮	…	

(b) densité de l'état noté « ? » obtenue après apprentissage

FIG. 1 – *Topologie d'un modèle où chaque état est connecté à tous les autres. Cinq états différents n'émettent qu'une culture. Un état général susceptible d'émettre n'importe quelle culture est ajouté afin d'autoriser n'importe quelle succession de cultures.*

1.3 Apprentissage automatique d'un HMM_2

Différents algorithmes permettent l'apprentissage d'un HMM_2, une fois donnés un corpus de données et la topologie du graphe des transitions entre états. Nous utilisons l'algorithme Forward-Backward [Kriouile, 1990; Mari *et al.*, 1997] qui est une variante de l'algorithme EM [Dempster *et al.*, 1977]. L'apprentissage se fait itérativement en partant d'un modèle où toutes les transitions et les distributions de cultures $\mathtt{b}_i(\mathtt{l})$ sont équiprobables. 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 $\mathtt{a}_{\mathtt{ijk}}$ et des densités $\mathtt{b}_{\mathtt{i}}(.)$ (cf. figure 1(b)).

1.3.1 L'algorithme Forward-Backward pour le second ordre

Pour estimer les paramètres d'un HMM2, nous définissons les nouvelles fonctions *forward* et *backward* étendues au deuxième ordre.

La fonction $\alpha_t(j,k)$ définit la probabilité $Prob(O_1,O_2,...,O_t, q_{t-1} = s_j, q_t = s_k/\lambda)$ d'avoir simultanément la suite d'observations partielle $O_1,...,O_t$ et la transition $s_j \rightarrow s_k$ entre les instants $t-1$ et t. Comme pour le premier ordre, $\alpha_t(j,k)$ peut être calculée à l'aide de $\alpha_{t-1}(i,j)$ où $s_i \rightarrow s_j$ et $s_j \rightarrow s_k$ sont deux transitions entre les instants $t-2$ et $t-1$ d'une part et $t-1$ et t d'autre part.

$$\alpha_t(j,k) = \sum_{i=1}^{N} \alpha_{t-1}(i,j).a_{ijk}.b_k(O_t), \quad \begin{array}{lll} 2 & \leq t & \leq T \\ 1 & \leq j & \leq N \\ 1 & \leq k & \leq N \end{array} \tag{1}$$

On définit aussi la fonction $\beta_t(i,j)$, comme étant la probabilité de la suite d'observations depuis $t+1$ à T, connaissant le modèle λ et la transition $s_i \rightarrow s_j$ entre les instants t-1 et t.

$$\beta_t(i,j) = Prob(O_{t+1},...O_T/q_{t-1} = s_i, q_t = s_j,\lambda), \quad \begin{array}{lll} 2 & \leq t & \leq T-1 \\ 1 & \leq i & \leq N \\ 1 & \leq j & \leq N \end{array}$$

Tout comme pour le premier ordre, cette quantité peut être calculée récursivement :

$$\beta_{t+1}(j,k) = \sum_{m=1}^{N} \beta_{t+2}(k,m)a_{jkm}b_m(O_{t+2}), \quad \begin{array}{lll} 1 & \leq t & \leq T-2 \\ 1 & \leq j & \leq N \\ 1 & \leq k & \leq N \end{array} \tag{2}$$

Étant donné un modèle λ et une suite d'observations O, on définit par $\eta_t(i,j,k)$ la probabilité de la transition $s_i \rightarrow s_j \rightarrow s_k$ entre les instants $t-1$ et $t+1$ pendant l'émission de la suite d'observations O.

$$\eta_t(i,j,k) = Prob(q_{t-1} = s_i, q_t = s_j, q_{t+1} = s_k/O, \lambda), \quad 2 \leq t \leq T-1$$

D'après Kriouile [Kriouile, 1990], on peut déduire :

$$\eta_t(i,j,k) = \frac{\alpha_t(i,j)a_{ijk}b_k(O_{t+1})\beta_{t+1}(j,k)}{P(O/\lambda)}, \quad 2 \leq t \leq T-1 \tag{3}$$

Tout comme dans le premier ordre, nous pouvons définir les quantités $\xi_t(i,j)$ et $\gamma_t(i)$:

$$\xi_t(i,j) = \sum_{k=1}^{N} \eta_t(i,j,k) \tag{4}$$

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \tag{5}$$

$\xi_t(i,j)$ représente la probabilité $\texttt{Prob}(\texttt{q}_{\texttt{t}-1} = \texttt{s}_\texttt{i}, \texttt{q}_\texttt{t} = \texttt{s}_\texttt{j}/\texttt{O})$ *a posteriori* que le processus stochastique effectue la transition $s_i \rightarrow s_j$ entre les instants $t-1$ et t connaissant toute la suite d'observations. Ces probabilités sont représentées par des épaisseurs proportionnelles associées aux traits matérialisant les transitions sur la figure 2.

$\gamma_t(i)$ représente la probabilité *a posteriori* que le processus soit à l'état s_i au temps t connaissant toutes les observations.

D'après l'énoncé de l'algorithme EM [Dempster *et al.*, 1977], l'estimée de $\overline{a_{ijk}}$ au sens du maximum de vraisemblance est donnée par la formule :

$$\overline{a_{ijk}} = \frac{\sum_t \eta_t(i,j,k)}{\sum_{k,t} \eta_t(i,j,k)} \tag{6}$$

Dans un HMM2 discret, tel qu'il est défini dans le paragraphe 1.2, les observations sont issues de densités discrètes définies sous la forme d'un tableau de probabilités $\texttt{b}_\texttt{j}(\texttt{l})$. Au sens du maximum de vraisemblance, leurs estimées sont données par les formules :

$$\overline{b_j(l)} = \frac{\sum_{t,\ o(t)=c_l} \gamma_t(j)}{\sum_t \gamma_t(j)} \tag{7}$$

2 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.

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. Ce tri nous a permis de construire un modèle où les états sont « clonés » en plusieurs états représentant les occupations intéressantes et en un état « fourre-tout », représentant l'ensemble des autres occupations (cf. figure 1).

La lecture de la matrice `A` entre les états du modèle donne la valeur de la probabilité d'une succession de trois états pour un HMM_2. Sa lecture n'est pas aisée ; aussi préférons nous afficher la valeur de $\texttt{Prob}(\texttt{q}_{\texttt{t}-1} = \texttt{s}_\texttt{i}, \texttt{q}_\texttt{t} = \texttt{s}_\texttt{j}/\texttt{O})$ (cf. formule 4) en fonction du temps qui représente l'évolution des probabilités de transitions pendant la période d'étude (cf. figure 2). Sur ce schéma, l'expert a 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, blé-blé. Ces successions sont perturbées en 93 par un passage dans l'état « fourre-tout » : cet état représente majoritairement les terres en jachère (ou gel de terre) et on retrouve donc là l'effet de la Politique Agricole Commune. Dans ce passage en jachère on observe également que l'orge disparaît au profit du blé et du colza, économiquement plus intéressants. À partir de ce schéma, l'expert s'est intéressé également à l'évolution et aux transitions entre successions, avec pour hypothèse une simplification des types de successions.

Nous avons alors tenté de déterminer les successions majoritaires en fixant leur taille à 3. Tous les triplets possibles de la base sont considérés (5 triplets pour 7 années, donc 81915 pour 16383 points, mais seulement 1109 triplets différents). Chaque triplet constitue une observation

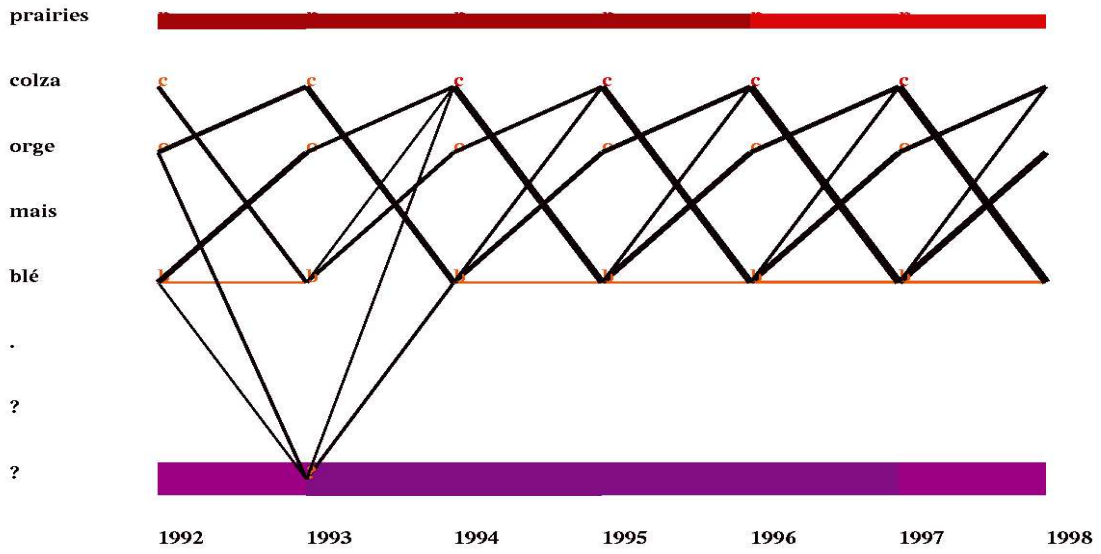


FIG. 2 – *Représentation des probabilités de transitions en fonction du temps dans le modèle de la figure 1. La ligne dénotée « ? » correspond à l'état « fourre-tout ». Les prairies sont largement dominantes par rapport aux différentes cultures (pays de Montmédy dans la Meuse).*

faite pendant la période 1992 – 1996. 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 2 successions (colza-blé-orge, colza-blé) représentent une grosse partie des terres cultivées (environ 28%) ;
- 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 réparti en grandes classes (colza+2céréales, colza+1céréale, maïs+2céréales, maïs+1céréale, monocultures). Nous avons alors, comme lors de la première étape, défini des états n'émettant que ces triplets et leurs permutations circulaires (cf. figure 3). Les trajectoires entre états mises en évidence dans la figure 2 deviennent des lignes droites dans la figure 4. Les probabilités de transition entre ces états constituent les résultats qui sont à soumettre aux experts.

Conclusions et perspectives

Ces résultats montrent que les HMM sont des outils d'extraction de régularités temporelles prometteurs. Nous avons extrait et quantifié les successions de cultures sur trois ans à l'aide de modèles d'ordre deux. Nous avons décomposé la période d'étude en périodes plus courtes pendant lesquelles les cultures sont issues d'une loi dont on peut estimer les paramètres. L'enchaînement

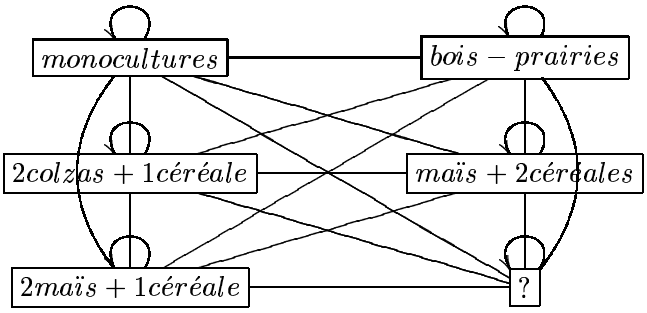


FIG. 3 – *Topologie d'un modèle de triplets. Un état particularise une succession de trois cultures définie a priori par l'expert. Toutes les transitions sont bidirectionnelles.*

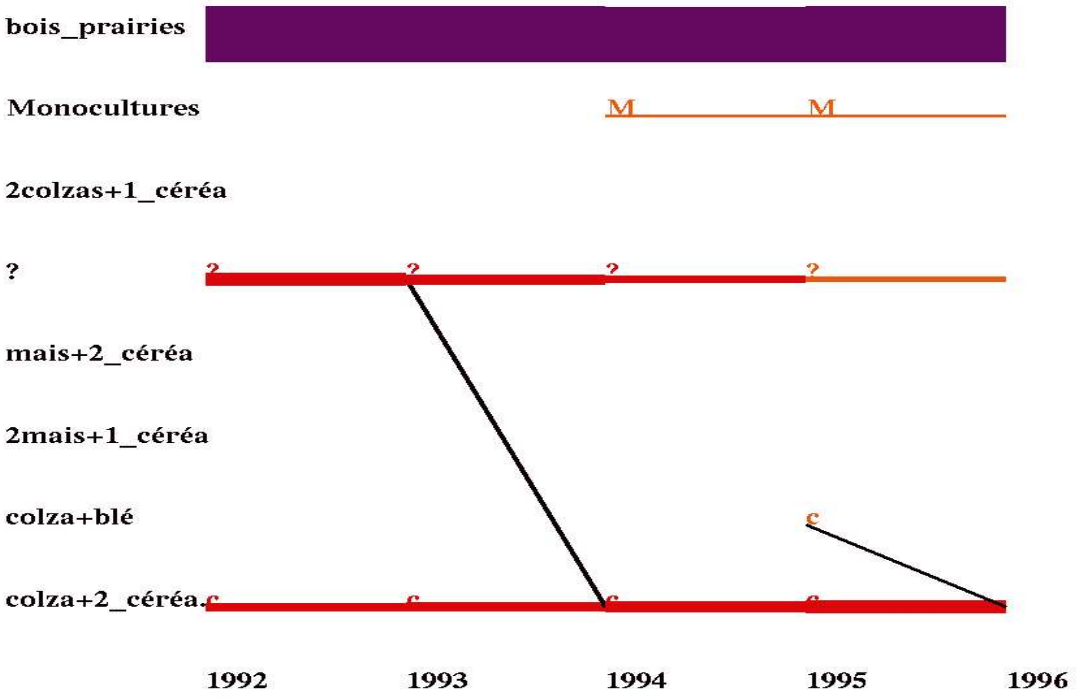


FIG. 4 – *Le modèle de la figure 3 met en évidence deux successions majoritaires : colza+2 céréales et la monoculture dont l'importance va en montant au fil du temps. La transition descendante entre 1993 et 1994 représente la disparition de la jachère au profit des successions de cultures du type colza + 2céréales (pays de Montmédy dans la Meuse). On voit l'état « ? » diminuer au profit des successions majoritaires, signe d'une simplification globale des systèmes de production agricoles.*

des répartitions est une chaîne de Markov et montre les évolutions quantitatives à la fois des cultures et des successions.

Nous travaillerons ultérieurement sur d'autres jeux de données et en segmentant ces jeux de données afin de permettre des regroupements plus aisés. La difficulté consistera alors à choisir un (ou des) modèle(s) adéquat(s) pour effectuer des études sur des régions agricoles cohérentes. Ceci nous conduit à souligner l'intérêt d'un travail qui porterait sur les déterminants spatiaux des successions culturales, afin de les relier à des caractéristiques locales connues.

Enfin, il faut souligner que, pour les agronomes, les successions culturales sont des objets de recherche centraux [Sébillotte, 1988], 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. Finalement ces connaissances pourront entrer dans les modèles actuellement en cours de développement [Le Ber et Benoît, 1998] 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

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

[Dempster *et al.*, 1977] 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.

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

[Kriouile, 1990] A. Kriouile. *La reconnaissance automatique de la parole et les modèles de Markov cachés : modèles du second ordre et distatnce de Viterbi à optimalité locale*. PhD thesis, Université de NANCY 1, 1990.

[Le Ber et Benoît, 1998] 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.

[Mari *et al.*, 1997] 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.

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

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