CarottAge Windows pour les données Teruti : manuel d'utilisation

Jean-François Mari

CarottAge Windows pour les données *Ter-Uti* : manuel d'utilisation

Mari Jean-François
Loria/Inria-Grand Est
615 rue du Jardin Botanique,BP 101,
F-54600 Villers-lès-Nancy, France

February 21, 2014

Chapter 1

CarottAge

1.1 Présentation de CarottAge

CarottAge est un rétro acronyme construit à partir du mot carotte qui se traduit en *markov* en russe et du mot Âge. C'est aussi un procédé d'analyse de la constitution des sols. Faire un carottage d'un sol, c'est extraire par forage un cylindre représentatif des couches traversées afin d'étudier leurs successions et les dater.

CarottAge est le résultat d'un travail de fouille de données effectué par des agronomes de l'Inra SAD ASTER (Mirecourt) et des informaticiens du projet Orpailleur Loria et Inria Grand Est pour extraire des bases de données agricoles *Ter-Uti* des informations sur les successions de cultures pratiquées dans une région.

CarottAge s'appuie sur la théorie des chaînes de Markov cachées - HMM comme Hidden Markov Model - pour permettre l'analyse de successions d'observations quelconques continues ou discrètes. Ces modèles 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 voisins.

Le logiciel calcule et affiche un signal dont l'analyse permet l'extraction et la datation de régularités temporelles et spatiales. Il est fourni sous forme d'une boîte à outils comportant plusieurs programmes indépendants ainsi qu'une application graphique qui permet de les enchaîner d'une façon interactive.

La première publication majeure de CarottAge se trouve dans la revue Ecological Modelling : *Studying crop sequences with* CarottAge*, a HMM-based data mining software* [39] dont le pre-print est donné en annexe. Sa

lecture est vivement conseillée (désolé !) avant toute expérimentation.

1.2 Présentation des données *Ter-Uti*

Notre ensemble de données est constitué de l'enquête *Ter-Uti* qui est réalisée par un sondage à deux niveaux de granularité. Un premier tirage, réalisé par l'IGN, consiste à sélectionner des photos aériennes régulièrement réparties sur l'ensemble du territoire métropolitain. Les photos représentent chacune un carré de 2 km de côté et sont séparées en moyenne par 6 km. Un deuxième tirage, réalisé par les DRAF[1], 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 sites est visité annuellement par des enquêteurs qui relèvent les occupations des sites. Pour plus de détails sur la grille *Ter-Uti*, on peut se reporter à [41].

Outre la séquence temporelle des occupations de chaque point, nous savons à quelle PRA il appartient et nous connaissons ses voisins, c'est-à-dire la disposition relative de chaque point et de chaque photo aérienne. En revanche, nous ignorons la localisation précise des points pour des raisons de secret statistique.

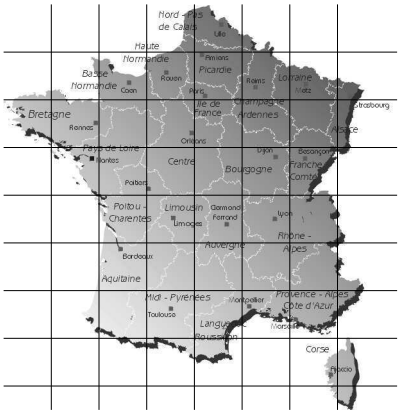
Les services de statistique de la DRAF ont réparti les occupations en différentes classes (environ 80) qui vont de "marais salants, étangs d'eau saumâtre" à " peupliers épars" en passant par "superficie en herbe à faible productivité potentielle". Certaines de ces classes ne sont pas ou peu présentes dans les régions étudiées considérées aussi avons nous restreint le nombre de classes à 49, par regroupement ou suppression [7].

nLig=112806, annee1=1992, anneen=2003, nAttr=1, indeter=95, isHeader=1

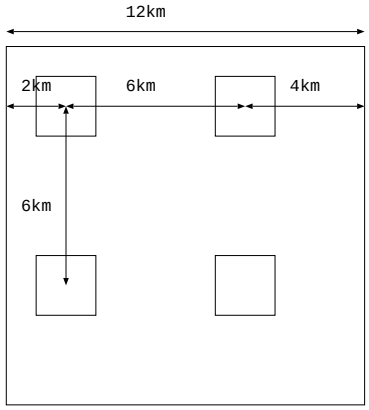
pt	dep	pra	photo	pti	92	93	94	...	00	01	02	03
1	2	2034	8885	1	27	28	42	...	42	27	27	27
2	2	2034	8885	2	27	33	27	...	40	27	27	42
3	2	2034	8885	3	27	40	52	...	27	40	27	33
						...						

Table 1.1: Chaque point est étiqueté par son département (dep) et sa PRA (petite région agricole)

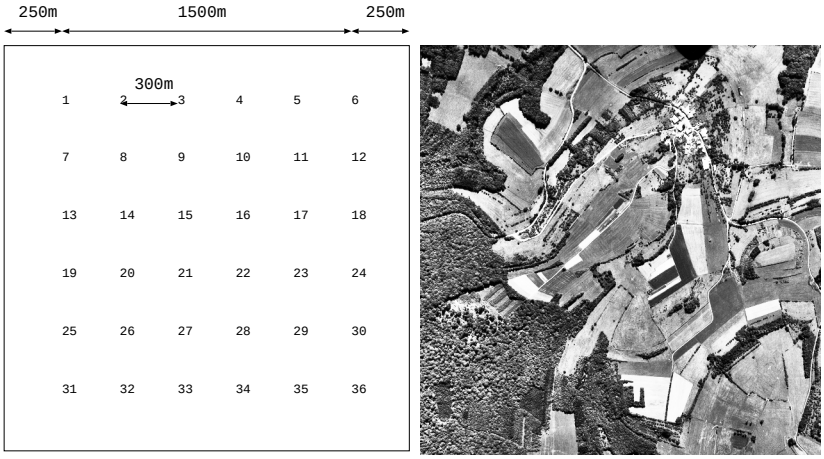
[1] Direction Régionale de l'agriculture et de la Forêt.



(a) Le principe du maillage de base du territoire



(b) Les 4 photos aériennes choisies dans une maille



(c) La photo aérienne et sa grille 6x6

Figure 1.1: Description des données *Ter-Uti*: 3820 mailles quadrillent la France (toutes ne sont pas représentées), 4 photos aériennes sont choisies dans une maille, une grille 6x6 détermine 36 sites

1.3 Installation

On trouve CarottAge pour traiter les données *Ter-Uti* sur le site du Loria [2] sous la forme d'une archive `carottage-windows-teruti-V1.zip` (Carottage pour windows et données *Ter-Uti*). Cette version est paramétrée pour traiter des données *Ter-Uti*. Plusieurs dossiers sont fournis :

SrcQt contient l'exécutable graphique `carrotage.exe`[3] dans le sous-répertoire `SrcQt/debug` ;

SrcPirenSpatial contient les binaires exécutables compilés correspondant à tous les outils nécessaires pour traiter les données *Ter-Uti* pour la France entière ;

config contient les fichiers de configuration *Ter-Uti*. Il s'agit de fichiers donnant la définition de la classification *Ter-Uti*: blé, orge, ... ainsi que les regroupements que nous avons opérés comme "bois" et "eau" qui regroupent toutes les superficies en bois et eaux respectivement ;

Mod contient les fichiers de l'espace de travail (description de modèles, modèles initiaux et finaux; ...) ;

Corpus contient le fichier de données *Ter-Uti* `short-example.txt`. Le fichier de données *Ter-Uti* `NouvelleFrance.txt` construit à partir des données *Ter-Uti* fournies par le Service de la Statistique et de la Prospective (SSP) du Ministère en charge de l'agriculture n'est pas inclus dans ce dossier car il n'est pas en *Open Access*. Toutefois, pour permettre une démonstration et vérifier la bonne installation du logiciel, ce fichier artificiel est fourni ;

DLL est un dossier qui contient les DLL (Dynamic Link Libraries Qt pour Windows Xp ou 7). Pour faire fonctionner CarottAge, il faut modifier la variable PATH dans le menu *System* de Windows pour ajouter le chemin d'accès à ce dossier. On peut aussi copier les fichiers du répertoire DLL dans le répertoire `SrcQt/debug`.

1.4 Exemple pour débuter

Dans cette section, nous allons exécuter les différents outils de CarottAge sur le fichier de démonstration fourni `short-example.txt`.

[2] http://www.loria.fr/~jfmari/App
[3] attention à l'orthographe anglaise

1.4.1 Segmentation de la période d'étude

Il s'agit d'étudier la dynamique de l'assolement de notre région. Une première solution consiste à déterminer autant d'assolements qu'il y a d'années de collecte d'occupation *Ter-Uti*. Une autre solution consiste à se limiter à un nombre limité de périodes – disons trois pour avoir une vue plus concise de l'évolution – et de laisser les `HMM` effectuer la meilleure[4] segmentation. Nous utiliserons le fichier `short-example.txt` pour obtenir des résultats comparables à ceux de la publication [39]. Dans cette fouille de données, on s'intéresse aux observations formées d'une seule occupation du sol. Leurs définitions sont regroupées dans le fichier `teruti1.cfg`. Si on avait voulu travailler avec des triplets d'occupation, on aurait utilisé `teruti3.cfg`.

L'archive contient un fichier de commandes Windows `do_example.bat` qui enchaîne les commandes :

création de la description du HMM : le fichier `lin3.lst` est crée par la commande `model-lin-gen.exe 3` comme décrit page 22. Ce fichier décrit la topologie du `HMM` (linéaire à trois états) ainsi que les densités de probabilités (*pdf* comme *probability density function*) qui sont ici uniformes ;

inventaire des observations : le programme `ter2indice-tempo` parcourt le fichier `short-example.txt` afin d'inventorier toutes les observations possibles. L'inventaire est représenté par la liste `bin-teruti1.lst` ;

création du Hmm : le programme `editmodel` crée le HMM à partir des fichiers `lin3.lst` et `bin-teruti1.lst` ;

estimation du Hmm : le programme `fwtInra` joue le rôle de la commande `estimate` évoquée page 22 ;

visualisation : le programme `gviewmod` construit le fichier `lin3.txt` qui donne des résultats comparables à ceux de la figure 2.3 page 25.

Comme décrit dans l'article [39], la figure 1.2 montre bien la progression puis disparition de la jachère ainsi que l'érosion des prairies.

1.4.2 Visualisation de transitions entre cultures

Pour obtenir un résultat comparable à celui de la figure 2.1 page 23, il faut spécifier un nouveau modèle `HMM` dit " HMM ergodique avec états de Dirac".

[4] au sens du maximum de vraisemblance

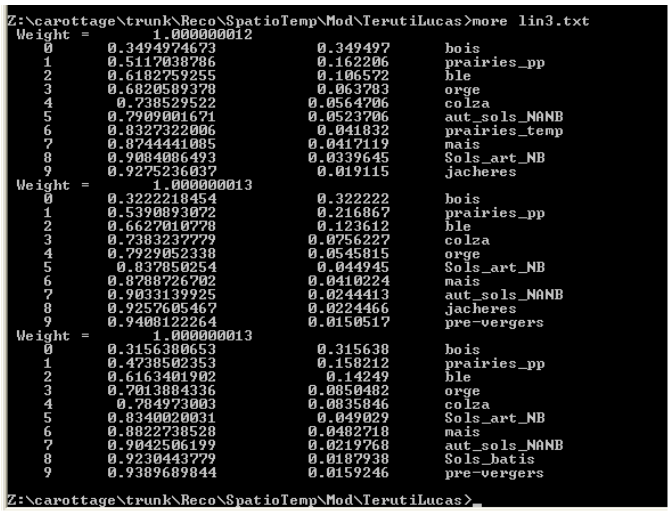


Figure 1.2: Visualisation des 3 *pdf* de `lin3.txt`

L'adjectif ergodique signifie ici que toutes les transitions entre états sont possibles. Le terme "état de Dirac" a été emprunté à la théorie des distributions. Il signifie que la densité de probabilité associée à cet état a la forme d'une impulsion de Dirac : un pour une occupation, zéro ailleurs. La construction du modèle ergodique dans lequel les états associés au blé, orge et colza sont différenciés se fait en plusieurs temps :

1. spécification d'un modèle linéaire avec le même nombre d'états, à savoir 6 états : "blé", "orge", "colza", "maïs", "prairies + forêts" ainsi que l'état "?" qui joue le rôle de "container" et qui capturera toutes les exceptions (cf. Tab. 1.2).

2. transformation de ce `HMM` linéaire en ergodique par la commande `lin_to_ergo`.

3. estimation par la commande `fwtInra`

4. visualisation du diagramme de Markov de la figure 2.3 page 25.

Toutes ces étapes sont regroupées dans le fichier `do_markov.bat` (cf. Tab. 1.3).

```
0 1 1
1 2 1
2 3 1
3 4 1
4 5 1
5 6 1
6 7 1
7 8 1
2 2 1
3 3 1
4 4 1
5 5 1
6 6 1
7 7 1
-1 -1 -1
{ble}
{orge}
{colza}
{mais}
{prairie +
bois}
equiprobable
```

Table 1.2: Description du HMM linéaire à 6 états dont 5 états de Dirac (fichier `bocm.lst`)

```
rem commandes pour realiser les exemples
rem de Ecological Modeling "Studying Crop Sequence whith Carottage ..."
rem Leber, Benoit, Schott, Mari, Mignolet
rem 2006
rem cultures simples teruti
rem fichier teruti1.cfg
rem executer dans Mod\TerutiLucas
set CORPUS=../../Corpus/TerutiLucas/short-example.txt

rem remplacer teruti1.cfg par le fichier de configuration correspondant
rem creation du Hmm lineaire a 6 etats
start /W ../../SrcPirenSpatialWindows/ter2indice-tempo.exe  -t ../../config/teruti1.cfg %CORPUS% -o bin-teruti1.lst
start /W ../../SrcPirenSpatialWindows/editmodel.exe -t ../../config/teruti1.cfg -d bocm.lst -i bin-teruti1.lst -o lin-bocm.mod
rem transformation en ergodique
start /W ../../SrcPirenSpatialWindows/lin_to_ergo.exe -t ../../config/teruti1.cfg lin-bocm.mod ergo-bocm.mod
start /W  ../../SrcPirenSpatialWindows/fwtInra.exe  -t ../../config/teruti1.cfg -n 6 ergo-bocm.mod  -o ergo-bocm.mod1 %CORPUS%
start /W ../../SrcPirenSpatialWindows/gviewmod.exe -t ../../config/teruti1.cfg ergo-bocm.mod1  -o ergo-bocm.txt -m 10
start /W  ../../SrcPirenSpatialWindows/fwtInra.exe  -t ../../config/teruti1.cfg -n 1 -x 2 ergo-bocm.mod1  -o ergo-bocm.gph %CORPUS%
start /W  ../../GviewGraph2_Qt\debug\GviewGraph.exe  ../../config/teruti1.cfg ergo-bocm.gph  0.01 1991 2003
```

Table 1.3: do_markov.bat: fichier de commandes pour créer le diagramme de Markov de la figure 2.1

1.5 Utilisation de l'interface graphique CarottAge

L'archive contient une application graphique qui permet aussi d'enchaîner manuellement ces étapes en dispensant l'utilisateur de l'écriture des fichiers de commandes. Les résultats sont les mêmes dans les deux modes de fonctionnement : fichier "bat" ou interface graphique.

1.5.1 Première utilisation

A la première utilisation, CarottAge demande de choisir deux répertoires : un répertoire de travail qui contiendra les fichiers de données ainsi qu'un répertoire de binaires. Il est possible de revenir sur ces choix grâce à l'option `Fichier`.

Le choix du répertoire de travail

Le bon choix est le répertoire Mod/TerutiLucas ;

Le choix du répertoire des binaires

Le bon choix (et c'est le seul) est SrcPirenSpatial.

1.5.2 Le sous menu : données

Cette version de CarottAge traite des fichiers de données *Ter-Uti* élaborés à partir de données fournies par le Service central de la statistique agricole. Dans ce menu, il faut ici préciser où se situe le fichier `short-example.txt` (cf. Fig.1.3). Ce menu permet aussi de se limiter à une période d'étude et d'appliquer un filtre d'extraction de points *Ter-Uti*, par exemple en précisant une liste de PRA ou de départements. Il faut pour cela avoir le fichier `nouvelleFrance.txt`.

1.5.3 Le sous menu : configuration

L'enquête *Ter-Uti* fournit une classification très précise de l'occupation du territoire. La centaine d'étiquettes différentes *Ter-Uti* doit être regroupée en un nombre bien inférieur de classes d'occupation du sol. Commencer par choisir dans ce sous menu : `importer une configuration` et choisir le fichier `teruti1.cfg` qui se trouve dans le dossier `config`. Par une série de glisser / insérer, on peut modifier ce regroupement. Avant de sortir de ce sous menu, `valider la configuration`, ce qui créera le fichiers des observations possibles.

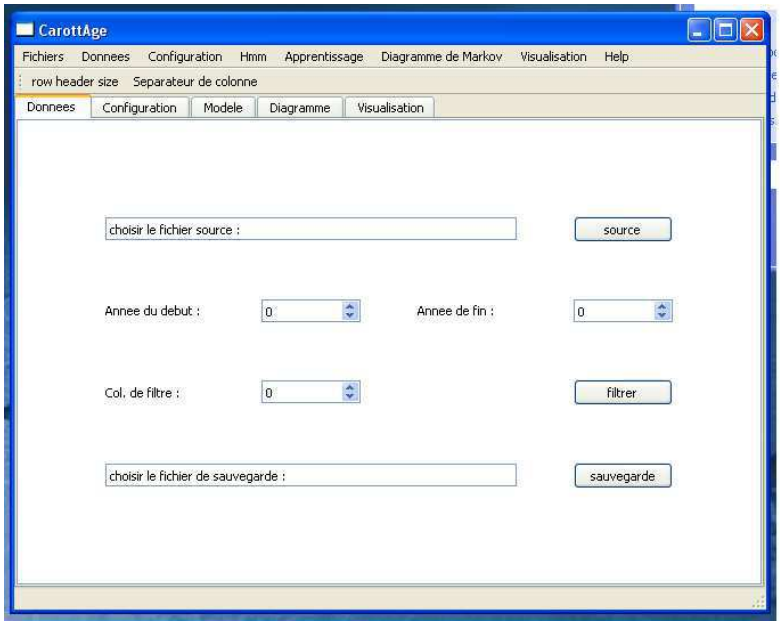


Figure 1.3: Le menu : données

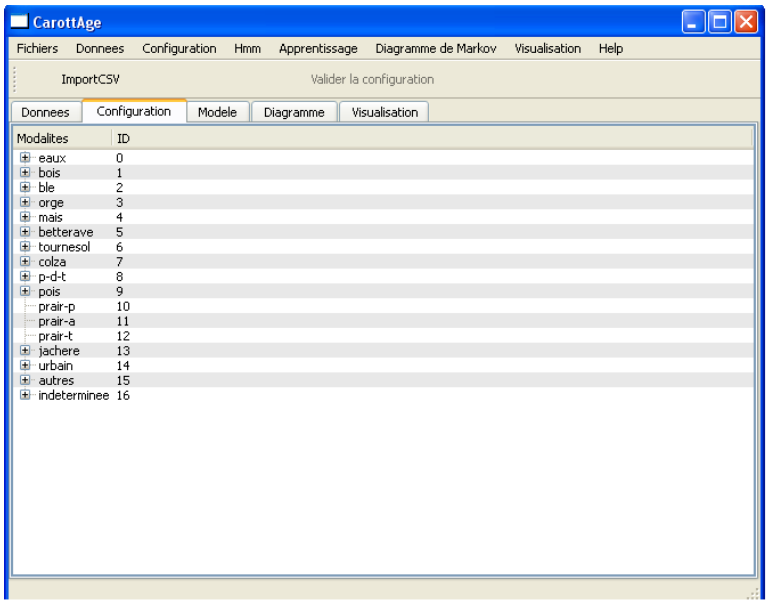


Figure 1.4: Le menu : configuration après avoir importé `teruti1.cfg`

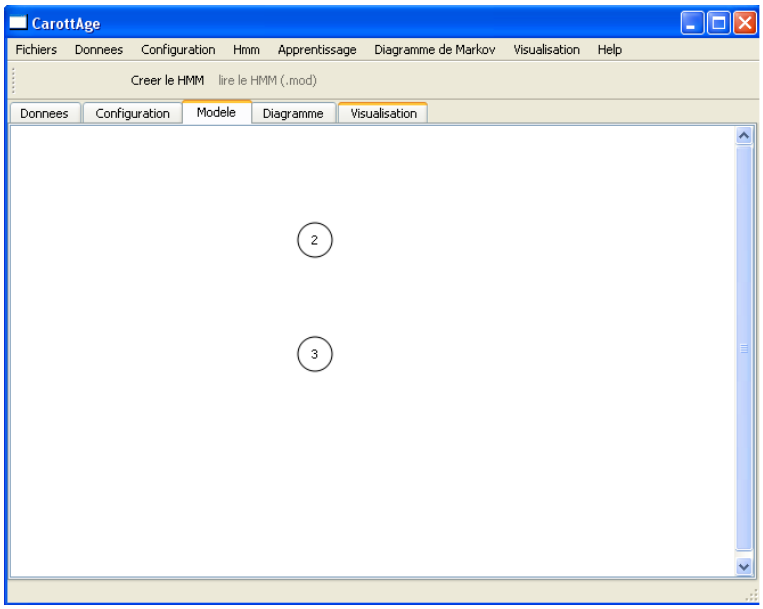


Figure 1.5: Le menu : modèle. Visualisation d'un HMM ergodique à une colonne de 2 états

1.5.4 Le sous menu : modèle

Ce sous menu (cf. Fig. 1.5) permet de construire les fichiers de description des `HMM`. Il fera appel dans votre dos aux programmes `model-lin-gen`, `editmodel`, ...

Deux familles de topologies sont possibles par l'option créer un modèle : linéaire et à colonne d'états. La topologie choisie dans le fichier `do_markov.bat` correspond à une seule colonne de 6 états. Par défaut, les états sont associés à des lois uniformes. En cliquant sur chaque état, on peut choisir les occupations pour les transformer en état de Dirac. Dans notre cas, il faut choisir une état blé, un état orge, un état colza, un état maïs, un état prairies et forêts en sélectionnant dans la liste toutes les occupations que l'on souhaite capter par cet état. Le dernier état reste équiprobable. Une fois la description spécifié, l'option `créer le HMM` crée la forme interne du `HMM`. La description est un fichier texte (cf. Tab 1.2), alors que le `HMM` a un format interne binaire stocké dans un fichier d'extension `.mod`.

1.5.5 Le sous-menu : Apprentissage

C'est le moins fourni de tous, mais celui qui en fait le plus. Choisir un nombre d'itérations égal aux nombres d'états sauf si vous savez ce que vous faites ! et lancer l'apprentissage par la commande `fwtInra`.

1.5.6 Le sous-menu : diagramme

Ce sous-menu permet la visualisation des diagrammes de Markov (cf. Fig. 1.6) et leurs sauvegardes dans différents formats.

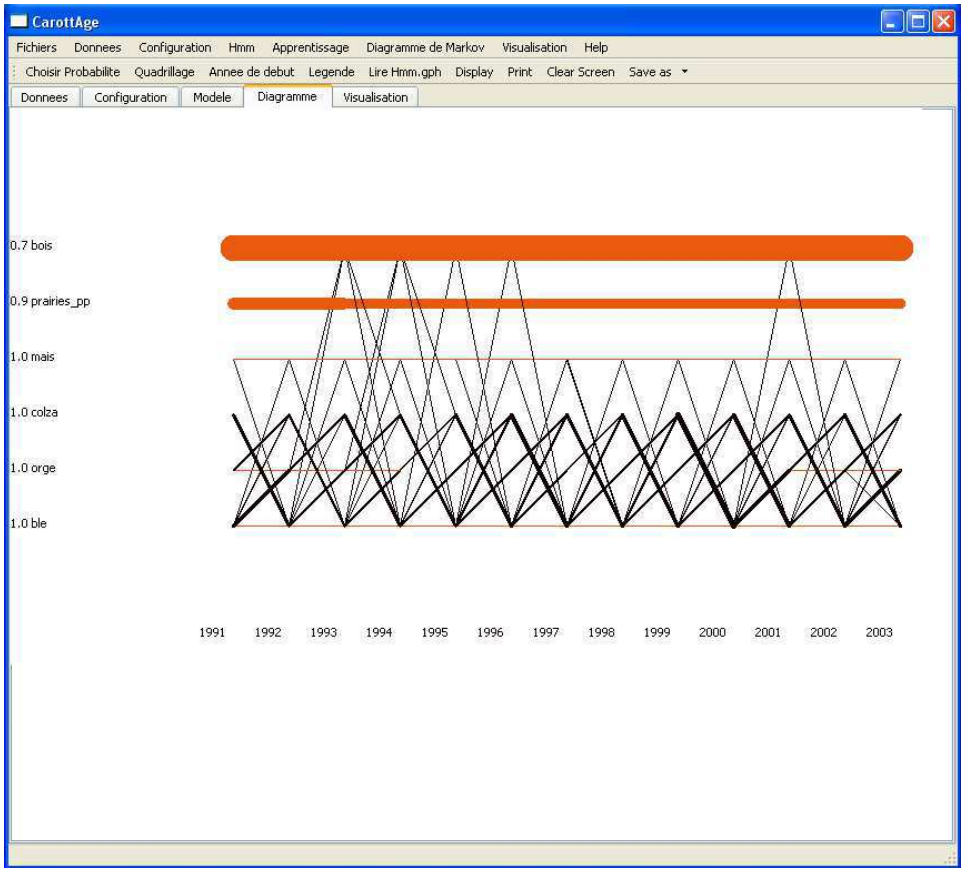


Figure 1.6: Visualisation des transitions entre cultures

1.5.7 Le sous-menu : visualisation

Ce sous-menu permet la visualisation des *pdf* associées aux états (cf. Fig. 1.7).

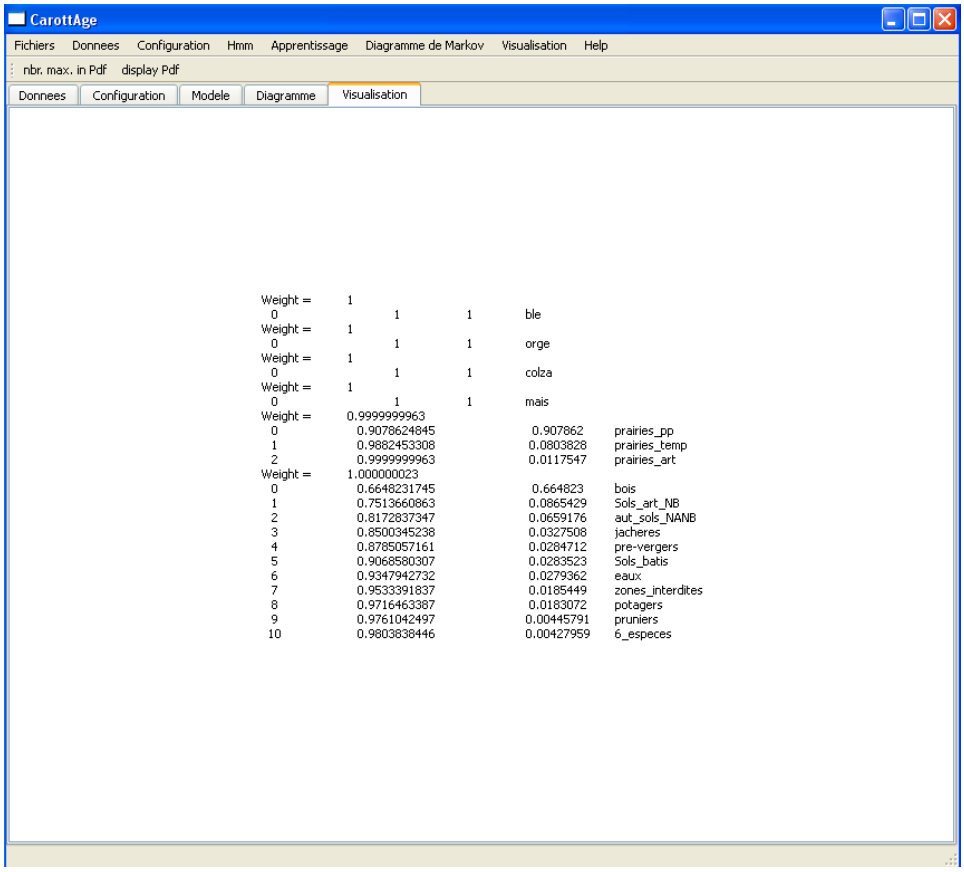


Figure 1.7: Visualisation des *pdf* du modèle ergodique

La figure 1.7 mérite quelques explications. Elle représente le `HMM` après apprentissage. On remarque que les *pdf* associées aux états de Dirac sont restées dans leur définition initiale. En revanche, l'état "container" – initialement loi uniforme – s'est peuplé des occupations qui ne pouvaient être captées par les états de Dirac. Il s'agit du cas idéal. Lorsque le modèle ne correspond pas à la réalité, on assiste à un phénomène de dérive dans lequel les états de Dirac se peuplent d'occupations majoritaires bien différentes de ce qui était prévu au départ. Tout l'art de la fouille de donnée par modélisation stochastique consiste à spécifier des modèles qui convergeront

vers un modèle utile à l'extraction de connaissances.

1.6 Développements futurs

CarottAge a donné naissance à ARPEnTAg [46] car la recherche des successions de cultures dans un territoire a vite fait apparaître le besoin de pouvoir les localiser et faire apparaître des quartiers culturaux comme l'avait fait remarquer J.-P. Deffontaines [20].

Pour utiliser CarottAge sur d'autres jeux de données, il est nécessaire de créer un répertoire `SrcMonProjet` pour y dériver une classe à partir de `Corpus0`. Cela nécessite un travail de programmation en C++. Les explications pour y arriver dépassent le cadre de ce tutoriel et feront l'objet du *Manuel du programmeur*.

Chapter 2

Annexe

Studying crop sequences with CarrotAge, a HMM-based data mining software

F. Le Ber[1,2] M. Benoît[3] C. Schott[3]
J.-F. Mari[2] C. Mignolet[3]

[1] ENGEES, 1 quai Koch, BP 1039, F-67070 Strasbourg Cedex
fleber@engees.u-strasbg.fr
phone: 33 388248230; fax: 33 388248284
[2] UMR 7503 LORIA, BP 239, F-54506 Vandœuvre-lès-Nancy Cedex
jfmari@loria.fr
[3] INRA SAD, Domaine du Joly, F-88500 Mirecourt
{benoit,schott,mignolet}@mirecourt.inra.fr

also in Ecological Modelling, 191(1):170 – 185, Jan 2006

2.1 Introduction

Sixty years after its launching through the "Marshall Plan", the European agriculture revolution is up again, but with some strong contradictions: water pollution, landscape uniformization, ethical crisis [26]. These harmful side-effects of agriculture could be aggravated if the evolution of agricultural practices continues following the current trends towards greater concentration, intensification and technicality. We focus on agricultural practices, from their choice by farmers decisions to their effects, as they continuously remodel the agricultural landscapes. The

approach of farming systems as landscapes "builders" is a new one, but its background is the vision of land as resource for agriculture [19, 37]. Agronomic measures specifically designed to maintain soil, water and air quality are necessary, including more severe regulations restricting intensification and the agricultural use of chemicals. For instance, keeping the nitrate content of drainage water to less than 50 mg.l-1 requires not only an optimized and reduced application of fertilizers, but also the planting of catch crops during the winter. Parts of the hydrological basins in many areas should be withdrawn from arable cropping and turned into grasslands or forests (several authors in [43]). Preventing runoff erosion and the associated pollution of surface water (especially by pesticides) needs grassland strips, ditches, or other structures placed in suitable strategic locations in a catchment. Again, similar conclusions could be drawn about many other environmental targets, such as biodiversity, or landscape quality and accessibility [13].

The farmer practices are the focus point of researchers who built tools to help their changes [9]. In this paper we propose a methodological approach of farmer practices involved in the land designing through land uses and land pattern changes. Actually, our approach combines agronomic and artificial intelligence methods. We rely on a land use data base, that we explore with a data mining approach, to find out spatial and temporal land patterns.

Mining sequential and spatial patterns is an active area of research in artificial intelligence. One basic problem in analyzing a sequence of items is to find frequent episodes, i.e. collections of events occurring frequently together. We rely on new numerical algorithms, based on high-order stochastic models – the second-order hidden Markov models (`HMM2`) – capable to discover frequent sequences of events in temporal and spatial data. These algorithms can extract spatial and temporal regularities that can be explained by human experts and may constitute elements of a *knowledge discovery process* [48]. Thus, agronomists and computer scientist have designed a data mining software, named CarottAge, in order to extract crop sequences and patterns from land-use data bases. This software allows the user to specify the architecture of the Markov model according to the data and his objectives. Displaying tools have also been defined. CarottAge is used in several research projects, e.g. agronomists try to find out crop sequences in order to model nitrate loss due to agricultural activities.

The paper is organized as follows. Part one is about the relationship between land and farmer practices and the modeling of crop rotations.

Part two is about HMM2 and the CarottAge software. Part three presents some results obtained by CarottAge on a French data base, and their analysis. Then we conclude and propose some perspectives.

2.2 An agronomic question

2.2.1 The relationships between land and agriculture

We want to focus on the mutual relationship between land and farmer practices: on the one hand, the current state of the land is a result of farming practices and changes in landscapes could not be decided without farmers participation, but on the other hand, the choice and location of cropping and grassland systems by farmers all over the world takes into account their own land characteristics [37, 38]. This management of land by farmers is a part of the global technical management building agriculture [9] and is a factor of farm economical effectiveness [2]. The future of European land is based on this management [25].
Environmental issues may be converted into farming systems questions in which the activities of farmers and their changing location from the new picture is the focus point of problem solving [27, 11]. A number of new research tools such as remote-sensing data and Geographical Information Systems are now available to address this type of research [8].
In most cases, farmers are seen to take into account the properties and layout of their land in deciding about the location of their cropping and grassland systems [53]. This relationship between farmers and their territory could be an individual or a collective one [40].

2.2.2 Land-use is managed by farmers

The land used by agriculture can be modeled as a complex and dynamic pattern of fields, including tilled plots and pastures. Sebillotte, in the 1st European Society of Agronomy Congress, defined the *cropping system* as a set of crop management procedures used on a homogeneously treated space inside a farm, which can be a field, or a part of field, or several fields. According to this definition, a given cropping system is a component of a farming system, and is identified (characterized) by the sequence of crops and corresponding technical operations [54].
The cropping system is a tool to characterize land use on the tilled part of farms [55]. However many farms have not only tilled crops but comprise

also pastures. So if we want to reason at farm scale, it is necessary to generalize the concept of cropping system by including grasslands [28].
So we propose to name *Agricultural Land Management System* (ALMS) the system of crop and grassland management procedures used on a portion of land (which can be a field, or a part of field including its boundaries, or several fields). According to this definition, a given ALMS is a component of a farming system, and is identified (characterized) by the choices of the rotation of crops or grassland uses, the farmland structure and the location rules of the crop rotations and grassland uses. This definition should be completed by including also common items such as hedges, fences etc. that are components of the landscape and play a role in farm management [15]. For us, the ALMS is the basic unit of landscape design at farm scale. At a regional scale, other land uses and actors outside farms should be taken into account (forests, waters, *wild areas*) to complete the ALMS, according to the aims of the models (biodiversity management, water protection, leisure) as well as collective farmers' organizations [17, 18, 30, 58, 57].

2.2.3 A proposal of European notation for crop sequences identification

As a tool of representation and understanding of the interactions between agriculture, land and environment, agricultural land use management could be used as well for research as for management and negotiations in agro-environmental policies. The main topic in this way should be focused on land use changes [35, 36]. Although the agricultural practices we are familiar with are far from covering the whole range of existing systems, we shall propose a method for establishing a nomenclature of ALMS.
The origin of these proposals lies in a number of monographs done for a large diversities of farms in a European research project[1] [5]. This first large range of landscape building monographs meets the work described in [53].
So, we propose a common notation of land use descriptions (Table 2.1) with two characteristics (i) description of the land uses as they are described, managed and decided by the actors, (ii) account of time scales as first organizational factor.
All over Europe and each year, the farmers have to allocate their crops and grassland uses in their territory. This allocation is an important part of farmer decision that we have to model [16]. This annual adjustment

[1] Regional Guidelines to Support Sustainable Land Use by EC-Agri-environmental Programs (EAP), AIR 3 CT94-1296.

between chosen crops and field plots results in different perennial rotations of crops and grassland use types [34]. Examples are:

- in Denmark: maize / maize / winter wheat / barley
- in south west France without irrigation: sunflower / winter wheat / barley
- in the East region of France: oil rapes / winter wheat
- in the plain of Rhine in Vorarlberg (Austria): maize / maize / temporary grassland for mowing (3 years).

These notations describe yearly sequence of crops or pasture uses as they are conceived by farmers: this has the advantages of corresponding to the planning structure of the farmer, which reasons rotations over several years, and to allow a stability of land use descriptions over years, whereas crop by crop descriptions would vary each year.
However they lack the account of the logic behind the simple crop rotation description, although some hints may be given (such as maize for silage *versus* maize for sale) which complete the raw fact description, so these notations cannot yet be fully counted as ALMS nomenclature. In the future, our aim is to contribute to build a framework of farmer rules used to build rotations [13]. The first work done by [3] shows the importance of delay between two crops, sowing and harvesting dates, machinery choices. Examples of use of the proposed European cropping/grassland management systems are given in table 2.1.

For crops:	
	`M / wW / wB..ic /`
means	Maize / winter Wheat / winter Barley with intermediary crops in autumn after harvesting
For grassland:	
	`.../ hC - tPH2 /...`
.../ /... means	each year the uses are the same
`hC` means	mowing for hay making
`tPH2` means	turning Pasture For Heifers 2 years old

Table 2.1: Nomenclature of crops and grassland uses sequences. Each crop name (e.g. M, wW) or grassland-use cluster (e.g. hC - tPH2) represents a year of the sequence.

2.3 Temporal Data Mining with HMM2

The purpose of pattern recognition is to specify as much models as there are classes to recognize. As opposite to pattern recognition, we do not have the knowledge of what to recognize but rather look for something regular to extract, hence the name *data mining*. Actually, data mining can be defined as the use of algorithms to extract information and patterns from databases [23, 22]. These algorithms are able to search the data and attempt to fit a model to the data, using some preference criteria. Data mining is a part of knowledge discovery processes that include four other steps: the selection of data, the preprocessing of data, the transformation of data, and the interpretation of the data mining results [22].
In the present work, we specify one second-order Hidden Markov Model (HMM2) in order to model, in a more simple way, the unknown behavior of a crop sequence. We rely on the assumption that the land-use of a field at time t depends on the land-use of the same field at time $t-1$, $t-2$, etc. Each state of the HMM2 captures a stationary behavior and represents a class (a crop or a cropping pattern) where the observations are drawn with a known probability density function. Furthermore, we compute the a posteriori probabilities that the Markov chain goes through some states between certain time slots. These a posteriori probabilities can be plot as a function of time and determine a fuzzy classification in the states space. This classification can be interpreted by the agronomists *wrt* the evolution of crop patterns and crop sequences.

2.3.1 HMM2 definition and automatic estimation

The second order Hidden Markov Models are based on the probabilities and statistics theories. They are implemented with unsupervised training algorithms (like the EM algorithm [21]) that allow to estimate a model parameters from a corpus of observations and an initial model. The resulting model is capable to segment each sequence in stationary and transient parts and to build up a classification of the data together with the a posteriori probability of this classification. This characteristic makes the HMM2's appropriate to discover temporal and spatial regularities as it is shown in various areas (e.g. [6, 12, 24, 31]). Furthermore, the very success of the HMMs is based on their robustness: even when the considered data do not suit a given HMM, its use can give interesting results.
In a HMM2, the underlying state sequence is a second-order Markov chain. Therefore, the probability of a transition between two states at time t

depends on the states in which the process was at time $t-1$ and $t-2$. A Markov chain is defined over a set of states – the crops in a field, or more generally the land-use categories in a place – that are unambiguously observed. The Markov chain specifies only one stochastic process, whereas in a HMM, the observation of a land-use category is not uniquely associated to a state but is rather a random variable whose conditional density depends on the current state at time t [4]. There is a doubly stochastic process:

- the former is hidden from the observer and is defined on a set of states;
- the latter is visible. It produces an observation, the land-use of a parcel, at each time slot depending on the probability density function that is defined on the state in which the Markov chain stays at time t. It is often said that the Markov chain governs the latter.

Thus, a HMM2 is specified by:

- a set of N states called $S = \{s_1, \dots s_N\}$;
- a three dimensional matrix (a_{ijk}) over S^3

$$\begin{aligned} a_{ijk} &= \text{Prob}(q_t = s_k / q_{t-1} = s_j, q_{t-2} = s_i) \\ &= \text{Prob}(q_t = s_k / q_{t-1} = s_j, q_{t-2} = s_i, q_{t-3} = ...) \end{aligned} \tag{2.1}$$

  with the constraints $\sum_{k=1}^{N} a_{ijk} = 1,\ \forall (i,j) \in [1,N]^2$, and where q_t is the current state at time t;

- a set of N discrete distributions: $b_i(.)$ is the distribution of observations associated to the state s_i. This distribution may be parametric, non parametric or even given by an HMM.

The probability of the state sequence $Q_1^T = q_1, q_2, ..., q_T$ is defined as:

$$\text{Prob}(Q_1^T) = \Pi_{q_1} a_{q_1 q_2} \prod_{t=3}^{T} a_{q_{t-2} q_{t-1} q_t} \tag{2.2}$$

where $\forall j, q_j \in S$, Π_{q_1} is the probability of state q_1 and $a_{q_1 q_2}$ is the probability of the transition $q_1 \to q_2$ (initialization of the model at times $t = 1$ and $t = 2$).

Given a sequence of observations $O_1^T = o_1, o_2, ..., o_T$, the joint state-output probability $\text{Prob}(Q_1^T, O_1^T)$, is defined as:

$$\text{Prob}(Q_1^T, O_1^T) = \Pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \prod_{t=3}^{T} a_{q_{t-2} q_{t-1} q_t} b_{q_t}(o_t). \quad (2.3)$$

The estimation of a `HMM1` is usually done by the Baum-Welch algorithm which is related to the EM algorithm [21]. We have shown that a `HMM2` can be estimated following the same way [45]. The estimation is an iterative process starting with an initial model and a corpus of sequences of observations that the `HMM2` must fit. Usually, the initial model has equiprobable transition probabilities and an uniform distribution in each state. At each step, the Baum-Welch algorithm determines a new model in which the likelihood of the sequences of observation increases. Hence this estimation process converges to a local maximum, according to the maximum likelihood (ML) estimation criteria [21, 47]. To assess the final model, we use the Kullback-Leibler distance between the distributions associated to the states [56]. Two states that are too close are merged and the resulting model is re-trained.
Intuitively, the Baum-Welch algorithm counts the number of occurrences of each transition between the states and the number of occurrences of each observation in a given state in the training corpus. Each count is weighted by the probability of the alignment between the states and the observations (cf. Equation 2.3). The principles of this algorithm are detailed in the appendix.

2.3.2 CarottAge

CarottAge[2] is a free software under a *Gnu Public License* that takes as input an array of discrete data – the rows represent the spatial sites and the columns the time slots – and builds a partition together with its a posteriori probability. CarottAge is written in C++ and runs under Unix and X11R6 systems. It has been designed specifically for mining land use data, based on `HMM2`. It is able to analyze temporal and spatial sequences of land use in a territory. Several models are available, we describe a few of them below. The CarottAge software is now used by agronomists – and also by geneticians for mining genomic data [29] – without any assistance of the designers.

[2] http://www.loria.fr/~ jfmari/App/

The functionalities of CarottAge

CarottAge get as an input preprocessed or transformed discrete data, represented within text files. Data mining is performed in four steps:

1. the editing of the initial model;
2. the iterative ML estimation using the Baum-Welch algorithm based on a corpus of sequences of observations;
3. the display of the model's parameters;
4. the display of the a posteriori transition probabilities.

The user has to write the initial model in two parts. The first part specifies the model's topology by means of a list of transitions between the states together with their relative weights. The second part defines the observations and gives the discrete probabilities over this set of observations. An example of a text file specifying a simple three states - left to right – self loops `HMM2`, where the three states have a uniform distribution, is described in Table 2.2.

```
 2  3  1
 3  4  1
 2  2  1
 3  3  1
 4  4  1
-1 -1 -1
equiprobable  #state 2
equiprobable  #state 3
equiprobable  #state 4
```

Table 2.2: Initial model (`lin3.mod`): the first lines describe the transitions (a line is structured like: <origin> <extremity> <weight>), the last lines describe the distributions associated to the states. Here the hidden states are called 2, 3 and 4. The distributions are uniform.

Non-uniform distributions can be also defined. Then the state is described with a list of observations and their probabilities as follows:

```
1 wheat #state n
```

which means that the state `n` contains only wheat, and that the probability of the other observations is null. CarottAge provides a program that builds a file containing the `HMM2` according to the text file used as input.

The model is then estimated on a corpus of sequences represented by a matrix of observations. A typical command line is:

```
estimate -n 3 lin3.mod -o lin3.mod1 lorraine.xls
```

The input file `lin3.mod` (cf. Table 2.2) is estimated using the corpus specified by the file `lorraine.xls`. Three iterations are performed. The resulting model is stored in the output file `lin3.mod1`. Actually, this file records the a posteriori transition probabilities (see Equation 2.10 in the Appendix) between the states, and the distributions of observations associated to the states.

A specific program has been developed for displaying the results of the model estimation (Figure 2.1). It displays both the model's parameters (especially the distributions) and the a posteriori probabilities of transitions between the states.

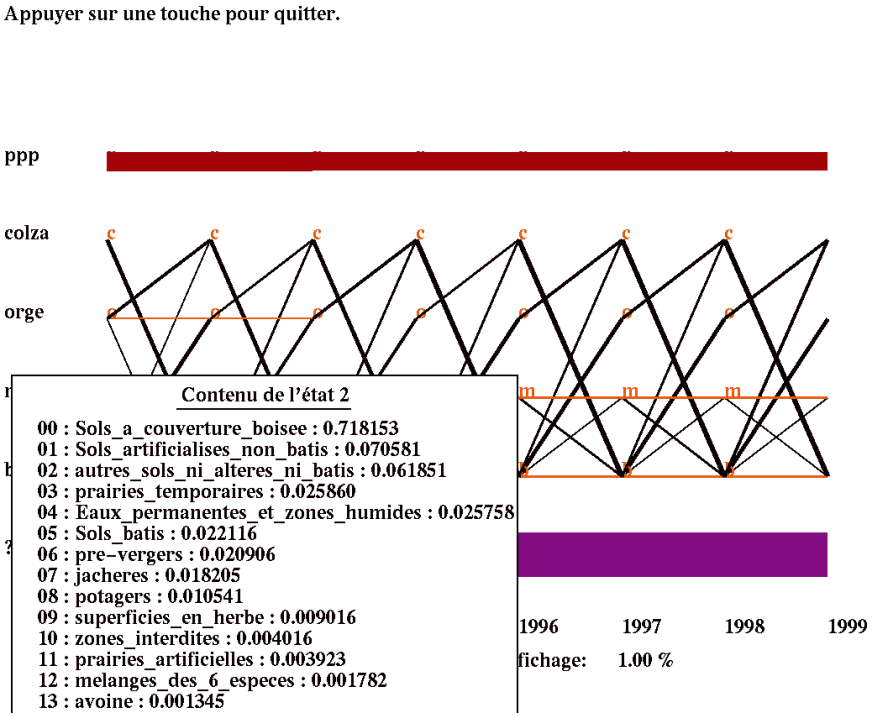


Figure 2.1: Displaying the results of CarottAge: the user can see the distributions associated to the states (table) and the a posteriori probabilities of transitions between states (diagonal and horizontal lines).

Models for mining land-use data

The role of the user is obviously crucial: it has to preprocess and transform the data, to define the initial model and to interpret the data mining results. Furthermore, these actions can be combined in a knowledge discovery process, where the data can be transformed in several ways and mined with various models. Actually, at the beginning of our work, the models were defined and experimented by the users (agronomists) and the computer scientists together [49]. Then, the agronomists used CarottAge by themselves, and designed their own models, as shown in Section 2.4.2.

A first model can be used for the extraction of temporal segments in which the distribution of the land-use categories is stationary. To do so, we have specified a `HMM2` with n states with a left to right, self loops topology (see Figure 2.2). This means that we attempt to capture n periods of evolution in the land use dynamics, where n is chosen according to the length of the period.

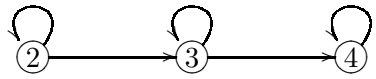


Figure 2.2: Model 1: the `HMM2` performs a data segmentation in three periods in which the observations are supposed stationary. This model is defined in Table 2.2.

The results of this model are displayed within a table where the evolution of the cropping pattern of a region is visible (Figure 2.3). Here we see that the pastures are dominant at the beginning of the period, and then decrease and are replaced by wheat at the end of the period, while the surface of rapeseed is continuously growing. This table is actually a synthetic view of the eight years (1992 – 99), pointing out the stable patterns and the main transitions.

Another model has been designed for measuring the probability of a succession of three[3] land-use categories. Actually, we have defined a specific state, called the *Dirac state*, whose distribution is zero except on a particular land-use category. Therefore, the transition probabilities between the Dirac states measure the probabilities between the land-use categories during a three years period. Figure 2.4 shows the topology of a `HMM2` that has two kinds of states: Dirac states associated to the most

[3]The number of steps is constrained by the memory of the `HMM2` (2).

state 2		state 3		state 4	
pastures	0.31	pastures	0.29	wheat	0.29
wheat	0.22	wheat	0.26	pastures	0.27
barley	0.16	rapeseed	0.14	rapeseed	0.17
rapeseed	0.12	barley	0.11	barley	0.12
maize	0.07	maize	0.08	maize	0.06
set-aside	0.05	set-aside	0.05	orchard	0.02

Figure 2.3: Viewing the results of model 1 applied on land-use data of the Lorraine Region (years 1992 – 1999).

frequent land-use categories (wheat, maize, barley, ...) and *container states* associated to uniform distributions over the set of observations. The estimation process usually empties the container state of the land-use categories associated with Dirac states.

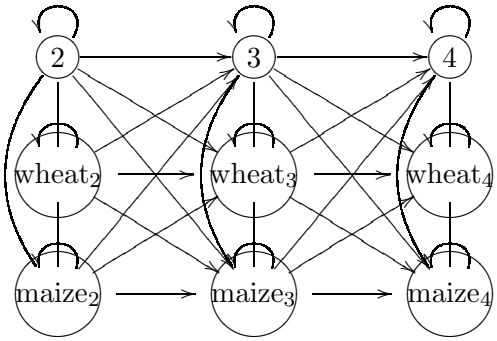


Figure 2.4: Model 2: the states denoted 2, 3 and 4 are associated to a distribution of land-use categories, as opposite to the states denoted with a specific land-use category. The number of columns determines the number of time intervals (periods). A connection without arrow means a two directional connection.

As results the user obtains a graphic showing the main transitions between Dirac and container states, i.e. the crop sequences in a region (Figure 2.5). The user can choose the resolution level, and see all transitions or only the main transitions. In the graphic shown figure 2.5, six crops have been individualized (the container state is denoted by **?**). The thickness of the lines represents the a posteriori probability of the transition between two crops (cf. Equation 2.10 in the Appendix). Diagonal lines mean that a

crop is followed by another crop, e.g. rapeseed (denoted by **colza**) to wheat (denoted by **ble**), while horizontal lines mean that a crop is followed by itself, e.g. pastures (denoted by **ppp** in the figure).

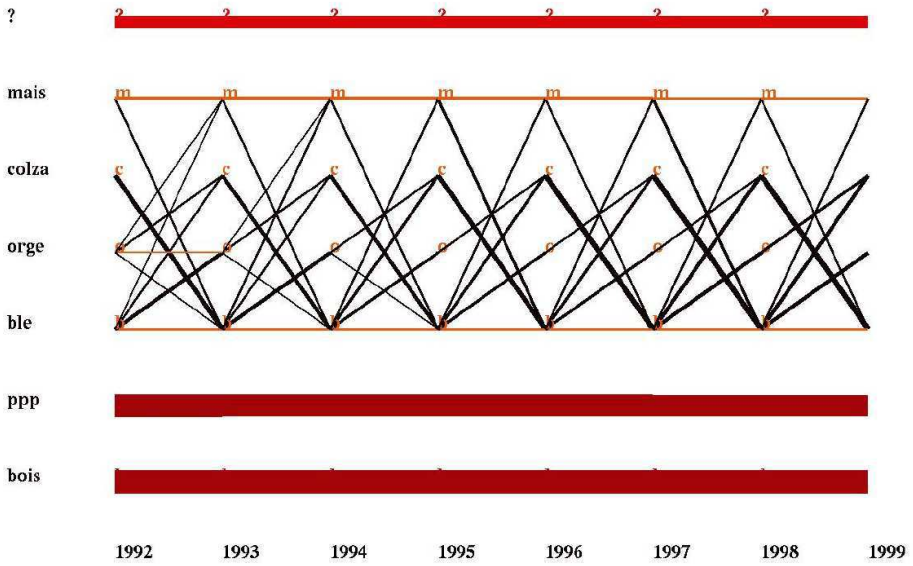


Figure 2.5: Viewing the results of model 2 applied on land-use data of the Lorraine Region (years 1992 – 1999).

These tables are very useful for seeing the evolution of the land use in a region and for comparing regions. The models can be used on crop data, but also on sequences of crops, and allow to produce sets of tables showing the evolution or stability of land use. Compared to `HMM1`, `HMM2` have the capability to model the transitions between Dirac states over a longer period –according to the farmer's practices–: three years compared to two years. Furthermore, these tables can be used as a support for field inquiries.
Finally, CarottAge allows the user to define various models, according to the data format and his purpose, as we see in the next part.

2.4 Using CarottAge for finding out crop sequences

2.4.1 The data base

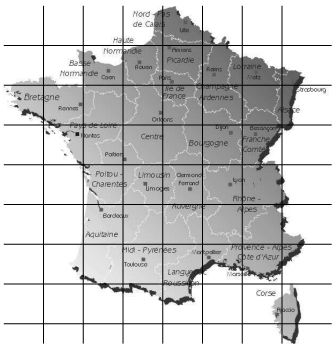
The *Ter-Uti* data are collected by the French agriculture administration on the whole metropolitan territory. They represent the land use of the country on a one year basis. Two levels of resolution are achieved (Figure 2.6). A first sample consists in selecting aerial photographs. The French territory is segmented into 3820 meshes. Each of the meshes contains four photographs that cover each one only a square of 2 km. Secondly, on each photography, a 6 by 6 grid determines 36 sites that are inquired every year in June. The land-use category of these sites (wheat, corn, potato, forest, rocks ...) is logged in a matrix in which the rows are the sites of the country and the columns the time slots (from 1992 to 2003). Finally, one *Ter-Uti* site represents roughly 100 hectares [42].

2.4.2 Analyzing crop sequences in the Seine Basin

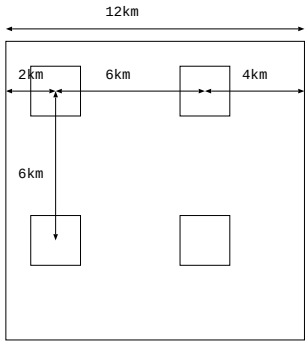
For thirty or forty years, the increasing human activities (domestic, industrial, agricultural) have gradually degraded the hydro-system of the Seine river, regarding water quality and biological population [50]. The nitrate contamination of groundwater and surface water is mainly caused by the evolution of agricultural activities, and related to their nature and to their organization inside the river watershed. The INRA team in Mirecourt is member of an interdisciplinary research program which aims to develop a tool for forecasting water quality in the Seine river watershed, based on assumptions upon agricultural changes. Thus, the INRA team analyses the agricultural activities in the watershed, their dynamics and their spatial organizations, focusing on the crop (temporal) rotations that are able to explain the risk of nitrate loss [52]. The data mining software CarottAge has been used on *Ter-Uti* data from the Seine watershed. Results are presented and analyzed below for a small district.

Crop sequences in Saint-Quentinois

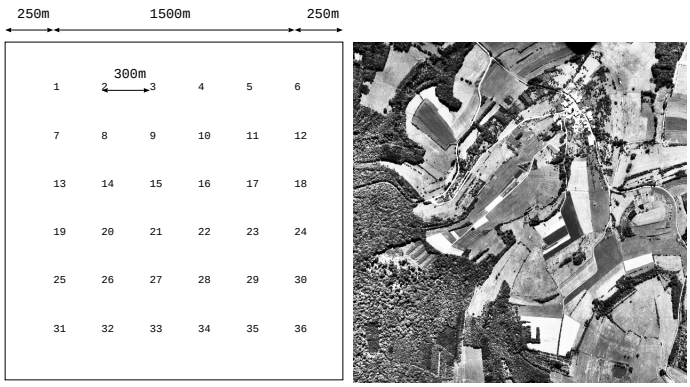
The diagram shown in figure 2.7 displays the main annual transitions between crops and their evolutions. The importance of the transition between two crops is expressed with the thickness of the line joining the two crops. One can see that, in this district, the wheat-based rotations are in a majority:



(a) the basic grid covering France territory



(b) the 4 air photos in a mesh



(c) an air photography and its 6x6 grid

Figure 2.6: Collecting the *Ter-Uti* data: 3820 meshes square France, 4 air photographs are sampled in a mesh, a 6x6 grid determines 36 sites.

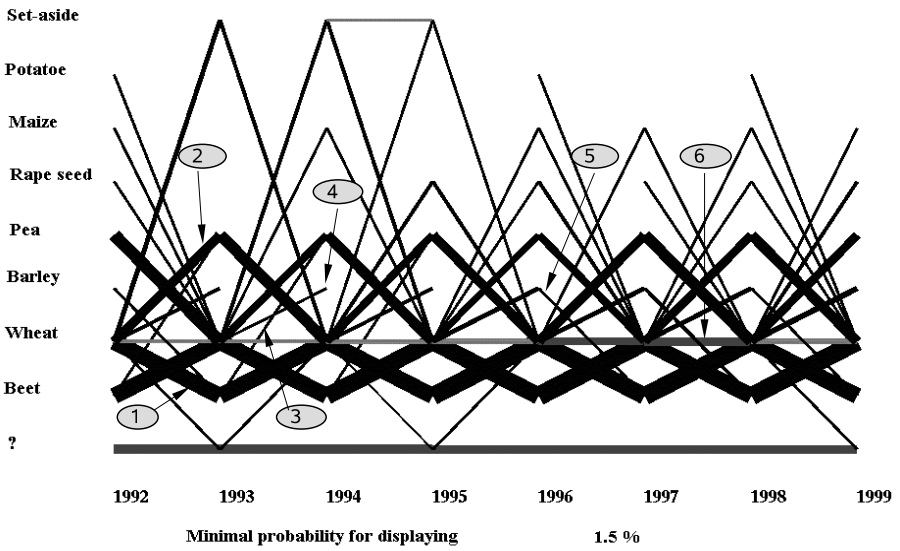


Figure 2.7: Crop transitions between 1992 and 1999 in the district of Saint-Quentinois (North-east of France). Only the transitions whose probability is greater than 1.5% are displayed. The question mark denotes the container state.

- the main transitions are wheat-beet-wheat (1) and wheat-pea-wheat (2), which have the thickest lines (bottom of the diagram).
- the transitions beet-pea (3) appear between 1992 and 1995 and then disappear.
- transitions like wheat-barley (4), or barley-beet (5), appear from 1996 (actually, they exist before 1996 but with a probability smaller than 1.5%).

One can also notice that the other crops, like rapeseed, maize, potatoes or set-aside are mainly followed or preceded with wheat. Furthermore, the transitions wheat-wheat (6) seem to grow between 1996 and 1998.

Three-crop sequences in Saint-Quentinois

In order to better examine the crop transitions, we transform the *Ter-Uti* data and apply CarottAge on tables representing couples, triples or even quadruples of crops. To minimize the data set, we have to select the main

rotations, based on our first analysis, e.g., for crop triples, wheat-beet-wheat, wheat-pea-wheat, etc.
Thus, we obtain a second diagram where the states represent triples of crops, which is more difficult to explain but confirms our first analysis (Figure 2.8).

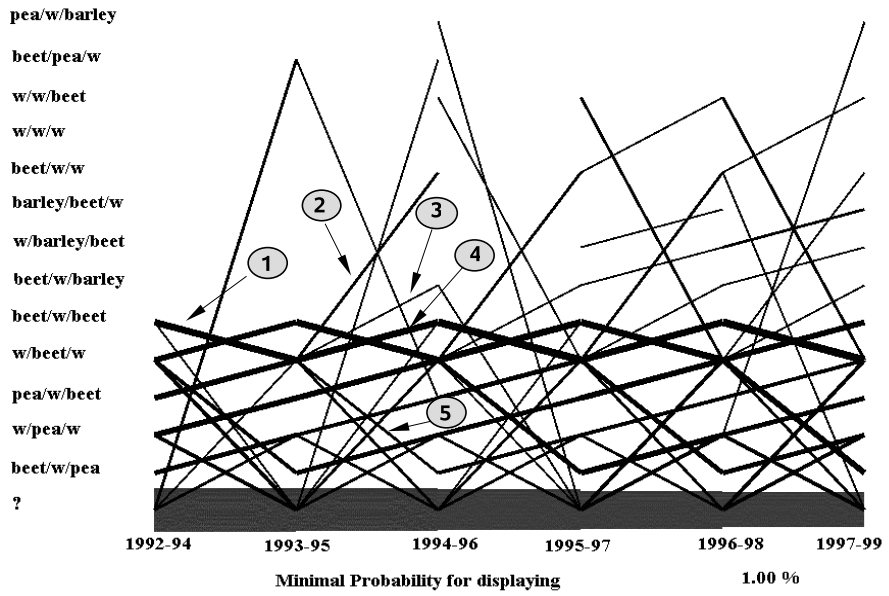


Figure 2.8: Transitions between triples of crops, 1992 – 1997, in the district of Saint-Quentinois. Only the transitions whose probability is greater than 1.0 are displayed. The question mark denotes the container state.

For example, if we follow the sequence of the crop triples starting from beet-wheat-beet in 1992–94, the main transition (1) leads to the wheat-beet-wheat triple. From this last triple, there are several possibilities: a first one goes towards the triple beet-wheat-wheat (2), a second one towards beet-wheat-barley (3), a third one, which has the greatest probability, towards beet-wheat-beet (4) and finally a fourth one towards beet-wheat-pea (5). Knowing that two triples are connected when they share two crops, we can synthesize the last transitions in the following way: (2) beet-wheat-beet-wheat-wheat, (3) beet-wheat-beet-wheat-barley, (4) beet-wheat-beet-wheat-beet, (5) beet-wheat-beet-wheat-pea.
Furthermore we notice a repeated pattern in this diagram, that looks like a

chain link: this pattern is composed with the repeated transitions between the triples wheat-beet-wheat and beet-wheat-beet, and reveals the existence of the quadruple succession beet-wheat-beet-wheat. Another pattern made of oblique lines can be found in this diagram: for example, the line starting from the beet-wheat-beet triple in 1992–94 connects to wheat-pea-wheat, then pea-wheat-beet, wheat-beet-wheat and finally to beet-wheat-beet or again to beet-wheat-pea. This connected sequence proves that all these triples belong to the same beet-wheat-pea-wheat four-crops succession.

Clustering the districts of the Seine Basin

The analysis of crop sequences and the determination of the main successions (double, triple or even quadruple successions, as shown before) are a basis for comparing and classifying agricultural territories. The small districts and the sub-watersheds of the Seine basin were compared and clustered thanks to statistical methods applied on the sets of crop triples that characterized each district or watershed. Finally we built a district typology clustering the similar districts *wrt* the crop successions.
More precisely, the analysis of *Ter-Uti* data in the Seine Basin was performed following these steps.

1. Determination of the main crop successions in each small district, using CarottAge (model 2) as explained in Sections 2.4.2 and 2.4.2. The whole basin was characterized with 64 3 or 4-crops successions, for 143 districts. The crop successions were clustered within 6 main categories, according to their agronomic function (cereals, break crops, etc.).

2. Computation of the distribution of the crop successions in each small district, using CarottAge (model 1). The districts were thus characterized –for a period– with sets of crop successions and their probabilities.

3. Analysis of the table *(districts × probabilities of crop-successions)* using the Principal Component Analysis method. The projections of the districts on the fifteenth first eigenvectors were used to design a new table with 15 variables characterizing the districts.

4. Clustering of the districts on the basis of this last table, using the *Hierarchical Ascendant Clustering* method. The districts were clustered within twenty classes, which represented a good

segmentation according to agronomists' expertise. The map of the Seine Basin, where the districts are colored according to this segmentation, is displayed in figure 2.9.

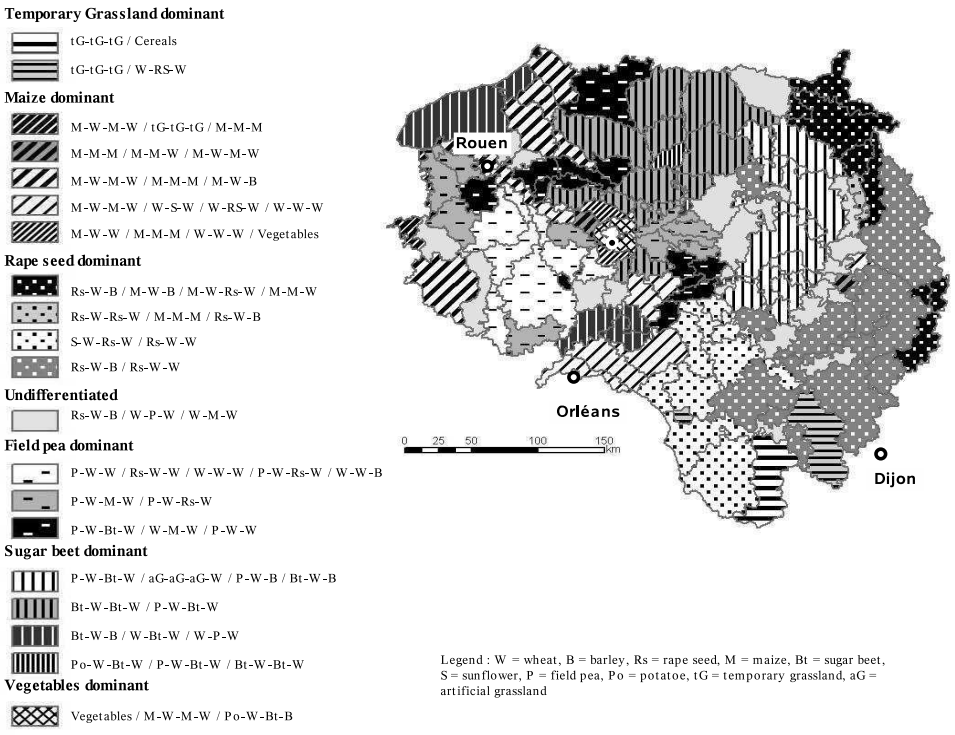


Figure 2.9: Map of the Seine Basin districts classified *wrt* their main crop sequences.

This map highlights the strong spatial structure of the distribution of crop sequences in the Seine Basin. This structure is to be related to big geological forms, as for example:

- The districts classified into "Rs-W-B dominant" are on the Jurassic calcareous plateaux (east of the basin).
- The districts of the classes "temporary grassland dominant" correspond to the granitic mountains of Morvan (south-east of the basin).

- The districts of the class "Bt-W-Bt-W dominant" occupy the silty plateaux of Picardy (north of the basin).

2.5 Conclusion

We claim that the concept of crop sequences is relevant and can be useful: it will help research on agriculture/environment relations by providing types of land use that convey the farmer's strategy, independently of year to year changes that characterize crop rotations; these types of land use are stable over several years and can be related, on one side to field characteristics and constraints, and on the other side to environmental effects. It will facilitate discussions between farmers and other actors of rural territories by setting a common language and allowing an objective description of agricultural land use types. This concept, by considering the middle-term strategy of the farmer, frees itself from the infinite diversity of actual crop successions and facilitates the comparison between fields similarly managed in different farms, and hence facilitates the extension of cropping system research to the territorial and multi-year scales, which are relevant to environmental questions.
The concept of agricultural land management system is a first step towards the precise description and classification of all types of land uses intervening in a region. In order to understand and manage the evolution of landscapes, it will be necessary to include non agricultural uses: forests, waters (in marshes, waters are subject to a particular type of collective management), roads and roadsides, etc.
With respect to this purpose, CarottAge has proven useful for exploring large land use data bases and for revealing the temporal and spatial organization of land use, based on crop sequences [48]. Furthermore, CarottAge can also be used to investigate and visualize the crop sequences of a few specific farms or of a small territory. Besides, the diagrams resulting from CarottAge, showing the main crop transitions, are good graphical supports for discussing the evolution of land use. For example, they have been used during regional farm surveys to collect the knowledge of farmers and agricultural technicians about crop sequences. Finally, the results of our analysis can be linked to models of nitrate flow and used for the evaluation of water pollution risks in a watershed [52].
To resume, crop sequences are a pragmatic research object useful to explain land use changes, and we propose to apply our analysis method and the CarottAge software to understand the recent changes and to

forecast the future new land uses [25, 10]. So, logically, our work will take place in the international project LUCC[4] [35].
To go further, we have to enlighten the farmers about the links between their objectives, their practices and the consequences of their practices [28]. A possible approach is to test different scenarios for the actors. Two types of scenarios may be developed based on the following argumentation: "What... if...", and "How... to...". Research methods to address these two types of scenarios taking into account the analysis of farmer practices and modeling of decision making are to be developed [1, 2].
The model-building process itself can serve as a tool to construct and discuss scenarios with the actors [16]. Two main model-building procedures are used: mathematical ones involving methods used in landscape ecology and linear programming, and graphic ones. We shall elaborate on the second procedure, since the first one is well known. For example, one research approach developed by geographers is to define a dictionary of spatial graphic symbols or *chorems* [14]. Using this form of qualitative modeling proves most useful in discussions with a wide number of people and enables us to build models of farmer practices in their spatial dimension [20]. A potential further development in this direction is the use of 3D visualization tools to facilitate the understanding of the land use and landscape changes (see [44] for an example).
To end with an ethical posture [32], we propose a new researcher behavior: investigating this type of issue we must not set out from the assumption that a farmer has voluntarily deteriorated the landscape parameter that is being investigated. This corresponds to the development of a *decision agriculture* [51] that is increasingly knowledge-based, and increasingly rooted in the information and communication sciences and technologies and to a sustainability trend with a new weight of land capabilities [59, 33]. We agree with [13]: "This does not, however, mean a technology-driven process of innovation, but on the contrary increased feedback of action and decision into the design of innovation..." mainly on land design management innovation!

2.6 Appendix: The Baum-Welch Algorithm

The Baum-Welch (or Forward-Backward) algorithm implements a `HMM2`'s estimation following the maximum likelihood estimation criteria. Since many state sequences may generate a given output sequence, the

[4]Land Use and Cover Changes.

probability that a model λ generates a sequence $O_1^T = o_1,...,o_T$ is given by the sum of the joint probabilities (given in equation 2.3, section 2.3.1) over all state sequences (*i.e.*, the marginal density of output sequences). To avoid combinatorial explosion, a recursive computation can be used to evaluate the above sum. The forward probability is defined for all $(j,k) \in [1,N]^2$ as:

$$\alpha_t(j,k) = \text{Prob}(q_{t-1} = s_j, q_t = s_k, O_1^t = o_1, ..., o_t). \tag{2.4}$$

This value represents the probability of starting from the initial state (s_1) and ending with the transition (s_j, s_k) at time t and generating output $o_1,...,o_t$ using all possible state sequences in between. The Markov assumption allows the recursive computation of the forward probability for $t \in [3,T]$ as follows:

$$\alpha_t(j,k) = \sum_{i=1}^{N} \alpha_{t-1}(i,j).a_{ijk}.b_k(o_t). \tag{2.5}$$

Without any loss of generality, we can suppose that s_N is the only final state, then the probability that the model generates the sequence $O_1^T = o_1, ..., o_T$ is $\text{Prob}(O_1^T = o_1, ..., o_T) = \sum_j \alpha_T(j,N)$. Another useful quantity is the backward function $\beta_t(i,j)$, defined as the probability of the partial observation sequence from $t+1$ to T, given the transition (s_i, s_j) between times $t-1$ and t. It can be expressed for all $t \in [2, T-1]$ and for all $(i,j) \in [1,N]^2$ by:

$$\beta_t(i,j) = \text{Prob}(O_{t+1}^T = o_{t+1}, ..., o_T / q_{t-1} = s_i, q_t = s_j). \tag{2.6}$$

The Markov assumption allows also the recursive computation of the backward probability as:

1. Initialization
$$\beta_T(i,j) = 1 \quad \forall (i,j) \in [1,N]^2$$

2. Recursion for $T-1 \geq t \geq 1$
$$\beta_t(i,j) = \sum_{k=1}^{N} \beta_{t+1}(j,k).a_{ijk}.b_k(o_{t+1}) \quad \forall (i,j) \in [1,N]^2. \tag{2.7}$$

Given an observation sequence $o_1, ..., o_T$, we define for all $t \in [2, T-1]$ and for all $(i,j,k) \in [1,N]^3$, the value $\eta_t(i,j,k)$ as the probability of the

transition $s_i \longrightarrow s_j \longrightarrow s_k$ between $t-1$ and $t+1$ during the emission of the observation sequence:

$$\eta_t(i,j,k) = \text{Prob}(q_{t-1} = s_i, q_t = s_j, q_{t+1} = s_k / O_1^T = o_1, ..., o_T). \tag{2.8}$$

We deduce for all $t \in [2, T-1]$ and for all $(i,j,k) \in [1,N]^3$,

$$\eta_t(i,j,k) = \alpha_t(i,j) a_{ijk} b_k(o_{t+1}) \beta_{t+1}(j,k) \,/\, \text{Prob}(O_1^T = o_1, ..., o_T). \tag{2.9}$$

As in the first order, we define
$\text{Prob}(q_{t-1} = s_i, q_t = s_j / O_1^T = o_1, ..., o_T) = \xi_t(i,j)$ as the a posteriori probability that the stochastic process accomplishes the transition $s_i \rightarrow s_j$ between $t-1$ and t assuming the whole sequence. We obtain for all $t \in [2, T-1]$ and for all $(i,j) \in [1,N]^2$:

$$\xi_t(i,j) = \sum_{k=1}^{N} \eta_t(i,j,k). \tag{2.10}$$

When the training corpus is a set of sequences, we sum $\xi_t(i,j)$ over this set and plot this value as a function of t (i and j are dropped in the Y-axis). This illustrates the behavior of the stochastic process between states s_i and s_j at time t (see Figure 2.5).
The second-order ML estimate of $\overline{a_{ijk}}$ is given by the equation:

$$\overline{a_{ijk}} = \sum_t \eta_t(i,j,k) \,/\, \sum_{k,t} \eta_t(i,j,k). \tag{2.11}$$

If N is the number of states and T the sequence length, the Baum-Welch algorithm has a complexity of $N^3 \times T$ for a `HMM2`. Interested readers may refer to [21, 47] to find more specific details of the implementation of this algorithm.

Bibliography

[1] F. Affholder, P. Bonnal, D. Jourdain, and E Scopel. Small scale farming diversity and bioeconomic variability: a modelling approach. In *Proceedings of the 15th International Symposium of the association for farming systems research-extension. Pretoria*, pages 952–959, 1998.

[2] J.-M. Attonaty, M.-H. Chatelin, and F. Garcia. Interactive simulation modelling in farm decision-making. *Computers and Electronics in Agriculture*, 22:157–170, 1999.

[3] C. Aubry, F. Papy, and A. Capillon. Modelling decision-making processes for annual crop management. *Agricultural systems*, 56:45–65, 1998.

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

[5] P. Baudoux, G. Kazenwadel, and R. Doluschitz. On-farm effects and farmer attitudes towards agri-environmental programs: a case study in baden-wüttemberg. *Études et Recherches sur les Systèmes Agraires et le Développement*, 1998:333–356, 1998. Brossier J., Dent B. (eds) : Gestion des exploitations et des ressources rurales. Entreprendre, négocier, évaluer. Farm and Rural Management. New context, new constraints, new opportunities.

[6] B. Benmiloud and W. Pieczynski. Estimation des paramètres dans les chaînes de Markov cachées et segmentation d'images. *Traitement du signal*, 12(5):433 – 454, 1995.

[7] Marc Benoît, Florence Le Ber, and Jean-François Mari. Recherche des successions de cultures et de leurs évolutions : analyse par HMM des données Ter-Uti en Lorraine. *Agreste Vision - La statistique agricole*, (31):23–30, June 2001.

[8] M. Benoît, J.-P. Deffontaines, F. Gras, E. Bienaimé, and R. Cosserat. Agriculture et qualité de l'eau. une approche interdisciplinaire de la pollution par les nitrates d'un bassin d'alimentation. *Cahiers Agricultures*, 6:97–105, 1997.

[9] M. Benoît, J.-L. Fiorelli, P. Morlon, and Y. Pons. Technical management: a central point for agronomy challenge. First European Congress of Agronomy, Session V-01. Paris, 1990.

[10] M. Benoît and M.C. Muhar. Farmers, landuse and groundwater quality: an interdisciplinary approach. Congress "Future of the Land", Wageningen, 1993.

[11] M. Benoît and F. Papy. La place de l'agronomie dans la problématique environnementale. *Les dossiers de l'environnement de l'INRA*, 17:53–62, 1998.

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

[13] J. Boiffin, E. Malézieux, and D. Picard. Cropping systems for the future. 3rd International crop science congress, 17-22 august 2000, Hambourg (Germany), 2000. 13 pages.

[14] R. Brunet. La carte-modèle et les chorèmes. *Mappemonde*, 86(4):3–6, 1986.

[15] F. Burel and J. Baudry. Hedgerow network patterns and processes in france. In Zonneveld and Forman, editors, *Changing Landscape: an ecological perspective*, pages 99–120. Springer, 1990.

[16] P.G. Cox. Some issues in the design of agricultural decision support systems. *Agricultural systems*, 52:355–381, 1996.

[17] S. Dabbert, S. Herrmann, G. Kaule, and M. Sommer, editors. *Landschafts-modellierung für die Umweltplanung*. Springer Verlag, 1999. 260 pages.

[18] S. Dabbert, S. Herrmann, T. Vogel, T. Winter, and H. Schuster. Socio-economic analysis and modelling of agricultural water demands and land use. In *German Programme on Global Change in*

Hydrological Cycle Status Report 2002 (Phase I, 2000-2003). 2002. 55 pages.

[19] C.T. de Wit. Resource use efficiency in agriculture. *Agricultural Systems*, 40:125–151, 1992.

[20] J.-P. Deffontaines, J.-P. Cheylan, S. Lardon, and H. Théry. Managing rural areas. From pratices to model. In J. Brossier, L. de Bonneval, and E. Landais, editors, *Systems studies in agriculture and rural development*, Science Update, pages 383–392. INRA, Paris, 1994.

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

[22] M. Dunham. *Data Mining*. Prentice Hall, 2003.

[23] U. Fayard, G. Piatesky-Shapiro, P. Smyth, and R. Uthurusamy, editors. *Advances in Knowledge Discovery in Data Minig*. AAAI/MIT Press, 1996.

[24] Shai Fine, Yoram Singer, and Naftali Tishby. The Hierarchical Hidden Markov Model: Analysis and Applications. *Machine Learning*, 32:41 – 62, 1998.

[25] L. Fresco, editor. *Future of the Land*. Wageningen, 1993.

[26] L. Fresco. Crop science: scientific and ethical challenge to meet human need. In *3rd International crop science congress, 17-22 august 2000, Hambourg (Germany)*, 2000. 11 pages.

[27] F. Gaury. Systèmes de culture et teneurs en nitrates des eaux souterraines : Dynamique passée et actuelle en région de polyculture élevage sur le périmètre d'un gîte hydrominéral. Doctorat de l'Ecole Nationale Supérieure Agronomique de Rennes, 1992.

[28] R. Gras, M. Benoît, J.-P. Deffontaines, M. Duru, M. Lafarge, A. Langlet, and P.-L. Osty. *Le fait technique en agronomie. Activité agricole, concepts et méthodes d'étude.* INRA - L'Harmattan, 1989. 160 pages.

[29] Sébastien Hergalant, Bertrand Aigle, Bernard Decaris, Jean-Francois Mari, and Pierre Leblond. HMM, an Efficient Way to Detect Transcriptional Promoters in Bacterial Genomes. In *European*

Conference on Computational Biology - ECCB'2003, Paris, France, pages 417–419, Sep 2003. poster in conjonction with the french national conference on Bioinformatics (JOBIM 2003).

[30] S. Herrmann, S. Dabbert, and H.-G. Schwarz von Raumer. Ecological threshold values as indicators for biodiversity - economic and ecological consequences. *Agriculture, Ecosystems and Environment*, pages 493–50, 2003.

[31] J. Adibi and W-M. Shen. Self Similar Layered Hidden Markov Model. In *5th European Conference on Principles of Knowledge Discovery in Databases*, Freiburg, Germany, September 2001.

[32] H. Jonas. *Le principe de responsabilité. Une éthique pour la civilisation technologique "Das Prinzip Verantwortung"*. Editions du Cerf, Paris, 1990 (1979). 336 pages.

[33] Michael Jordan and Zoubin Ghahramani. Factorial Hidden Markov Models. *Machine Learning*, 29(2 – 3):245 – 273, November 1997.

[34] D.L. Karlen, G.E. Varvel, D.G. Bullock, and R.M. Cruise. Crop rotations for the 21st century. *Advances in Agronomy*, 1994.

[35] E.F. Lambin, X. Baulies, N. Bockstael, G. Fischer, T. Krug, and *et al.* Land-use and land-cover change (LUCC): implementation strategy. IGBP Rep. 48, IHDP Rep. 10, Int. Geosph.-Biosph. Program., Int. Hum. Dimens. Glob. Environ. Change Program, 1999. Stockholm/Bonn.

[36] E.F. Lambin, H.J. Geist, and E. Lepers. Dynamics of land-use and land-cover change in tropical regions. *Annual Review of Environment and Resources*, 28:205–241, 2003.

[37] S. Lardon, J.-P. Deffontaines, J. Baudry, and M. Benoît. L'espace est aussi ailleurs. In J. Brossier, B. Vissac, and J.-L. Le Moigne, editors, *Modélisation systémique et système agraire. Décision et organisation*, pages 321–337. INRA, Paris, 1990.

[38] F. Le Ber and 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.

[39] F. Le Ber, M. Benoît, C. Schott, J.-F. Mari, and C. Mignolet. Studying Crop Sequences With CarrotAge, a HMM-Based Data Mining Software. *Ecological Modelling*, 191(1):170 – 185, Jan 2006.

[40] P.-Y. Le Gal and F. Papy. Coordination processes in a collectively managed cropping system: double cropping of irrigated rice in senegal. *Agricultural systems*, 57:135–159, 1998.

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

[42] M. Ledoux and S. Thomas. De la photographie aérienne à la production de blé. *AGRESTE, la statistique agricole*, (5), 1992.

[43] G. Lemaire and B. Nicolardot, editors. *Maîtrise de l'azote dans les agrosystèmes*. INRA Éditions, Paris, 1997. 333 pages.

[44] A. Lovett, S. Herrmann, K. Appleton, and T. Winter. Landscape modelling and visualisation for environmental planning in intensive agricultural areas. In E. Buhmann and S. Ervin, editors, *Trends in Landscape Modeling*, pages 114–122. Wichmann, Heidelberg, 2003.

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

[46] Jean-François Mari, El-Ghali Lazrak, and Marc Benoît. Time space stochastic modelling of agricultural landscapes for environmental issues. *Environmental modelling & software*, 46:219–227, August 2013. http://hal.inria.fr/hal-00807178/PDF/arpentage_hal.pdf.

[47] Jean-François Mari and René Schott. *Probabilistic and Statistical Methods in Computer Science*. Kluwer Academic Publishers, January 2001.

[48] Jean-Francois Mari and Florence Le Ber. Temporal and spatial data mining with second-order hidden markov models. In Mohamed Nadif, Amedeo Napoli, Eric San Juan, and Alain Sigayret, editors, *Fourth International Conference on Knowledge Discovery and Discrete Mathematics - Journées de l'informatique Messine - JIM'2003, Metz, France*, pages 247–254. IUT de Metz, LITA, INRIA, Sep 2003.

[49] J.F. Mari, F. Le Ber, and M. Benoît. Fouille de données agricoles par modèles de markov cachés. In *IC'2000, Journées Francophones d'Ingénierie des Connaissances, Toulouse*, pages 197–205. AFIA, ERSS, IRIT, GRACQ, 2000.

[50] M. Meybeck, G. De Marsilly, and E. Fustec. *La Seine en son bassin, fonctionnement d'un système fluvial anthropisé*. Elsevier, 1998. 750 pages.

[51] B.J. Miflin. Sugar beet production: strategies for the future. In *Proceedings of the 60th IIRB Congress*, pages 253–262. IIRB, Brussels, 1997.

[52] C. Mignolet, C. Schott, and 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(2004):219–236, 2004.

[53] P. Morlon and M. Benoît. Étude méthodologique d'un parcellaire d'exploitation agricole en tant que système. *Agronomie*, 6:499–508, 1990.

[54] M. Sebillotte. Some concepts for analysing farming and cropping systems and for understanding their different effects. In A. Scaife, editor, *Proceedings of the first Congress of European Society of Agronomy, Colmar*, volume 5, pages 1–16. European Society of Agronomy, 1990.

[55] M. Sebillotte. Système de culture, un concept opératoire pour les agronomes. In L. Combe and D. Picard, editors, *Les systèmes de culture*, pages 165–196. INRA éditions, Paris, 1990.

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

[57] P.M. van Dijk, F.J.P.M. Kwaad, and M. Klapwijk. Retention of water and sediment by grass strips. *Hydrological Processes*, 10(8):1069–1080, 1996.

[58] P.M. Van Dijk, M. Van der Zijp, and F.J.P.M. Kwaad. Soil erodibility parameters under various cropping systems of maize. *Hydrological Processes*, 10(8):1061–1067, 1996.

[59] P. Vereijken. A methodic way to more sustainable farming systems. *Netherlands Journal of Agricultural Science*, 40:209–223, 1992.