

CarrotAge, un logiciel pour la fouille de données agricoles

Marc Benoît, Florence Le Ber, Jean-François Mari, Catherine Mignolet,
Céline Schott

► **To cite this version:**

Marc Benoît, Florence Le Ber, Jean-François Mari, Catherine Mignolet, Céline Schott. CarrotAge, un logiciel pour la fouille de données agricoles. Colloque STIC et Environnement SE'2003, INSA Rouen, Jun 2003, Rouen, France, 4 p. inria-00107651

HAL Id: inria-00107651

<https://hal.inria.fr/inria-00107651>

Submitted on 19 Oct 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CARROTAGE, un logiciel pour la fouille de données agricoles

M. Benoît¹

F. Le Ber^{2,3}

J.-F. Mari³

C. Mignolet¹

C. Schott¹

¹ INRA SAD

² CEVH, ENGEES – ULP

³ UMR 7503 LORIA

INRA SAD, Domaine du Joly, F-88400 Mirecourt

{benoit,mignolet,schott}@mirecourt.inra.fr

ENGEES, 1 quai Koch, BP 1039, F-67070 Strasbourg CEDEX

UMR 7503 LORIA, BP 239, F-54506 Vandœuvre-lès-Nancy

{leber,jfmari}@loria.fr

Résumé :

Ce texte présente le logiciel CARROTAGE qui a été développé pour aider à l'analyse de bases de données agricoles. CARROTAGE utilise les modèles de Markov cachés pour segmenter des séquences temporelles ou spatiales en zones homogènes et interprétables. Ce logiciel a été mis en œuvre dans différents projets de recherche qui concernent des problématiques agro-environnementales. Nous présentons deux exemples, l'un en Midi-Pyrénées, l'autre dans le bassin de la Seine.

Mots Clef :

Intelligence artificielle, fouille de données, agronomie.

Abstract:

This paper presents the software CARROTAGE, that has been developed for helping the analysis of agricultural data bases. CARROTAGE is based on high-order Hidden Markov Models and performs the segmentation of spatial or temporal sequences in order to find regularities. This software has been used in several research projects in concern with agro-environmental problems. Two examples are described, in the Midi-Pyrénées region and in the watershed of the Seine river.

Keywords:

Artificial intelligence, data mining, agronomy.

1 Introduction

Nous présentons dans ce texte un logiciel de fouille de données développé dans le cadre d'une collaboration entre agronomes et informaticiens. Ce logiciel, dénommé CARROTAGE, permet aux agronomes d'analyser des bases de données spatiales et temporelles sur l'occupation du sol. Les connaissances ainsi extraites interviennent dans de nombreuses problématiques liant l'uti-

lisation du territoire et la gestion ou la protection de ressources naturelles. Ainsi CARROTAGE est utilisé actuellement dans le cadre de recherches sur la qualité de l'eau dans le bassin de la Seine et dans le cadre d'un projet concernant la gestion de l'irrigation en Midi-Pyrénées.

La méthode de fouille de données mise en œuvre dans CARROTAGE s'appuie sur les modèles de Markov cachés qui sont utilisés dans de nombreux travaux en fouille de données, reconnaissance de la parole ou segmentation d'images. CARROTAGE peut traiter à la fois des signaux temporels ou spatiaux. Il prend en entrée une base de données, qui est constituée par un recueil systématique depuis 20 ans des occupations du sol sur un ensemble de points du territoire français. En sortie, CARROTAGE affiche des cartes et des indices que les agronomes interprètent en termes de successions de cultures et qu'ils peuvent confronter aux connaissances acquises sur le terrain et intégrer dans d'autres modèles.

Ce texte est organisé comme suit. Les deux premières parties précisent les fondements pratiques et théoriques du logiciel CARROTAGE. Les deux suivantes présentent des exemples d'utilisation.

2 Problématique agronomique

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. Cette évolution se traduit principalement dans le changement des successions de cultures, par l'introduction de nouvelles cultures, le raccourcissement des rotations, la prise en compte de la réglementation de la politique agricole commune, etc. L'étude de ces successions est un enjeu pour l'agronomie car :

– la connaissance des successions et de leur évolution

est un outil pour la gestion de certaines ressources de l'agriculture (par exemple l'irrigation) ;

- c'est également un moyen de prévoir et de prévenir des effets environnementaux majeurs tels que la pollution des ressources en eau et la structuration des paysages ;
- à un niveau prospectif, cette connaissance est utile pour analyser les dynamiques en cours dans l'agriculture.

Différents moyens existent pour connaître les successions pratiquées dans une région : enquêtes chez les agriculteurs, recueils d'expertise auprès de techniciens agricoles ou exploitation de bases de données statistiques.

CARROTAGE a été développé pour permettre l'exploitation de la base de données issue de l'enquête *Ter-Uti* sur l'utilisation du territoire [4]. Cette base de données est constituée de la façon suivante. Un premier tirage, réalisé par l'IGN, permet de sélectionner des photos aériennes régulièrement réparties sur l'ensemble du territoire métropolitain. Un deuxième tirage, réalisé par les services régionaux de l'agriculture, consiste à superposer sur chaque photo, une grille de 36 points. Compte tenu de la distance entre les photos, la représentativité d'un point est proche de 100 hectares. L'ensemble de ces points est visité annuellement par des enquêteurs qui en relèvent l'occupation. Environ 80 classes d'occupations du sol sont recensées, mais on peut les réduire à une cinquantaine si on ne considère que les occupations agricoles. Finalement, l'information disponible sur chacun des points est la suivante : la séquence temporelle des occupations depuis 1992, le département et la petite région agricole (PRA) d'appartenance, les points voisins.

3 Principes de la méthode

Le logiciel CARROTAGE a pour but d'aider à l'analyse de données spatio-temporelles. Il s'appuie sur une méthode de classification numérique, les modèles de Markov cachés (HMM comme *Hidden Markov Model*). Un HMM permet de représenter deux processus stochastiques, le premier gouvernant le second [1] :

- le premier processus est caché et défini sur un ensemble d'états auxquels sont associés des densités de probabilité. La transition d'un état à un autre dépend des n états parcourus précédemment suivant l'ordre du modèle.
- le deuxième processus est qualifié de visible. Il émet une observation à chaque pas de temps en fonction de la densité de probabilité associée à l'état courant du processus caché.

Dans le contexte présent, on considère que la distribution des occupations du sol dans une région donnée évolue selon un processus de Markov. La distribution une année donnée dépend de la distribution les années précédentes. On suppose que l'occupation d'une parcelle une année donnée suit la densité de probabilité définie par la distribution des occupations du sol dans la région la même

année. Cette modélisation ignore les raisons qui produisent les occupations du sol observées sur une région, mais permet d'utiliser des algorithmes d'apprentissage rapides sur des bases de données : CARROTAGE utilise l'algorithme Forward - Backward [5].

CARROTAGE traite selon cette approche deux familles de tableaux de données numériques :

- les séquences temporelles d'observations, telles que la suite des relevés annuel des occupations du sol sur un point précis du territoire,
- les séquences spatiales d'observations, telles que le relevé de la répartition spatiale des occupations sur un territoire une année donnée.

Dans les deux cas, la classification d'un point s'effectue en ne prenant en compte que l'occupation actuelle de ce point et son occupation les deux années précédentes ou – pour les séquences spatiales – les occupations de un ou deux points voisins. Ces hypothèses nous permettent d'utiliser les modèles de Markov cachés d'ordre un et deux (HMM1 et HMM2). Un HMM1 ou HMM2 se définit concrètement comme un ensemble d'états, auxquels sont associés des lois de distributions, et une matrice des transitions entre les états. Le choix des paramètres de départ (nombre d'états, transitions autorisées, lois de distribution) permet de segmenter différemment les données et donc de faire apparaître différentes informations. Ainsi, CARROTAGE propose plusieurs types de modèles. En particulier, un état correspond en général à une distribution de cultures (états dits "*de réserve*"), mais nous avons défini des états associés à une seule culture (états dits "*de Dirac*") et des états associés à des distributions de successions de cultures [2, 6].

4 Étude de successions de cultures dans le bassin de la Seine

Depuis plusieurs dizaines d'années, l'hydrosystème du bassin de la Seine s'est progressivement altéré du point de vue de la qualité de l'eau et de ses peuplements biologiques, suite à l'activité humaine domestique, industrielle et agricole [8]. La contamination des eaux souterraines et de surface par les nitrates résulte principalement de l'évolution des activités agricoles, de leur nature mais aussi de la façon dont elles s'organisent dans le territoire du bassin versant. C'est pourquoi le Programme Interdisciplinaire de Recherche en ENvironnement sur la Seine (PIREN-Seine) développe un outil de prévision de la qualité des eaux du bassin en fonction d'hypothèses sur des changements de pratiques agricoles. Dans ce cadre, les chercheurs de l'INRA Mirecourt travaillent à l'analyse de l'agriculture du bassin, de ses dynamiques et de son organisation spatiale [9]. En particulier, ils s'intéressent aux successions de cultures qui, plus que les assolements, constituent une variable explicative des risques de pertes en nitrates. CARROTAGE a

été utilisé pour cela sur les données *Ter-Uti* des différents départements du bassin de la Seine. Nous présentons ci-dessous un exemple de l'analyse menée et des résultats obtenus sur les données d'une petite région agricole de l'Aisne, la PRA St-Quentinoise et Laonnoise [10].

Différents modèles disponibles dans CARROTAGE ont été appliqués sur les données *Ter-Uti* de la PRA St-Quentinoise et Laonnoise dans le but de :

- calculer l'assolement moyen pour une période donnée, ici de 1992 à 1999 (Figure 1, gauche); le HMM utilisé est constitué uniquement d'états de réserve.
- visualiser les principales transitions annuelles entre cultures (Figure 1, droite); le HMM est constitué d'états de réserve et de Dirac (betterave, blé, pois, etc.). Sur le graphe affiché en figure 1, les traits liants les différents états de Dirac sont d'épaisseur proportionnelle à la probabilité de transition entre les cultures.

Ces résultats sont analysés de la façon suivante. On remarque que les principales têtes de rotation – la betterave et le pois – sont majoritairement suivies et précédées par un blé. On peut donc en déduire des “portions” de successions de cultures sur trois ans, de type *blé-betterave-blé* ou *blé-pois-blé*, indiquant des successions quadriennales de type *?-blé-betterave-blé* ou *?-blé-pois-blé*, mais rien ne permet d'affirmer la nature de la tête de rotation suivant ou précédant ce triplet de cultures. Sur le graphe, les traits pointillés indiquent les cultures possibles après le triplet *blé-betterave-blé* : une betterave, un pois, un blé, de l'orge, du colza ou de la jachère.

D'autres modèles ont été utilisés pour rechercher les différentes successions. Ce travail a été mené sur toutes les petites régions agricoles du bassin de la Seine et a permis d'en définir une typologie en fonction des successions pratiquées et de leurs évolutions [10].

5 Aide à l'interprétation d'images satellitaires en Midi-Pyrénées

CARROTAGE a été utilisé dans le cadre d'un projet mené par le centre INRA de Toulouse sur la prévision des besoins en eau pour l'irrigation en Midi-Pyrénées. Habituellement, les besoins en eau sont évalués dans une région en fonction des cartes d'occupation du sol obtenues à partir d'images satellitaires [3]. Cette méthode est cependant problématique car à la date où les prévisions doivent être établies, en début de saison, les images disponibles ne sont pas suffisantes pour cartographier avec certitude les occupations du sol. La connaissance des successions de cultures pratiquées dans la région concernée permet alors de lever certaines ambiguïtés : connaissant l'occupation d'une parcelle l'année précédente, on infère les occupations possibles pour l'année courante, et une image de début de saison peut alors suffire pour conclure.

Dans le projet qui nous occupe, CARROTAGE a permis

d'effectuer une segmentation spatiale des données *Ter-Uti* de la région Midi-Pyrénées pour dégager trois régions homogènes du point de vue des successions de cultures (Figure 2). L'étude s'est ensuite focalisée sur une seule des trois régions, où CARROTAGE a été utilisé pour analyser les évolutions temporelles des successions. Finalement des probabilités de transition entre cultures ont été calculées pour les 10 années couvertes par les données. Ces probabilités sont ensuite utilisées avec une carte d'occupation du sol de l'année $n - 1$ et une image satellitaire de début de saison (avril) de l'année n , afin de produire une estimation de la carte d'occupation du sol à l'année n . L'évaluation de l'estimation obtenue s'appuie sur une carte d'occupation du sol de l'année n établie *a posteriori* [7].

Les trois régions de la figure 2 ont été analysées de façon différenciée :

- au Nord, une première zone (gris clair) contient les cultures. Une segmentation temporelle de cette région révèle trois périodes. On remarque dans les deux tableaux (à droite) la progression de la rotation *blé-tournesol* et de la monoculture du maïs depuis 10 ans ;
- une deuxième zone (gris foncé) dans laquelle la distribution des couverts est stationnaire et constituée de bois, de surfaces en herbe et d'alpages ;
- au Sud, une dernière zone (en noir) contient les massifs montagneux, peu susceptibles d'évolution.

Sur cet exemple, les modèles de segmentation spatiale et temporelle sont utilisés de façon complémentaire. Les premiers permettent de définir des régions homogènes vis-à-vis des successions sur une longue période, puis les seconds permettent une étude plus fine des régions.

6 Conclusion et perspectives

Dans l'état actuel, CARROTAGE permet aux agronomes d'obtenir des tableaux représentant des assolements sur une région et des graphes représentant des transitions entre cultures ou entre successions de cultures. Ces résultats peuvent ensuite être analysés et intégrés dans différents modèles, comme le montrent les exemples présentés ci-dessus. CARROTAGE est ainsi mis en œuvre dans différents projets concernant l'agriculture et l'environnement et permet d'exploiter efficacement des bases de données jusque là peu utilisées.

CARROTAGE peut être amélioré afin d'offrir un outil de fouille de données plus complet et convivial. En particulier, pour aider les agronomes dans leur analyse, il faudrait développer des méthodes pour classer et comparer automatiquement les graphes de transitions entre cultures obtenus sur les petites régions. Des approches qualitatives, classification ou raisonnement à partir de cas, peuvent être utilisées en ce sens.

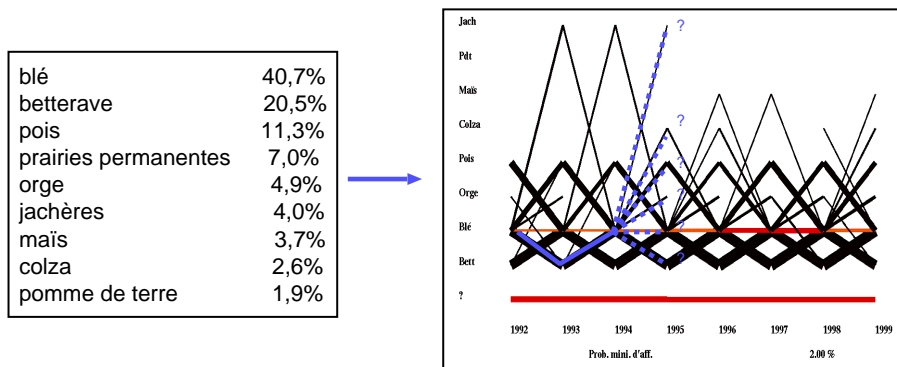


FIG. 1 – Résultats de CARROTAGE pour les données de la PRA St-Quentinoise et Laonnoise (Aisne, années 1992-2000).

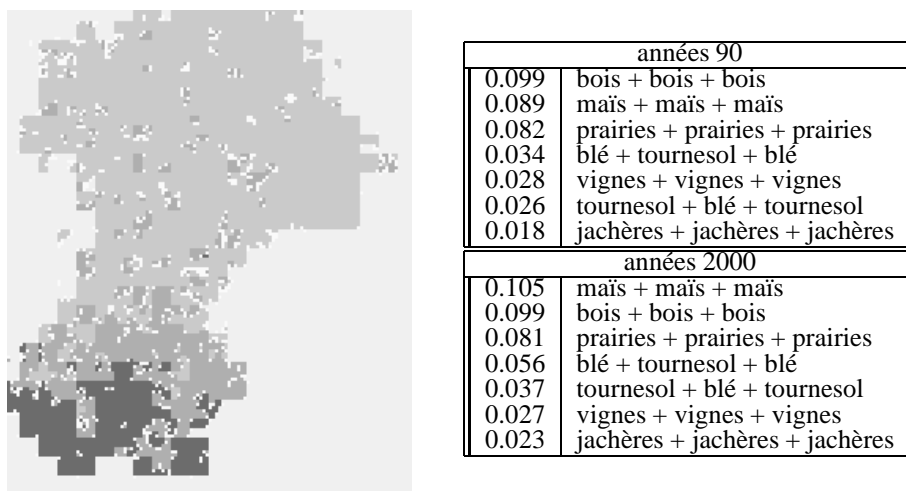


FIG. 2 – Cartographie des successions culturales dans les départements du Gers et des Hautes-Pyrénées et table des évolutions pour la région nord (années 1992-2000).

Références

- [1] Baker (J. K.). – Stochastic Modeling for Automatic Speech Understanding. In : *Speech Recognition*, pp. 521 – 542. – Academic Press, New York, 1974.
- [2] Benoît (M.), Le Ber (F.) et Mari (J.-F.). – Recherche des successions de cultures et de leurs évolutions : analyse des données Ter-Uti en Lorraine. *AGRESTE, la statistique agricole*, n° 31, 2001, pp. 23–30.
- [3] Casterad (M.A.) et Herrero (J.). – Irrivol : A method to estimate the yearly and monthly water applied in an irrigation district. *Water Resources Research*, vol. 34, 1998, pp. 3045–3049.
- [4] Ledoux (M.) et Thomas (S.). – De la photographie aérienne à la production de blé. *AGRESTE, la statistique agricole*, n° 5, 1992.
- [5] Mari (J.-F.), Haton (J.-P.) et Kriouile (A.). – Automatic Word Recognition Based on Second-Order Hidden Markov Models. *IEEE Trans. on Speech and Audio Processing*, vol. 5, 1997, pp. 22 – 25.
- [6] Mari (J.-F.), Le Ber (F.) et Benoît (M.). – Segmentation temporelle et spatiale de données agricoles. *Revue Internationale de Géomatique*, vol. 12, n° 4, 2002, pp. 439–460.
- [7] Mesmin (D.). – Estimation de l’assolement d’un territoire. – Mémoire de DESS, Université Blaise Pascal, Clermont-Ferrand, 2002. 73 p.
- [8] Meybeck (M.), De Marsilly (G.) et Fustec (E.). – *La Seine en son bassin, fonctionnement d’un système fluvial anthropisé*. – Elsevier, 1998. 750 pages.
- [9] Mignolet (C.), Bornerand (C.) et Benoît (M.). – Dynamique spatiale et temporelle de l’activité agricole dans le bassin de la Seine au cours des trente dernières années. *C.R. Acad. Agric. Fr.*, vol. 87, n° 1, 2001, pp. 99–109.
- [10] Mignolet (C.), Schott (C.), Mari (J.-F.) et Benoît (M.). – Typologies des successions de cultures et des techniques culturales dans le bassin de la Seine. Rapport intermédiaire du contrat PIREN-Seine. – INRA SAD Mirecourt, 2003, 23 pages.