

Modélisation stochastique pour l'observation du territoire agricole

Jean-François Mari, Florence Le Ber, Loria, Équipe Orpailleur

21 novembre = présentation de Marie

Présentation de l'exposé

- ▶ Principes et méthodes de la fouille de données
- ▶ Une approche : les modèles de Markov cachés
- ▶ Observation du territoire agricole : données Ter Uti et successions de cultures
- ▶ Étude de successions de cultures dans le bassin de la Seine
- ▶ Aide à l'interprétation d'images satellitaires en Midi-Pyrénées
- ▶ Première analyse du territoire de Chizé (projet ANR:Ecoger)

Fouille de données : le principe

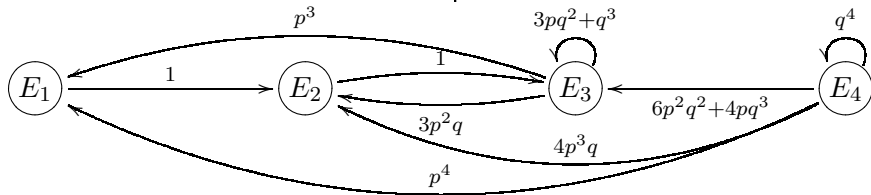
Sur des bases de données quantitatives ou qualitatives, structurées ou non, temporelles, spatiales, etc.

Objectif = Élaborer des indices concis et expurgés du bruit de l'application qui permettent à un analyste d'avoir une vue claire d'un tableau de nombres ou de données qualitatives.

À partir de cette vue, l'analyste peut extraire quelques éléments de connaissance, débiter un raisonnement et formuler d'autres expérimentations susceptibles de compléter sa connaissance du domaine.

Modèles de Markov : exemple

Un groupe de 4 chercheurs effectue des missions dans une campagne lointaine. Certains s'égarent. Soit p la probabilité qu'un chercheur disparaisse. Le groupe ne repart qu'avec 3 ou 4 membres sinon il est mis au repos pendant que le laboratoire lui fournit d'autres chercheurs. Chaque soir, s'il y a 0, 1 ou 2 chercheurs valides, un nouveau chercheur arrive pendant la nuit.



Représentation à l'aide d'une chaîne de Markov ($q = 1 - p$) : les états représentent le nombre de chercheurs chaque matin.

Plus d'incertain : modèles de Markov cachés

On suppose maintenant que les états sont cachés :

On ne peut pas voir les chercheurs, mais on peut accéder aux cuisines de l'hôtel du village, dans lequel on observe une ambiance corrélée à la situation du moment (**quantité de vaisselle, nombre et type de bouteilles vides, etc.**).

À chaque état caché, on attache une distribution des observations qui le caractérise.

L'ensemble constitue un modèle de Markov caché (Hidden Markov Model = HMM)

Définition des HMM

Dans un HMM sont représentés deux processus stochastiques :

Un processus caché = la suite des distributions est un processus de Markov :

- ▶ $\mathbf{S} = \{O, s_1, s_2, \dots, s_N, F\}$ est un ensemble fini comprenant N états, un état initial O et un état final F ;
- ▶ $\mathbf{A}$ est la matrice donnant les probabilités de transition entre états ;

$\mathbf{A} = (a_{ij})$ pour un modèle d'ordre 1 (HMM1) ;

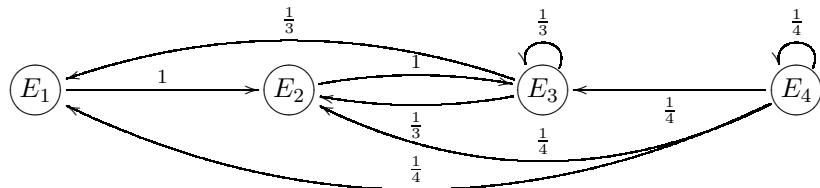
$\mathbf{A} = (a_{ijk})$ pour un modèle d'ordre 2 (HMM2) ;

Un processus observable : les bouteilles vides chaque jour.

- ▶ $\mathbf{B}$ est la matrice représentant les distributions de bouteilles $b_i(\cdot)$ associées aux états s_i .

Apprentissage par maximisation de la vraisemblance : algorithme E.M.

- On fixe un modèle initial.



état 1		état 2		état 3		état 4	
bdx	0.029	bdx	0.029	bdx	0.029	...	
bgne	0.029	bgne	0.029	bgne	0.029	...	
bière	0.029	bière	0.029	bière	0.029	...	
tokay	0.029	tokay	0.029	tokay	0.029	...	
eau	0.029	eau	0.029	eau	0.029	...	
...		...		...		...	

Apprentissage (suite)

- À partir des relevés quotidiens, pendant l'apprentissage, on estime les paramètres du modèle : distributions des états et transitions entre états.

état 1		état 2		état 3		état 4	
bdx	0.31	bdx	0.29	bgne	0.29		
bgne	0.22	bgne	0.26	bdx	0.27		
bière	0.16	cognac	0.14	cognac	0.17		
cognac	0.12	bière	0.11	bière	0.12		
tokay	0.07	tokay	0.08	tokay	0.06		
eau	0.05	eau	0.05	pastis	0.02		

On obtient dans **A** :

les nouvelles transitions entre états

→ quantification des probabilités de successions entre états

et dans **B** :

les nouvelles distributions de bouteilles associées

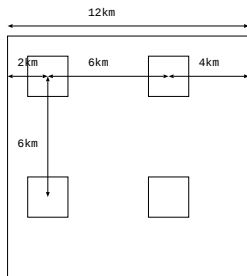
→ classification non supervisée de distributions temporelles de consommations.

Données spatiales et temporelles fournies par les services statistiques du Ministère français de l'Agriculture et de la Forêt qui effectuent un **relevé annuel** de l'occupation du territoire depuis 1992.

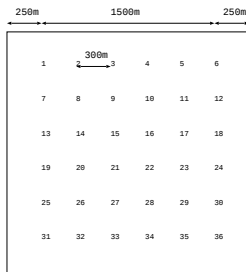
Exemples de catégories Ter Uti d'occupations du sol :

0	Eaux permanentes et zones humides	0.022
1	Sols à roche mère affleurante	0.022
2	Sols à couverture boisée	0.022
3	avoine	0.022
4	seigle	0.022
...	...	0.022
48	zones interdites	0.022

Les données Ter Uti: les photos aériennes



(a) 4 photos aériennes dans une maille



(b) une photo aérienne découpée selon 6x6 points



Les successions culturales effectivement pratiquées par les agriculteurs présentent une grande variabilité temporelle et spatiale. Elles sont mal connues des agronomes, alors qu'elles constituent un élément important du fonctionnement des territoires agricoles :

- ▶ le choix des successions des cultures relève des pratiques et contraintes des agriculteurs ;
- ▶ la nature de ces successions et de leur évolution est un indicateur des **dynamiques** en cours dans une région agricole ;
- ▶ les successions contribuent à des effets environnementaux majeurs :
 - ▶ l'érosion ;
 - ▶ la dégradation des qualités d'eaux souterraines ;
 - ▶ l'utilisation des ressources en eau (irrigation) ;
 - ▶ l'entretien des paysages.

Modélisation stochastique de l'occupation des sols : les principes

- ▶ La modélisation stochastique représente par une probabilité notre incertitude sur la réalisation d'un phénomène temporel et/ou spatial dont on cherche à extraire des régularités ou des **classes** pour le comprendre et **raisonner** dessus.

Modélisation stochastique de l'occupation des sols : les principes

- ▶ La modélisation stochastique représente par une probabilité notre incertitude sur la réalisation d'un phénomène temporel et/ou spatial dont on cherche à extraire des régularités ou des **classes** pour le comprendre et **raisonner** dessus.
- ▶ Technique de modélisation éprouvée en reconnaissance de la parole. “Une région nous parle”. Sa production “sonore” est représentée par la suite des occupations spatiales qui est une suite temporelle d'images où les sites spatiaux sont corrélés spatialement et temporellement.

Modélisation stochastique de l'occupation des sols : les principes

- ▶ La modélisation stochastique représente par une probabilité notre incertitude sur la réalisation d'un phénomène temporel et/ou spatial dont on cherche à extraire des régularités ou des **classes** pour le comprendre et **raisonner** dessus.
- ▶ Technique de modélisation éprouvée en reconnaissance de la parole. “Une région nous parle”. Sa production “sonore” est représentée par la suite des occupations spatiales qui est une suite temporelle d'images où les sites spatiaux sont corréllés spatialement et temporellement.
- ▶ exp : pas de maïs à proximité de la forêt, pas de succession temporelle parking + forêt + blé, ...

- ▶ CarottAge = carotte (qui se traduit en russe par markov) + âge (car il est question de datation).
- ▶ Développé en collaboration entre le LORIA (J.-F. Mari, F. Le Ber) et l'INRA (M. Benoît).
- ▶ exploite la base de données Ter Uti du ministère de l'agriculture pour fournir des éléments de connaissances sur les successions de cultures et leurs évolutions.
- ▶ utilise des méthodes de classifications probabilistes fondées sur les modèles de Markov cachés.

Principes de la fouille par CarottAge

CarottAge utilise les modèles de Markov cachés (HMM) pour fouiller la base de données Ter Uti en faisant les hypothèses suivantes :

- ▶ La production agricole d'une région est stationnaire pendant un intervalle de temps :
→ **une distribution de cultures** associée à une période temporelle caractérise un **état** de la région.
- ▶ L'état d'une région à l'instant t dépend des états aux instants précédents :
→ *on utilise un modèle de Markov d'ordre 1 ou 2 pour décrire l'évolution des distributions (ou états).*

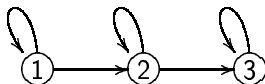
Ces hypothèses permettent la réalisation d'algorithmes d'apprentissage automatique à partir de tableaux de données.

Il est à destination de plusieurs utilisateurs :

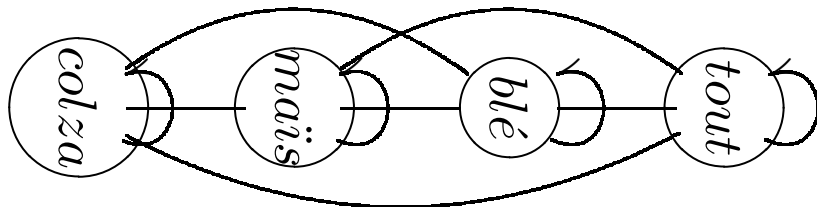
- ▶ les chercheurs agronomes, qui peuvent quantifier des connaissances qualitatives, infirmer ou confirmer des hypothèses et élaborer par exemple des modèles dynamiques de paysages ou des modèles de pollution de l'eau ;
- ▶ les décideurs ?? ou les enseignants des écoles d'agronomie qui cherchent à avoir une vue d'ensemble de l'évolution des phénomènes agraires.

Plusieurs modèles

Distributions de cultures sur plusieurs périodes (évolution des assolements) :

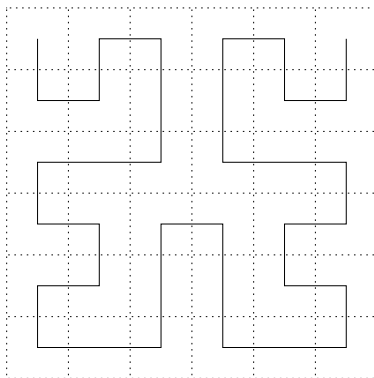


Introduction d'états spécifiques, états de 'Dirac'. On recherche ainsi les transitions entre cultures, donc les successions, et non pas l'évolution des assolements.



Modèle spatial

On utilise un parcours de l'espace pour le ramener à une dimension.
Utilisation d'une courbe fractale (courbe de Hilbert-Peano)



Les données sont pré-traitées et mises en forme selon les éléments recherchés :

- ▶ réduction du nombre d'occupations
- ▶ complétion pour la segmentation spatiale
- ▶ mise sous forme de tableau 1 culture
- ▶ mise sous forme de tableau 3 cultures
- ▶ regroupement de cultures (céréales, cultures de vente ...)
- ▶ etc.

Étude de successions de cultures dans le bassin de la Seine

Le programme PIREN-Seine développe un outil de prévision de la qualité des eaux du bassin de la Seine en fonction d'hypothèses sur des changements de pratiques agricoles.

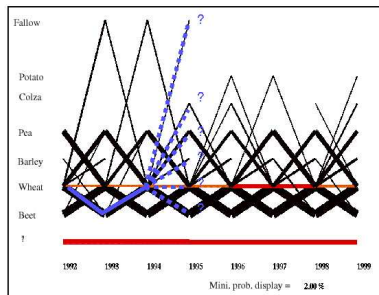
Les chercheurs de l'INRA Mirecourt travaillent à l'analyse de l'agriculture du bassin, de ses dynamiques et de son organisation spatiale.

Ils s'intéressent aux successions de cultures qui constituent une variable explicative des risques de pertes en nitrates.

CarottAge a été utilisé sur les données Ter Uti des petites régions agricoles du bassin de la Seine en différentes étapes.

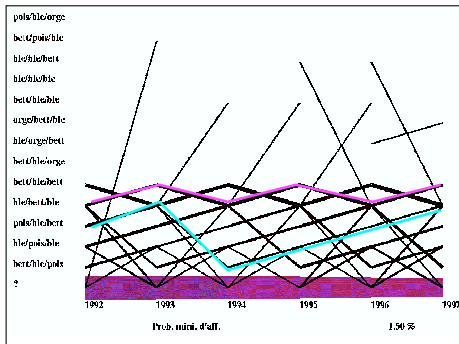
Modèles 1 culture – PRA St Quentinnois et Laonnois années 1992-2000

wheat	40,7%
beet	20,5%
pea	11,3%
permanent meadow	7,0%
barley	4,9%
fallow	4,0%
maize	3,7%
colza	2,6%
potato	1,9%



Modèles 3 cultures : découverte de successions sur 4 ans

blé-bett-blé	14,5%
blé-pois-blé	7,0%
bett-blé-bett	6,9%
pp-pp-pp	6,8%
pois-blé-bett	5,3%
bett-blé-pois	4,5%
bett-blé-blé	1,9%
blé-blé-blé	1,6%



succession bi-annuelle (*betterave* + *ble*)⁺

et succession sur 4 ans (*betterave* + *ble* + *pois* + *ble*)⁺

Aide à l'interprétation d'images satellitaires en Midi-Pyrénées (INRA Toulouse – LORIA)

- Pour évaluer les besoins en irrigation, on utilise des images satellitaires qui permettent d'établir des cartes d'occupations du sol sur une région.

Aide à l'interprétation d'images satellitaires en Midi-Pyrénées (INRA Toulouse – LORIA)

- ▶ Pour évaluer les besoins en irrigation, on utilise des images satellitaires qui permettent d'établir des cartes d'occupations du sol sur une région.
- ▶ À la date où les prévisions doivent être établies (au printemps), les images satellitaires disponibles ne sont pas suffisantes pour cartographier avec certitude les occupations du sol de l'année courante.

La connaissance des successions de cultures pratiquées permet alors de lever certaines ambiguïtés.

Aide à l'interprétation d'images satellitaires en Midi-Pyrénées (INRA Toulouse – LORIA)

- ▶ Pour évaluer les besoins en irrigation, on utilise des images satellitaires qui permettent d'établir des cartes d'occupations du sol sur une région.
- ▶ À la date où les prévisions doivent être établies (au printemps), les images satellitaires disponibles ne sont pas suffisantes pour cartographier avec certitude les occupations du sol de l'année courante.

La connaissance des successions de cultures pratiquées permet alors de lever certaines ambiguïtés.

- ▶ CarottAge est utilisé pour réaliser une segmentation spatiale et temporelle des données. La segmentation spatiale permet de définir des régions homogènes vis-à-vis des successions sur une longue période puis une segmentation temporelle permet d'établir des probabilités de transitions entre cultures, région par région.

Cartographie des successions culturales dans les départements du Gers et des Hautes-Pyrénées

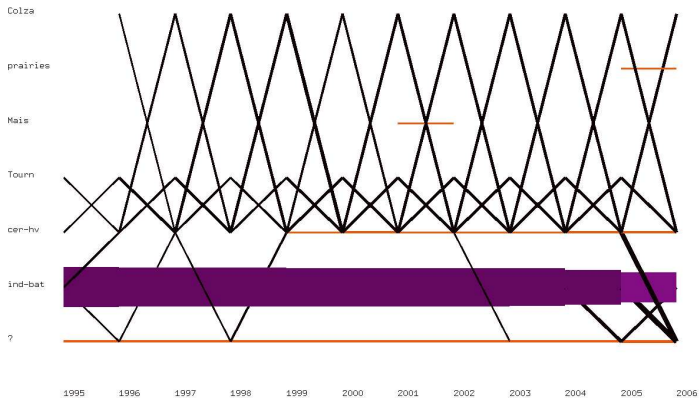


années 90	
0.099	bois + bois + bois
0.089	maïs + maïs + maïs
0.082	prairies + prairies + prairies
0.034	blé + tournesol + blé
0.028	vignes + vignes + vignes
0.026	tournesol + blé + tournesol
0.018	jachères + jachères + jachères
années 2000	
0.105	maïs + maïs + maïs
0.099	bois + bois + bois
0.081	prairies + prairies + prairies
0.056	blé + tournesol + blé
0.037	tournesol + blé + tournesol
0.027	vignes + vignes + vignes
0.023	jachères + jachères + jachères

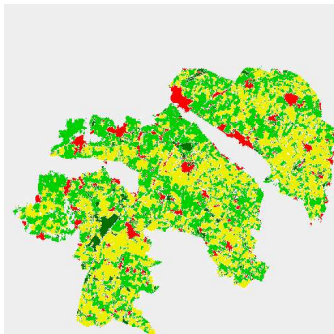
Analyse du territoire de Chizé : préliminaires

- ▶ Relevé des occupations au niveau des parcelles de 1995 à 2006;
- ▶ Application d'une grille spatiale à 1 niveau (10 mètres, 60 mètres ...) pour simuler Ter Uti;
- ▶ Analyse temporelle des successions;
- ▶ Classification de l'espace pour localiser les régions correspondant à ces successions.

Successions temporelles à Chizé



Analyse spatiale



jaune	{colza, tournesol, maïs } + blé
vert clair	les successions inexpliquées
vert foncé	forêts, prairies
rouge	bâti

Conclusions

Le logiciel CarottAge utilise

- ▶ Une méthode de classification de données temporelles et spatiales obtenues par un modèle probabiliste, le HMM.
- ▶ La même mesure probabiliste pour classer les données dans le temps et dans l'espace et donne une cohérence aux traitements.

Le logiciel CarottAge est utilisé dans différents programmes de recherche concernant l'agriculture :

- ▶ Analyse de l'effet des pratiques agricoles dans la pollution des eaux ;
- ▶ Prédiction des besoins en irrigation ;
- ▶ Simulation : Les résultats du logiciel CarottAge sont utilisés avec d'autres modèles pour effectuer des simulations prospectives sur l'occupation agricole du sol et ses effets en matière d'eau et d'environnement (diffusion des transgènes)

Aide à l'analyse :

- ▶ Couplage avec un système de raisonnement à bases de connaissances ou d'autres méthodes de classification pour une meilleure aide à l'exploitation des résultats (système d'ECBD).
- ▶ Classification symbolique à l'aide de treillis de Galois de régions sur la base de leurs successions de cultures.

Quelques références

U. Fayard, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), *Advances in Knowledge Discovery in Data Mining*, AAAI/MIT Press, 1996.

M. H. Dunham, *Data Mining. Introductory and Advanced Topics*, Prentice Hall, 2003.

J.-F. Mari, F. Le Ber, M. Benoît, Segmentation temporelle et spatiale de données agricoles, *Revue Internationale de Géomatique*, 12(4), pp. 439–460, 2002.

C. Mignolet, C. Schott, M. Benoît, Spatial dynamics of agricultural practices on a basin territory: a retrospective study to implement models simulating nitrate flow. The case of the Seine basin, *Agronomie*, 24, pp. 219–236, 2004.

J.-F. Mari, F. Le Ber, Temporal and Spatial Data Mining with Second-Order Hidden Markov Models, *Soft Computing*, 2005.
<http://dx.doi.org/10.1007/s00500-005-0501-0>