Time space stochastic modelling of agricultural landscapes for environmental issues

Jean-François Mari, El-Ghali Lazrak, Marc Benoît

Time space stochastic modelling of agricultural landscapes for environmental issues

Jean François Mari [*loria,cnrs,inria], El Ghali Lazrak [†inra], and Marc Benoît [‡inra]

[loria] *Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France*
[cnrs] *CNRS, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France*
[inria] *Inria - Nancy - Grand-Est, Villers-lès-Nancy, F-54600, France*
[inra] *Inra, SAD-ASTER, UPR 055, Mirecourt, F-88500, France*

Also in Environmental Modelling and Software, DOI information : 10.1016/j.envsoft.2013.03.014

Résumé

Since the initial point of [Lan93] saying that Geographic Information Systems (GIS) were poorly equipped to handle temporal data, many researchers have sought to integrate the time dimension into GIS [RHS01]. We present a time space modelling approach – and a generic software named ARPENTAGE– capable of clustering a territory based on its pluriannual land-use organization. By adding the ability to represent, locate and visualize temporal changes in the territory, ARPENTAGE provides tools to build a Time-Dominant GIS. One main Markovian assumption is stated : the land-use succession in a given place depends only on the land-use successions in neighbouring plots. By means of stochastic models such as a hierarchical hidden Markov model and a Markov random field, ARPENTAGE performs an unsupervised clustering of a territory in order to reveal patches characterized by time space regularities in the land-use successions. Two case studies are developed involving two territories carrying environmental issues. Those territories have various sizes and are parameterized using long term surveys and/or remote sensing data. In both cases, ARPENTAGE detects, locates and displays in a GIS the temporal changes. This gives valuable information on the spatial and time dynamics of the land-use organization of those territories.

keywords : Markov random field, MRF , second-order HMM , data mining , landscape organization , land-use successions , temporal GIS , Hierarchical Hidden Markov Model

[*] jfmari@loria.fr
[†] lazrak@mirecourt.inra.fr
[‡] benoit@mirecourt.inra.fr

1 Software availability

Name : ARPENTAGE

Programming language : C++

Libraries used : Gnu STL, shapelib, gen2shp, txt2dbf

Inputs : csv files holding landscape raster representation : Lambert conformal conic coordinates (Tab. 2) or 2 level sampling TER-UTI data (Tab. 3)

Outputs : ESRI shapefiles and DBF files

User interface : Unix / Windows scripts files

Availability : http ://www.loria.fr/~jfmari/App/

Licence : Gnu GPL.

Demo : http ://www.loria.fr/~jfmari/App/Arpentage/demo.zip

2 Introduction

Stochastic modelling is a convenient way of building statistical and probabilistic models for capturing the spatial and temporal variability that is not yet fully understood [HK08, SMDF+12] especially in all alive processes. In agricultural landscapes, land-use (LU) seem randomly distributed among different agricultural fields (plots) managed by farmers. Nevertheless, the landscape spatial organization and its temporal evolution reveal at various scales the presence of logical processes and driving forces related to the soil, climate, cropping system, and economical pressure whose understanding is a major challenge mainly for landscape agronomists [BRM+12]. The data mining approach involving spatial and temporal clustering methods to get a landscape description in terms of land-use patterns has already demonstrated its capabilities in knowledge extraction [LMB09, SLM+12]. Such a description is useful in various areas : [For95] has demonstrated that a concise description of the mosaic of plots in terms of patch arrangement is necessary for ecologists to understand the relationship between landscape organization and species flows or biotic diversity. [JSMP06] use such a plot mosaic description to lower runoff on agricultural land by spatially alternating different crops at the catchment level. This description is also of interest in landscape governance [SLOW11] issues because land-use location influences the assessment of the visual aesthetic of a landscape.

This paper presents a method – and a generic software named ARPENTAGE (*Analyse de Régularités dans les Paysages : Environnement, Territoires, Agronomie* or "Landscape Regularities Analysis : Environment, Territory, Agronomy", *arpenter* is a French verb meaning "to survey") – capable of clustering territories of various sizes into patches based on their pluri-annual LU organization. It provides a Geographic Information System (GIS) with a description of the main time changes in the landscape together with their localizations. The scope of this software is not restricted to agriculture but may extend to other fields whenever it comes to locate sequences in space like in time space epidemic or ecological species surveillance. It implements a Markov random field of sequences whose parameters can be estimated based on a stream of time space data : long term surveys or remote sensing data.

This paper is organized as follows : section 3 presents the stochastic models that ARPENTAGE implements : second-order Hidden Markov Models (`HMM2`),

Hierarchical Hidden Markov Models (HHMM2), and Markov Random Fields (MRF). Section 4 describes the method used by ARPENTAGE to cluster a 3-D stream of data representing a time sequence of landscapes. Section 5 evaluates ARPENTAGE on two different annual landscape raster representations : 2 level resolution surveys and remote sensing data. Section 6 compares ARPENTAGE with similar software programmes. Finally, Section 7 focuses on ARPENTAGE in the framework of temporal GIS.

3 Temporal and spatial modelling background

ARPENTAGE relies on a stochastic Markovian principle to model time space landscape regularities. In short, this framework is based on the two following assumptions in spatial and temporal domains respectively :

- the Markov random field assumption assumes that the land-use of a given field depends only on the land-use of the neighbouring fields ;
- the Markov chain assumption assumes that the land-use of a given field in a year depends only on the land-use of the recent previous years in the same field.

Therefore, these two assumptions may be summarized by assuming in turn that the land-use succession of a given field only depends on the land-use successions in the neighbouring fields.

3.1 Hidden Markov Models

A Hidden Markov Model is a Bayesian network which represents the sequence of observations as a doubly stochastic process : an underlying "hidden" process, called the state sequence of random variables $Q_0, Q_1, Q_2, \ldots Q_T$ and an output (observation) process, represented by the sequence of random variables $O_0, O_1, O_2, \ldots O_T$ over the same time interval.

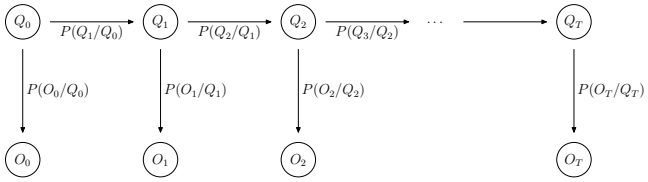


FIGURE 1 – Conditional dependencies in an HMM1 represented as a Bayesian network. The hidden variables (Q_t) govern the observable variables (O_t)

We define a discrete hidden Markov model (HMM) by giving :

- $\mathcal{E} = \{e_1, e_2, \ldots, e_K\}$, a finite set of K states that are the outcomes of Q_t ;
- $\mathbf{A}$ a matrix defining the transition probabilities between the states. These probabilities are time independent.
  - $\mathbf{A} = (a_{ij})$ for a first-order HMM (HMM1). a_{ij} is the probability $P(Q_t = e_j / Q_{t-1} = e_i), \quad \forall t = 1, T$ that the Markov chain is in state e_j at index t assuming it was in state e_i at index $t-1$ (see Fig. 1) ;
  - $\mathbf{A} = (a_{ijk})$ for a second-order HMM (HMM2). a_{ijk} is the probability $P(Q_t = e_k / Q_{t-1} = e_j, Q_{t-2} = e_i), \quad \forall t = 2, T$ that the Markov chain is in

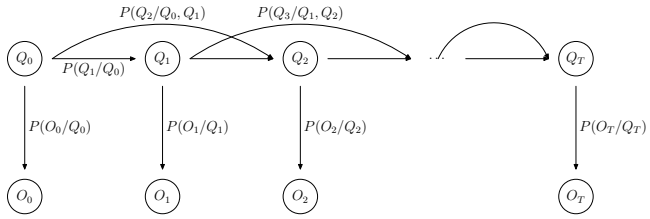


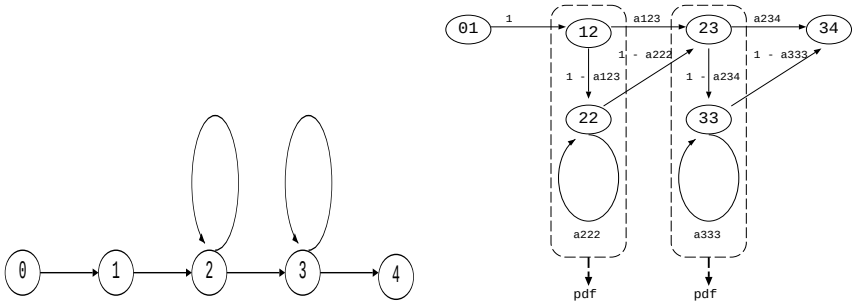
FIGURE 2 – Conditional dependencies in an HMM2 represented as a Bayesian network

state e_k at index t assuming it was in state e_j at index $t-1$ and e_i at index $t-2$ (see Fig. 2) ;

- $\mathcal{O} = \{o_1, o_2, \dots o_L\}$ a set of L observations that are the outcomes of O_t ;
- $\mathcal{B} = \{b_1(), b_2(), \dots, b_K()\}$ a set of K probability density functions (pdf) over $\mathcal{O}$, each of them being associated to a state $e_i, \quad i = 1, K$.

3.1.1 HMM2 properties

Each second-order Markov model has an equivalent first-order model on the 2-fold product space $\mathcal{E} \times \mathcal{E}$ but going back to first-order increases dramatically the number of states. For instance, figure 3(b) shows the equivalent HMM1 associated with the HMM2 depicted in figure 3(a). In this model the states in the same column share the same pdf.



(a) original second-order model (b) first-order equivalent model

FIGURE 3 – Decreasing the order of a HMM2.

The transition probabilities determine the characteristic of the state duration model. In a HMM1, whose topology is depicted in the figure 3(a) : linear, left-to-right, self-loop, the probability $d_j(l)$ that the stochastic chain loops l times in the state j follows a geometric law of parameter a_{jj} :

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

In the model depicted in figure 3(b), in which the successive states are indexed by $i = j-1,\ j,\ k = j+1$, the duration in state e_j may be defined as :

$$d_j(0) \quad = \quad 0 \tag{2}$$

$$
\begin{aligned}
d_j(1) &= a_{ijk}, \quad i \neq j \neq k \\
d_j(n) &= (1 - a_{ijk}) \times a_{jjj}^{n-2} \times (1 - a_{jjj}), \quad n \geq 2.
\end{aligned}
$$

These models achieved interesting results in pattern recognition and knowledge extraction in areas such as : speech recognition [MHK97, PBW98, EdP10], hydrology [LABP12], biology [EAA+09, ETD+11] and agronomy [MLB06, LBBS+06].

3.1.2 Hierarchical hidden Markov Models

We define a discrete hierarchical hidden Markov model (`HHMM`) as a `HMM` whose states are `HMM` [FST98]. Therefore, a second-order hierarchical `HMM` (`HHMM2`) is a 2-level hierarchical hidden Markov model in which the main `HMM` is a `HMM2` (see Fig. 4).

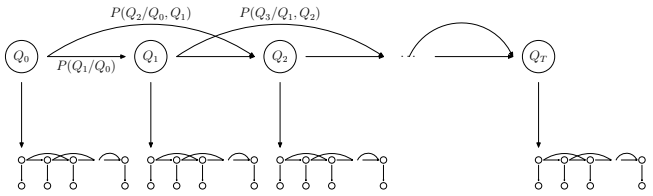


FIGURE 4 – 2-level second-order hierarchical hidden Markov model (`HHMM2`) represented as a Bayesian network. The observation probabilities are given by the `HMM2` depicted in Fig. 2

The second-order Hidden Markov Models implement an unsupervised training algorithm – the Baum-Welch algorithm [DLR77] – that can tune the `HHMM2` parameters from a corpus of observations in order to fit the model to the observations. The estimated model enables to segment each sequence in stationary and transient parts and to build up a classification together with its *a posteriori* probability

$$
P\left(Q_0^T = q_0^T \mid O_0^T = o_0^T\right), \tag{3}
$$

while the uncertainty of the class assignment of observation o_t in class e_k is measured by the *a posteriori* probability

$$
P\left(Q_t = e_k \mid O_0^T = o_0^T\right), \quad t = 0, T, \quad e_k \in \mathcal{E}. \tag{4}
$$

3.2 Stochastic spatial modelling

In the space domain, the `MRF` theory is an elegant mathematical way for accounting neighbouring dependencies [GG84, Bes86] between plots. A landscape representation is given by a set $\mathcal{S}$ of sites (*eg* plots) and a relation of neighbourhood on $\mathcal{S}$ (Fig. 5). $\mid \mathcal{S} \mid$ denotes the number of sites and $\mathcal{N}(i)$ the set of neighbours of site i. As in section 3.1, we call $\mathcal{E} = \{e_1, e_2, \ldots, e_K\}$, a set of K different classes that will play the role of patches. $Z_i = e_k$ means that site i is assigned to class e_k. The collection of outcomes $\{Z_i = z_i\}$ is called a *configuration*. In the following, the random variables Z_i will belong to $\mathcal{R}^K$. In particular, e_k is a binary vector of $\mathcal{R}^K$ having its k^{th} component set to 1, all the others being 0.

3.2.1 The Potts model with external field

In a Potts model with external field, a unique parameter $\beta > 0$ controls the pair-wise interaction – aggregation *versus* dispersion – between the patches whereas an additional vector V_i weights the values of z_i. The probability of a configuration $Z = z$ is given by :

$$P(Z = z) = \frac{\exp\left(-\sum_{i \in S}\left[z_i^t V_i - \beta \sum_{j \in \mathcal{N}(i)} z_i^t z_j\right]\right)}{W}.$$

W is a normalizing factor involving all the possible configurations. Its computation is intractable, hence the need of approximations such as the mean field approximation. z^t denotes the transpose of vector z and the product $z_i^t z_j$ is equal to 1 if the sites i and j are in the same class, 0 otherwise.

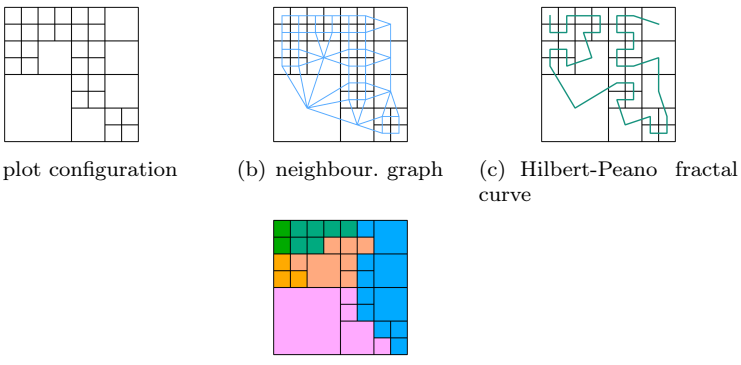
3.2.2 The Mean Field approximation

The mean field theory applied to `MRF` provides an approximation of the distribution of a `MRF` that allows the design of fast algorithms in image segmentation. In this theory, the class assignment probabilities of the neighbours of site i are set constant and replaced by their mean value. In this framework, [CFP03] introduce the self-consistency equation :

$$P_i^{mf}(e_s) = \frac{\exp\left[-V_i(e_s) + \beta \sum_{j \in \mathcal{N}(i)} P_j^{mf}(e_s)\right]}{\sum_{k=1}^{K} \exp\left[-V_i(e_k) + \beta \sum_{j \in \mathcal{N}(i)} P_j^{mf}(e_k)\right]}, \tag{5}$$

$V_i(e_k)$ being the k^{th} component of V_i. This equation says that the mean computed based on the mean field approximation must be equal to the mean used to define the approximation.

3.3 Approximation of a `MRF` by a `HMM`



(a) plot configuration (b) neighbour. graph (c) Hilbert-Peano fractal curve

(d) LU allocation

FIGURE 5 – Simple landscape and its neighbourhood graph

`HMM` can approximate efficiently `MRF` [BP95, GP97] by means of a Hilbert-Peano fractal curve (cf. Fig. 5(c)) that introduces a total order in the lattice of

sites [Ska92, DCOM00]. The 2-D landscape is first sampled using a $2^n \times 2^n$ grid. A scan is next performed using the Hilbert-Peano curve. To take into account the irregular neighbour system, the variable plot size and the overall landscape shape, we adjust the fractal depth by removing the fractal motifs lying entirely in a plot. For example, figure 5(c) shows two successive merging in the bottom left field that yield to the agglomeration of 16 points. The "blank" pixels in the $2^n \times 2^n$ image that are not in the landscape are assigned to the same "blank" plot and are partly removed in such a way. Two successive points in the fractal curve represent two neighbouring points in the landscape but the opposite is not true, nevertheless, this rough modelling of the neighbourhood dependencies has shown interesting results compared to an exact Markov random field modelling [GP97].

4 ARPENTAGE description

ARPENTAGE is based on CARROTAGE [LBBS+06] : a data mining toolbox for mining temporal data. Therefore, these two software programmes share a great part of code. They have the same programs to edit and train the `HMM2`. ARPENTAGE produces ESRI shapefiles that represent the landscape by means of a mosaic of patches, each of them being characterized by a temporal `HMM2` that models the temporal dynamics. ARPENTAGE takes advantage of the CARROTAGE graphic user interface facilities to display the temporal changes involved in the extracted clusters.

An elementary observation can range from a LU (such as cereals in the Yar watershed case study) or a LU category (such as Wheat in the Seine watershed case study) to a fixed length LU succession spanning several years (usually 2, 3 or 4) on a plot. In the latter case, the observation time sequence over the study period is made of overlapping LU sub sequences. The length of the LU succession influences the interpretation of the final model. However, the total number of LU successions is a power function of the succession length, and memory resources required during the estimation of `HMM2` parameters increase dramatically. The user defines the LU categorization in a configuration file (box 2 in Fig. 6).

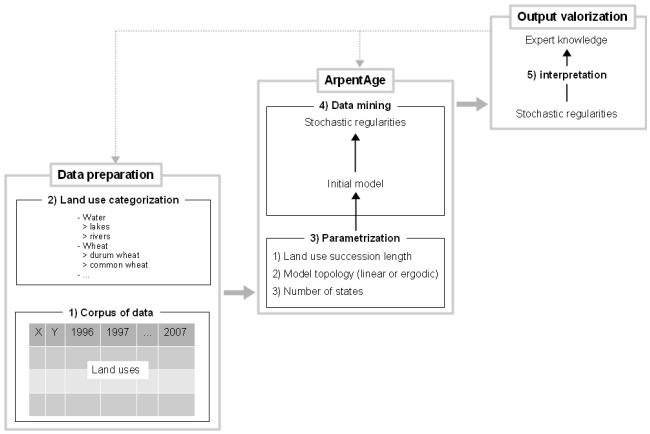


FIGURE 6 – Mining data with ARPENTAGE

ARPENTAGE implements hierarchical HMM2 as shown in Figure 7. A master

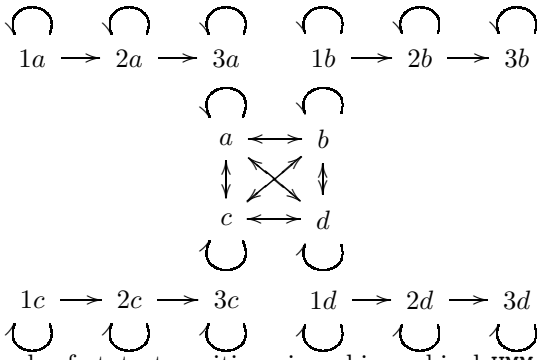


FIGURE 7 – Graph of state transitions in a hierarchical HMM. Each state $x \in \{a, b, c, d\}$ of the HHMM topology is a HMM whose states are $1x, 2x, 3x$. In this figure, the HHMM topology is ergodic (all states are inter connected) whereas the HMM topology is left-to-right and linear

HMM2– whose underlying transition graph is made up of states named a, b, c, d – approximates the MRF. In each master state x, the LU succession probabilities are given by a temporal HMM2 whose underlying graph contains states named $1x, 2x, 3x$. The editing and training of this HHMM2 is performed using the corresponding programs of the CARROTAGE toolbox.

In CARROTAGE, the design of the HMM2 is a crucial step. The user has to specify the underlying graph (linear, ergodic, ...) and to decide which state must be container or Dirac [LBBS+06]. In ARPENTAGE, the user must simply set the number of states (box 3 in Fig. 6) in the ergodic model (a,b,c,d in Fig. 7) related to the number of classes to extract and, in each class, the number of states of a linear model related to the number of steady periods or snapshots (1,2,3 in the same figure).

4.1 *a posteriori* decoding

ARPENTAGE regroups the territory sites into patches (box 4 in Fig. 6) by assigning a class to each site. The assignment is done in three steps :

1. Define a K state ergodic HHMM2 (see Fig. 7) that process the observations along the fractal curve. The observation on a given site is the temporal LU sequence observed on this site. This sequence is built up with single LU observed at time t such as $(lu_t), \quad t = 0, T$ or overlapping temporal n-uplets such as $([lu_t, lu_{t+1}, lu_{t+n-1}]), \quad t = 0, T - n + 1$.
2. Let CARROTAGE train this HHMM2 and compute during the last iteration of the EM algorithm the *a posteriori* class assignment probabilities

$$P\left(z_i = e_k \mid curve\right), \quad i \in S, \quad k = 1, K. \tag{6}$$

3. To take into account the full neighbourhood of each site, we next model the class assignment using a K-colour Potts model with a site-dependent external field whose mean field is the *a posteriori* probabilities computed in step 2 (Eq. 6). Finally, the ICM algorithm [Bes86] performs the class

assignment. Using one iteration, it scans the territory following the fractal curve and gets, for all $i \in S$, a new estimate of $P_i^{mf}(e_s)$ based on Equation 5. The site i is labelled by $\arg\max_k P_i^{mf}(e_k)$ and the mean field at site i is updated to be 1 on this component and 0 on the others.

It seems reasonable to set the external field using Eq. 6 :

$$V_i(e_k) = -\log\left(P(z_i = e_k/curve)\right).$$

The best results have been obtained by setting $\beta = 1$. Then Equation 5 introduces a smooth effect in Eq. 6 that eliminates the effect of the Peano curve in which only 2 neighbours – the previous and the next in the fractal – were taken into account.

5 Case studies

5.1 Data preparation

The corpus of spatial and temporal LU data is generally built either from remotely sensed LU data or from long-term LU surveys. Depending on the data source, several differences in the LU database may exist regarding the number of LU modalities and the representation of the spatial entities : polygons in vector data or pixels in raster data. In the following, the first data source (remotely sensed LU) is illustrated by the Yar watershed case study and the second (long-term LU field surveys) is illustrated by the Seine watershed case study. Principal characteristics of the two case studies are summarized in Table 1.

	Case study	
	Seine watershed	**Yar watershed**
Data source	TER-UTI surveys	Remote sensing
Surface (km^2)	112000	61
Study period	1992 – 2003	1997 – 2008
LU modalities	83 (reduced to 49)	6
Atomic spatial unit	TER-UTI point	polygon
Data base format	Excel data sheet	ESRI Shapefile

TABLE 1 – Comparison between 2 land-use databases coming from two different sources : land-use surveys and remote sensing

5.2 ARPENTAGE on remotely sensed LU data : the Yar watershed

This watershed – 61.5 km^2 – is known as being a place in Brittany where there is an important phytoplanktonic biomass and *Ulva species* mass proliferation risk. Using data obtained by remote sensing analysis and spanning the 1997 – 2008 period, we have distinguished only six LUs : Urban, Water, Forest, Grassland, Cereal and Maize.

On these data, using CARROTAGE, we have performed preliminary temporal segmentation tests with linear models having an increasing number of states

nLig=153157, y1=1997, yn=2008, nAttr=1, indeter=0, isHeader=1													
x	y	pt	poly	97	98	09	00	01	02	...	06	07	08
164603	2424461	1	4825	1	1	1	1	1	1	...	1	1	1
164623	2424461	2	4825	1	1	1	1	1	1	...	1	1	1
164643	2424461	3	4800	3	3	3	3	3	3	...	3	3	3
164663	2424461	4	5005	3	3	3	3	3	3	...	3	3	3
...													

TABLE 2 – First lines of the Yar data sheet. The first line is a header giving the file size (153157 lines), the study period (1997 – 2008), the number of attributes per site each year (=1), the value of the "blank pixel" (=0) and specifies that the next line gives the column's names : x, y coordinates (Lambert conformal conic), pixel Id, polygon Id, and the LU sequence between $year_1$ and $year_n$

on the whole territory. These tests showed that a 6-state linear `HMM2` was the best compromise to achieve an accurate time resolution with a small number of parameters. This defines 6 timestamps. Plotting together the surface size devoted to each LU on these 6 timestamps gives the major trends of the LU dynamics (Figure 8).

The patches shown in figure 8 are associated to a 5-state ergodic `HHMM2`. States 1 and 2, respectively represent Forest and Urban and are steady during the study period. The Urban state is also populated by less frequent LUs that constitute its privileged neighbours. Grassland is the first neighbour of Urban, but it vanishes over the time. The other 3 states exhibit a greater LU diversity and a more pronounced temporal variation. In state 3, Grassland, Maize and Cereal evolve together until the middle of the study period. Next, Grassland and Maize decrease and are replaced by Cereal. This trend may show that a change in the cropping system was undertaken in the patches belonging to this state and threaten the groundwater and surface water quality. State 4 and state 5 represent 2 steady areas populated mainly by Grassland and Forest.

5.3 ARPENTAGE on long-term LU surveys : TER-UTI data on the Seine watershed

The TER-UTI data are collected by the French agriculture administration on the whole French mainland territory. They represent the land use of the country on a one year basis. Two levels of resolution are achieved (see Fig. 9) and determine 2 fractal scans. The aerial photos are first ordered by a Hilbert-Peano scan while the 36 points inside a photo are ordered using a common space filling curve. This defines an extended fractal curve on which the *a posteriori* class assignment probabilities (Eq. 6) are computed. The mean field is defined at the photo level by averaging the mean field probabilities of the 36 points inside a photo. Finally, the ICM algorithm is run on the regular photo lattice and classifies each photo.

The 83 LU have been grouped with the help of agricultural experts into 49 categories following an approach based on the LU frequency in the spatial and temporal database and the similarity of crop management.

On the Seine watershed, represented by 112806 sites (see Tab. 3), ARPENTAGE exhibited patches whose spatial organization looks similar to the mosaic obtained by [MSB04, MSB07] on the same data in their work for modelling the spatial dynamics of farming practices in the Seine watershed and for understanding the relations in diffuse pollution observed in the ground waters and surface waters of the river Seine.

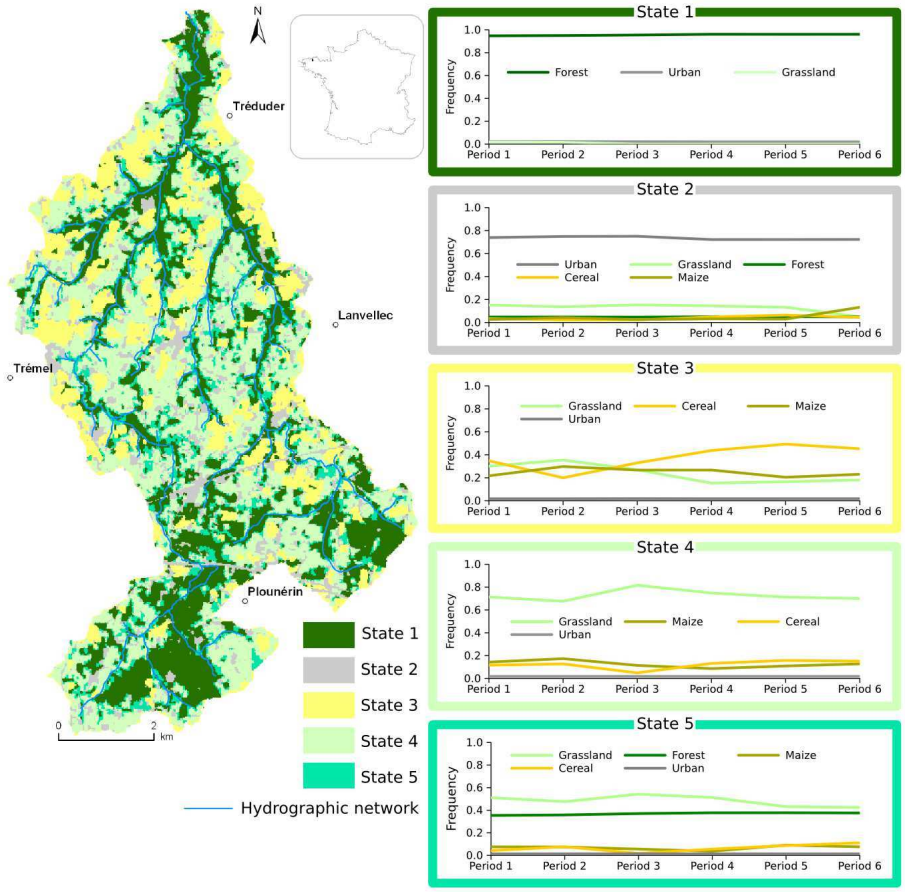
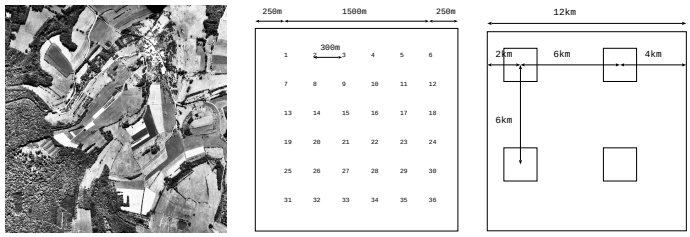


FIGURE 8 – The Yar watershed seen as patches of LU dynamics. Each map unit stands for a state of the `HHMM2` used to achieve the spatial segmentation. Each state is described by a diagram of the LU evolution. Location in France of the Yar watershed is shown by a black spot depicted in the upper middle box

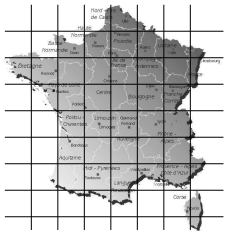
nLig=112806, y1=1992, yn=2003, nAttr=1, indeter=95, isHeader=1

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 3 – First lines of the Seine data sheet. The (x,y) coordinates have been replaced by the photo Id, the intra grid point Id (pti : 1 – 36). Each site is labelled by the administrative department (dep) and the agricultural district (pra = smaller agricultural region) where it is located



(a) an aerial photograph (b) the 6x6 grid inside a photo (c) the 4 aerial photos in a mesh



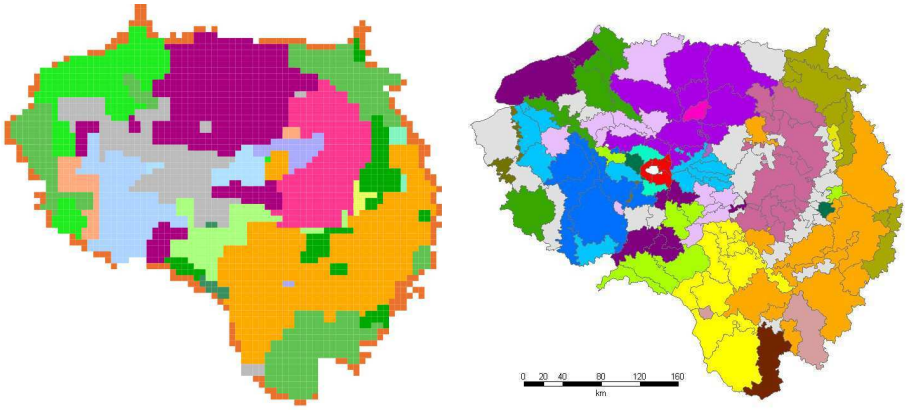
(d) the basic grid covering France

FIGURE 9 – Collecting the TER-UTI data : 3820 meshes square France, 4 aerial photographs are sampled in a mesh, a 6x6 grid determines 36 sites.

In this work, 147 districts were first labelled by their main crop successions using CARROTAGE. As the cropping system was assumed as being stationary over the whole study period, a one state `HMM2` was specified. The observations were temporal triplets of LU. Their distribution defined a cropping plant that was computed on each agricultural district. A linear component analysis (LCA) followed by a hierarchical classification (HC) using Ward's method identified homogeneous regions made up of groups of contiguous agricultural districts which exhibited similar combinations of crop successions (see Fig. 10(b)). It is interesting to note that ARPENTAGE produced roughly a similar mosaic without having to use the geographical limits of the agricultural districts. In both experiments, the observations were temporal triplets of LU, the number of states in the master `HHMM2` was set to the same number of classes found by the HC and the number of steady periods was set to 1 like in Mignolet's work in order to have a fair experiment.

6 Comparison with other similar software programmes

ARPENTAGE provides a stand-alone analysis tool to extract patches based on their pluri-annual LU organization, it can be seen as a GIS analysis tool. Since the initial point of [Lan93] saying that GIS were poorly equipped to handle temporal data, many researchers have sought to integrate the time dimension into GIS [RHS01]. It is now well accepted in many fields such as pedagogy [Pia73], anthropology [Hal90], GIS [Peu02] and agronomy [LMB09, SLM+12] that the temporal and spatial dimensions are interrelated and cannot be exchanged. This



(a) ARPEnTAge (b) LCA+HC

FIGURE 10 – Comparison between two clusterings of the Seine watershed. LCA+HC represents the map obtained by statistical methods [MSB07]. ARPEnTAge gives results directly from Ter-Uti data without considering the district borders. The patch's colours are characteristic of the LU succession distributions that are roughly the same in both maps

explains why a 3-D modelling approach provides a limited answer. Following Langran's and Peuquet's works, a GIS that handles time data can fall into three categories [Peu00, Wac00].

Space-Dominant Models : space is considered as a container in which events occur. A new snapshot is created every time a new event occurs. Time is frozen in each layer.

Time-Dominant Models : specific patterns that occur repeatedly or in sequence constitute units that are geographically located.

Relative Space-Time Models : the relations between entities determine their locations.

Several software programmes that implement Space-Dominant Models and having clustering capabilities, have been released in various domains :

- In the image segmentation area, SpaceEM3 [1] is used for clustering various data from hyper spectral satellite images, remote sensing and mapping epidemics of ecological species.
- In the space-time disease surveillance domain, ClusterSeer [2], SatScan [3], GeoSurveillance [4] and the Surveillance package for R [5] provide maps from disease data. More generally, the GNU R statistical tool provides access to Geoprocessing tools [6] (ArcGIS, QGIS, ...). R programmers can read shapefile, do unsupervised clustering on the spatial entities based on their attributes and represent the results as shapefiles. But, the time dimension of the attributes is not handled.

1. http ://spacem3.gforge.inria.fr
2. http ://www.terraseer.com
3. http ://www.satscan.com
4. http ://www.acsu.buffalo.edu/˜ rogerson/geosurv.htm
5. http ://cran.r-project.org/web/packages/surveillance/
6. http ://cran.r-project.org/web/views/Spatial.html

In our knowledge, no software provides a simultaneous analysis of time sequences and their spatial locations. Consequently, ARPENTAGE can be seen as the first software in the agronomic area implementing a Time-Dominant Model and processing time-space data.

7 Discussion and conclusions

We have presented a software programme called ARPENTAGE whose goal is to achieve an unsupervised clustering of a 2-D territory represented by its LU successions. This software is based on a stochastic modelling of the time space stream of data. The user controls the clustering through a limited set of parameters : the length of the elementary observation (1 LU for the Yar case study and 3 successive LUs in the Seine watershed case study), the number of states in the master `HHMM2` that specifies the number of clusters to be extracted and the number of states of the temporal `HMM2` that define the number of desired steady periods.

In the mean field paradigm applied to the Potts model, we have shown that the initialization of the mean field by the *a posteriori* probabilities given by a fractal scan provides a tractable opportunity to obtain patchy landscapes. These probabilities can be used to define an external field as well. But so far, the value β that controls the pixel interaction strength has not been not learnt and set by the expert. A logical continuation of this work would be to consider its learning.

ARPENTAGE rapidly produces patchy landscapes of various sizes whose classes can be analyzed more precisely by agronomists. As shown in the Yar case study, ARPENTAGE implements a Time-Dominant model and proposes a visualization of changes – *eg* where the grasslands are replaced by crops – by means of shapefiles and Markov diagrams that CARROTAGE can display. In the Seine case study, ARPENTAGE produced a clustering of the watershed based on 3 year successions and computed a shapefile that can be viewed as a snapshot showing clusters having stationary successions over the study period. In this case, the `HHMM2` acts as a Space-Dominant model in which the dominant successions are the themes to be located.

In a stochastic framework, a plot mosaic description is obtained by estimating as many probabilistic distributions as clusters that a clustering program can extract, each of them characterizing the content of a cluster. Only few works tackle the issue of describing the neighbour effects between clusters and their time dynamics. ARPENTAGE showed interesting capabilities in quantifying the neighbourhood effects between clusters. [LMB09], in their work to describe a patchy landscape having environmental issues, used CARROTAGE to determine the main LU successions and ARPENTAGE to locate them inside the territory. As they observed that LU successions were stationary over the 1996-2007 period, they used a simple temporal HMM2 to represent the states of the hierarchical HMM2 (see Fig. 11). This model had 2 states. One – $S(X)$ – described the distribution of the temporal quadruplets of interest related to the succession $S(X)$ involving the LU X, the other state – $N(X)$ – captured the distribution of the temporal quadruplets in the neighbourhood. The Markov field introduces a blur in the patch's frontier and in the patch estimation because a site is classified not only based on its temporal characteristics (the quadruplet succession) but depends now on the classification of the neighbouring sites. A

patch was then described by two stochastic pluri-annual LU distributions : one characterizing the inside and the second characterizing the border. The latter influenced their relative locations as in a relative time space model. That last point shows that ARPENTAGE brings a valuable help for creating shapefiles from time-space data in temporal GIS.

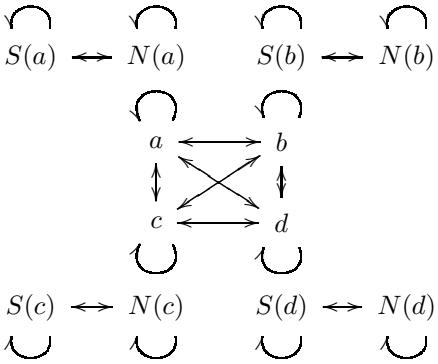


FIGURE 11 – Graph of state transitions in a `HHMM` that describe 4 kinds of patches based on the inside – S(X) – and border – N(X) – observation distributions

8 Acknowledgments

Many organizations had provided us with support and data. This work was mainly supported by the ministery of *Education nationale*, the *région Lorraine*, the API - ECOGER, the *Zone Atelier PIREN-Seine*, the ANR ADD-COPT, ANR BiodivAgrim and ANR PopSy projects. We thank the CNRS team UMR COSTEL (Rennes) for the "Yar database" and the anonymous reviewers who help us to improve the initial manuscript. This paper has been published (2013) in Environmental Modelling and Software. The original publication is available at www.elsevier.com under DOI 10.1016/j.envsoft.2013.03.014.

Références

[BCG+13] J.-E. Bergez, P. Chabrier, C. Gary, M.H. Jeuffroy, D. Makowski, G. Quesnel, E. Ramat, H. Raynal, N. Rousse, D. Wallach, P. Debaeke, P. Durand, M. Duru, J. Dury, P. Faverdin, C. Gascuel-Odoux, and F. Garcia. An open platform to build, evaluate and simulate integrated models of farming and agro-ecosystems. *Environmental Modelling & Software*, 39(0) :39 – 49, 2013.

[Bes86] Julian Besag. On the Statistical Analysis of Dirty Picture. *Journal of the Royal Statistical Society*, B(48) :259 – 302, 1986.

[BP95] 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.

[BRM+12] Marc Benoît, Davide Rizzo, Elisa Marraccini, Anna Camilla Moonen, Mariassunta Galli, Sylvie Lardon, HÃľlÃĺne Rapey, Claudine

Thenail, and Enrico Bonari. Landscape agronomy : a new field for addressing agricultural landscape dynamics. *Landscape Ecology*, 27(10) :1385 – 1394, 2012.

[CFP03] Gilles Celeux, Florence Forbes, and Nathalie Peyrard. EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recognition*, 36(1) :131–144, 2003.

[DCOM00] Revital Dafner, Daniel Cohen-Or, and Yossi Matias. Context-based Space Filling Curves. *Computer Graphics Forum*, 19(3) :209–218, 2000.

[DLR77] 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.

[EAA+09] Catherine Eng, Charu Asthana, Bertrand Aigle, Sébastien Hergalant, Jean-Francois Mari, and Pierre Leblond. A new data mining approach for the detection of bacterial promoters combining stochastic and combinatorial methods. *Journal of Computational Biology*, 16(9) :1211–1225, Sept. 2009. http ://hal.inria.fr/inria-00419969/en/.

[EdP10] H.A. Engelbrecht and J.A. du Preez. Efficient backward decoding of high-order hidden markov models. *Pattern Recognition*, 43(1) :99 – 112, 2010.

[ETD+11] Catherine Eng, Annabelle Thibessard, Morten Danielsen, Thomas Bovbjerg Rasmussen, Jean-Francois Mari, and Pierre Leblond. In silico prediction of horizontal gene transfer in Streptococcus thermophilus. *Archives of Microbiology*, 193(4) :287–297, January 2011.

[For95] Richard T.T. Forman. Some general principles of landscape and regional ecology. *Landscape Ecology*, 10(3) :133–142, 1995.

[FST98] Shai Fine, Yoram Singer, and Naftali Tishby. The Hierarchical Hidden Markov Model : Analysis and Applications. *Machine Learning*, 32 :41 – 62, 1998.

[GG84] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6, 1984.

[GP97] Nathalie Giordana and Wojciech Pieczynski. Estimation of Generalized Multisensor Hidden Markov Chains and Unsupervised Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(5) :465 – 475, May 1997.

[Hal90] Edward T. Hall. *The hidden dimension*. Anchor Books, 1990.

[HK08] Ruihong Huang and Christina Kennedy. Uncovering hidden spatial patterns by hidden markov model. In *Proceedings of the 5th international conference on Geographic Information Science*, GIScience '08, pages 70–89, Berlin, Heidelberg, 2008. Springer-Verlag.

[JSMP06] A. Joannon, V. Souchère, P. Martin, and F. Papy. Reducing runoff by managing crop location at the catchment level : considering agronomic constraints at farm level. *Land Degradation and Development*, 17(5) :467–478, 2006.

[LABP12] Thierry Leviandier, A. Alber, F. Le Ber, and H. Piégay. Comparison of statistical algorithms for detecting homogeneous river reaches along a longitudinal continuum. *Geomorphology*, 138(1) :130 – 144, 2012.

[Lan93] Gail Langran. *Time in Geographic Information Systems*. Taylor and Francis, 1993. ISBN : 0-7484-0003-6.

[LBBS+06] 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.

[LMB09] E.G. Lazrak, J.-F. Mari, and M. Benoît. Landscape regularity modelling for environmental challenges in agriculture. *Landscape Ecology*, 25(2) :169 – 183, Sept. 2009.

[MHK97] 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.

[MLB06] J.-F. Mari and F. Le Ber. Temporal and Spatial Data Mining with Second-Order Hidden Markov Models. *Soft Computing*, 10(5) :406 – 414, March 2006.

[MSB04] 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(4) :219 – 235, 2004.

[MSB07] C. Mignolet, C. Schott, and M. Benoît. Spatial dynamics of farming practices in the Seine basin : Methods for agronomic approaches on a regional scale. *Science of The Total Environment*, 375(1–3) :13–32, April 2007.

[PBW98] Johan A. Du Preez, Promoters Dr. E. Barnard, and Dr. D. M. Weber. Efficient high-order hidden markov modelling. In *in Proceedings of the International Conference on Spoken Language Processing*, pages 2911–2914, 1998.

[Peu00] Donna J. Peuquet. *Time in GIS ;Issues in spatio-temporal modelling*, chapter Space-time representation :An overview, pages 3 – 12. NCG Nederlandse Commissie voor Geodesie, 2000.

[Peu02] Donna J. Peuquet. *Representations of Space and Time*. The Guilford Press, 2002.

[Pia73] Jean Piaget. *Développement de la notion de temps chez l'enfant*. Presses universitaires de France, 1973.

[RHS01] J. F. Roddick, K. Hornsby, and M. Spiliopoulou. Yabtsstdmr - yet another bibliography of temporal, spatial and spatio-temporal data mining research. In K.P. Unnikrishnan and R. Uthurusamy, editors, *SIGKDD Temporal Data Mining Workshop*, pages 167–175, San Francisco, CA, 2001. ACM.

[Ska92] W. Skarbek. Generalized Hilbert Scan in Image Printing. *Theoretical Foundation of Computer Vision*, pages 45 – 57,, 1992. R. Klette and W.G. Kropetsh, eds., Akademik Verlag.

[SLM+12] Noémie Schaller, El Lazrak, Philippe Martin, Jean-François Mari, Christine Aubry, and Marc Benoît. Combining farmers' decision rules and landscape stochastic regularities for landscape modelling. *Landscape Ecology*, 27 :433–446, 2012.

[SLOW11] Adrian Southern, Andrew Lovett, Tim O'Riordan, and Andrew Watkinson. Sustainable landscape governance : Lessons from a catchment based study in whole landscape design. *Landscape and Urban Planning*, 101(2) :179 – 189, 2011.

[SMDF+12] Jordy Salmon-Monviola, Patrick Durand, Fabien Ferchaud, François Oehler, and Luc Sorel. Modelling spatial dynamics of cropping systems to assess agricultural practices at the catchment scale. *Computers and Electronics in Agriculture*, 81(0) :1 – 13, 2012.

[Wac00] Monica Wachowicz. *Time in GIS ;Issues in spatio-temporal modelling*, chapter The role of geographic visualisation and knowledge discovery in spatio-temporal data modelling, pages 13 – 26. NCG Nederlandse Commissie voor Geodesie, 2000.