Using Markov Models to Mine Temporal and Spatial Data

Jean-Francois Mari, Florence Le Ber, El-Ghali Lazrak, Marc Benoît, Catherine Eng, Annabelle Thibessard, Pierre Leblond

0

Using Markov Models to Mine Temporal and Spatial Data

Jean-François Mari[1], Florence Le Ber[1,2], El Ghali Lazrak[3], Marc Benoît[3], Catherine Eng[4], Annabelle Thibessard[4] and Pierre Leblond[4]
[1]*LORIA / Inria-Grand Est, Campus scientifique, BP 239, F-54500, Vandœuvre-lès-Nancy*
[2] *ENGEES,1 Quai Koch, F-67000, Strasbourg*
[3] *INRA, UR 055, SAD-ASTER, domaine du Joly, F-88500, Mirecourt*
[4] *Laboratoire de Génétique et de Microbiologie, UHP-INRA, UMR 1128-IFR110, F-54500, Vandœuvre-lès-Nancy*
France

1. Stochastic modelling, temporal and spatial data and graphical models

Markov models represent a powerful way to approach the problem of mining time and spatial signals whose variability is not yet fully understood. Initially developed for pattern matching (Baker, 1974; Geman & Geman, 1984) and information theory (Forney, 1973), they have shown good modelling capabilities in various problems occurring in different areas like Biosciences (Churchill, 1989), Ecology (Li et al., 2001; Mari & Le Ber, 2006; Le Ber et al., 2006), Image (Pieczynski, 2003; Forbes & Pieczynski, 2009) and Signal processing (Rabiner & Juang, 1995). These stochastic models assume that the signals under investigation have a local property –called the Markov property– which states that the signal evolution at a given instant or around a given location is uniquely determined by its neighbouring values. In 1988, Pearl (Pearl, 1988) shown that these models can be viewed as specific dynamic Bayesian models which belong to a more general class called graphical models (Whittaker, 1990; Charniak, 1991).

The graphical models (GM) are the results of the marriage between the theory of probabilities and the theory of graphs. They represent the phenomena under study within graphs where the nodes are some variables that take their values in a discrete or continuous domain. Conditional –or causal– dependencies between the variables are graphically expressed. As an example, the relation between the random variables U, V and W depicted by the figure 1 expresses that V and W are the reasons –more or less probable– of U. In a Bayesian attitude, the uncertainty about this relation is measured by the conditional probability $P(U/V,W)$ of observing U given V and W.

In graphical models, (see Fig. 2, 3 and 4), some nodes model the phenomenon's data thanks to adequate distributions of the observations. They are called "observable" variables whereas the others are called "hidden" variables. The observable nodes of the graph give a frozen view of the phenomenon. In the time domain, the temporal changes are modelled by the set of transitions between the nodes. In the space domain, the theory of graphs allows to take into account the neighbourhood relations between the phenomenon's constituents.

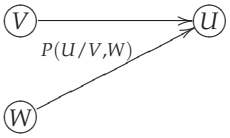


Fig. 1. Conditional dependency of U with V and W in a Bayesian network. The probability measures the confidence of the dependency

The mining of temporal and / or spatial signals by graphical models can have several purposes:

Segmentation : in this task, the `GM` clusters the signal into stationary (or homogeneous) and transient segments or areas (Jain et al., 1999). The term stationnary means that the signal values are considered as independent outcomes of probability density functions (`pdf`). These areas are then post-processed to extract some valuable knowledge from the data.

Pattern matching : in this task, the `GM` measures the *a posteriori* probability $P(model = someLabel/observedData)$. When there are as many `GM` as labels, the best probability allows the classification of an unknown pattern by the label associated with the highest probability.

Background modelling : in order to make proper use of quantitative data, the `GM` is used as a background model to simulate an averaged process behavior that corrects for chance variation in the frequency counts (Huang et al., 2004). The domain expert compares the simulated and real data frequencies in order to distinguish if he / she is facing to over- or under-represented data that must be investigated more carefully.

In this chapter, we will present a general methodology to mine different kinds of temporal and spatial signals having contrasting properties: continuous or discrete with few or many modalities. This methodology is based on a high order Markov modelling as implemented in a free software: CARROTAGE (see section 3). Section 2 gives the theoretical basis of the modelling. Section 3 describes a general flowchart for mining temporal and spatial signals using CARROTAGE. The next section is devoted to the description of three data mining applications following the same flowchart. Finally, we draw some conclusions in section 5.

2. The HMM as a graphical model

The Hidden Markov Model is a graphical model which represents the sequence of observations as a doubly stochastic process: an underlying "hidden" process, called the state sequence of random variables $Q_1, Q_2, ...Q_T$ and an output (observation) process, represented by the sequence $O_1, O_2, ...O_T$ over the same time interval (see Fig. 2 – 3). The sequence (Q_t) is a Markov chain and represents the different clusters that must be extracted.

2.1 `HMM` definition

We define a hidden Markov model by giving:

- $\mathbf{S} = \{s_1, s_2, \ldots, s_N\}$, a finite set of N states ;
- $\mathbf{A}$ a matrix defining the transition probabilities between the states:

  $\mathbf{A} = \left(a_{ij}\right)$ for a first order HMM (`HMM1`) (Fig. 2),

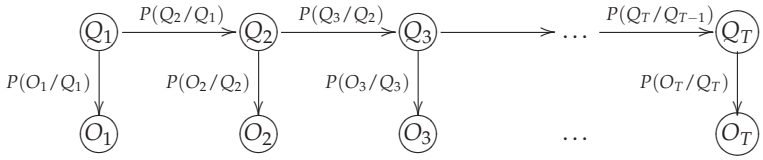


Fig. 2. Conditional dependencies in a `HMM1` represented as a Bayesian network. The hidden variables (Q_t) govern the observable variables (O_t)

$\mathbf{A} = \left(a_{ijk}\right)$ for a second order HMM (`HMM2`) (Fig. 3);

- $\mathbf{b_i}(.)$ the distributions of observations associated to the states s_i. This distribution may be parametric, non parametric or even given by an HMM in the case of hierarchical HMM (Fine et al., 1998).

As opposite to a Markov chain where the states are unambiguously observed, in a HMM, the observations are not uniquely associated to a state s_i but are drawn from a random variable that has a conditional density $\mathbf{b_i}(.)$ that depends on the actual state s_i (Baker, 1974). There is a doubly stochastic process:

- the former is hidden from the observer, is defined on a set of states and is a Markov chain;
- the latter is visible. It produces an observation at each time slot –or index in the sequence– 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.

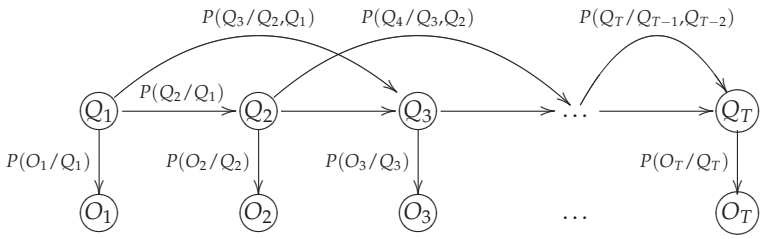


Fig. 3. Conditional dependencies in a `HMM2` represented as a Bayesian network. The hidden variables (Q_t) govern the observable variables (O_t)

2.2 Modelling the dependencies in the observable process

Defining the observation symbols is the first step of a HMM data processing. In this chapter, we will present our data mining work based on various `GM` applied on different kinds of signals having contrasting properties:

- genomic data characterized by long sequences (several millions) of the 4 nucleotides A, C, G, T (application 1);

- short temporal discrete sequences (around 10 value long) with a great number (around 50) of modalities like the temporal land use successions (LUS) of agricultural fields whose mosaic defines a 2-D spatial territory (application 2);
- continuous data like the values of a river width sampled from the river's source up to its end (application 3).

To take into account the correlations between successive or neighbouring observations, several options are possible.

2.2.1 Continuous observations

The usual way to model continuous random observations is to consider them as Gaussian distributed. When the observations are vectors belonging to $\mathbb{R}^d$, multivariate Gaussian `pdf` are used. The main reason of this consideration is that an unknown `pdf` can be approximated by a mixture of multivariate Gaussian `pdf`. To take into account the correlations between successive observations, first and second order regression coefficients (Furui, 1986) are stacked over the observation vector:

$$R(t) = \frac{\sum_{n=-n_0}^{n_0} nO(t+n)}{\sum_{n=-n_0}^{n_0} n^2} \tag{1}$$

where $O(t+n)$ is the observation (frame) $t+n$. The $2n_0+1$ frames involved in the computation of the regression coefficient $R(t)$ are centered around frame t. By this way, the vector at time t models the shape of the observation variations and incorporates information about the surrounding context.

2.2.2 Categorical observations

When the observations are discrete and belong to a finite set $C = \{c_1, c_2, ...c_M\}$, it is convenient to represent this correlation by adding new dependencies between the current observation and the previous observations. In the particular case shown in Fig. 4, the observation distribution is a conditional `pdf` $\mathbf{b_{iuv}}(o_t)$ that represents the conditional probability of observing o_t assuming the state s_i and the observations u and v that occured respectively at indices $t-1$ and $t-2$:

$$o_{t-1} = u,\ o_{t-2} = v \quad u, v \in C.$$

In the temporal domain, this leads to the definition of a `Mp-Mq HMM` where p is the order of the hidden Markov process and q refers to the dependencies in the observable process.

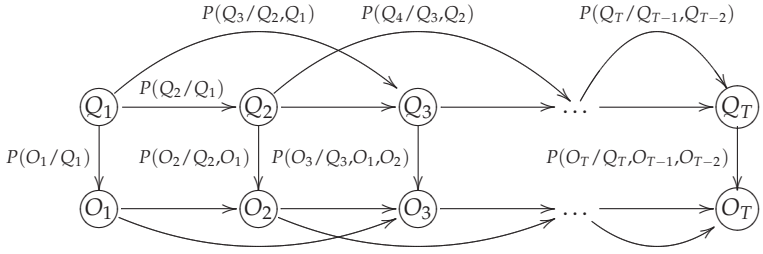


Fig. 4. Conditional dependencies of a `M2-M2 HMM` represented in a Bayesian network

Another way to take into account the correlation between successive (neighbouring) observations, is to consider composite observations drawn from the n-fold product $C^n = C \times C \ldots C$.

The elementary observation (for example, a nucleotide, a land use ...) is considered together with its context. This leads to the definition of *k-mer* (see section 4.1.1) in biology or land use succession in agronomy (see section 4.2.1.3). As a direct consequence, the `pdf` size will be changed from $|C|$ to $|C|^n$ where $|C|$ denotes the cardinality of C. It is then possible to control the balance between the parameter number assigned to the hidden variables and to the observable ones in the model.

2.3 Automatic estimation of a HMM2

The estimation of an `HMM1` is usually done by the forward backward algorithm which is related to the EM algorithm (Dempster et al., 1977). We have shown in (Mari et al., 1997) that an `HMM2` can be estimated following the same way. The estimation is an iterative process starting with an initial model and a corpus of sequences of observations that the `HMM2` must fit even when the insertions, deletions and substitutions of observations occur in the sequences. If `N` is the number of states and `T` the sequence length, the second-order forward backward algorithm has a $\mathtt{N}^3 \times \mathtt{T}$ complexity for an `HMM2`.

The very success of the HMM is based on their robustness: even when the considered data do not suit a given HMM, its use can give interesting results. The initial model has equi-probable transition probabilities and a uniform distribution in each state. At each step, the forward backward algorithm determines a new model in which the likelihood of the sequences of observation increases. Hence this estimation process converges to a local maximum. Interested readers may refer to (Dempster et al., 1977; Mari & Schott, 2001) to find more specific details of the implementation of this algorithm.

The choice of the initial model has an influence on the final model obtained by convergence. To assess this last model, we use the Kullback-Leibler distance between the distributions associated to the states (Tou & Gonzales, 1974). Two states that are too close are merged and the resulting model is re-trained. Domain experts do not interfere in the process of designing a specific model, but they have a central role in the interpretation of the results that the final model gives on the data.

3. CARROTAGE a general framework to mine sequences

We have developed a knowledge discovery system based on high-order hidden Markov models for analyzing temporal data bases (Fig. 5). This system, named CARROTAGE[1], takes as input an array of discrete or continuous data –the rows represent the individuals and the columns the time slots– and builds a partition together with its *a posteriori* probability. CARROTAGE is a free software [2] under a Gnu Public License. It is written in C++ and runs under Unix systems. In all applications, the data mining processing based on CARROTAGE is decomposed into four main steps:

Model specification. Even if CARROTAGE may use models of any topology, we mainly use two different graph topologies: linear and ergodic. In a linear model, there is no circuit between the nodes except self loops on some nodes. Whereas in an ergodic model, all the nodes are inter connected; a node can reach all the others. The first `HMM2` that

[1] CARROTAGE is a retro acronym that comes from the word carrot that can be translated by Markov in Russian and age to refer to the temporal component of the data. It is also a technique which consists in drilling a hole in some material (a tree or the ice of the Antarctic) to withdraw a cylinder that allows to date the process of creation

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

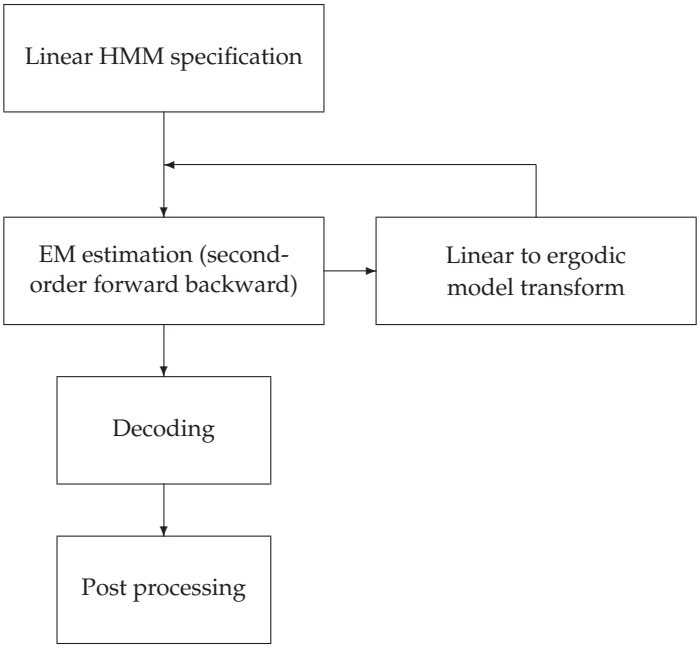


Fig. 5. General flow Chart of the data mining process using CARROTAGE

CARROTAGE has to estimate is linear with equi-probable transitions from each state and uniform distributions of observations in every states. The only parameter let to the user is the number of states.

Iterative estimate of the model parameters. The parameter estimation of the model is performed by the forward backward algorithm for `M2-Md HMM`. Basically, given a sequence of symbols $\left(o_1^T\right) = o_1, o_2, \ldots o_T$ the second-order forward backward algorithm computes the expected count of the state transition $s_{i_1} \rightarrow s_{i_2} \rightarrow s_{i_3}$

$$\eta_t(i_1, i_2, i_3) = P(Q_{t-2} = s_{i_1}, Q_{t-1} = s_{i_2}, Q_t = s_{i_3} / O_1^T = o_1^T) \tag{2}$$

at index $t-2, t-1, t$.

The first parameter estimate is performed on a linear model to acquire a segmentation of the sequence into as many homogeneous regions than there are states in the specified model.

Linear to ergodic model transform. The estimated linear model is transformed into an ergodic one by keeping the previously estimated `pdf` and interconnecting the states. This allows the stochastic process to re-visit the states and, therefore, segment the data into an unconstrained number of homogeneous regions, each of them associated to a state.

Decoding. The decoding state uses the last iteration of EM algorithm to calculate the *a posteriori* probability of the hidden states. It is possible to compute three types of *a posteriori* probability. In all the following definitions, we assume that the hidden state s_i is attained at time t and that we have a T length observation sequence $\left(o_1^T\right)$.

type 0

$$P_0(i,t) = \sum_{i_1,i_2} \eta_t(i_1,i_2,i) \tag{3}$$

The *a posteriori* probability of the state s_i at index t assuming the whole sequence $\left(o_1^T\right)$.

type 1

$$P_1(i,t) = \sum_{i_1} \eta_t(i_1,i,i) \tag{4}$$

The *a posteriori* probability of the 2 state transition $s_i \to s_i$ at index t assuming the whole sequence $\left(o_1^T\right)$. This probability can be computed either by a `HMM1` or by a `HMM2`.

type 2

$$P_2(i,t) = \eta_t(i,i,i) \tag{5}$$

The *a posteriori* probability of the 3 state transition $s_i \to s_i \to s_i$ at index t assuming the whole sequence $\left(o_1^T\right)$. This probability is typical of a `HMM2`.

In some applications, as the mining of crop successions (see section 4.2), the *a posteriori* transition probability (type 1) between 2 states can be used and gives an interesting information. In such a case, we use:

$$P_1(i,j,t) = \sum_{i_1} \eta_t(i_1,i,j) \tag{6}$$

Post processing: The post processing is application dependent and involves mostly a classification step of the different segments. Further ad-hoc treatments must be performed in order to extract valuable information as shown in the application section.

4. Applications

4.1 Mining genomic data

In this section, we describe a new data mining method based on second-order HMM and combinatorial methods for Sigma Factor Binding Site (SFBS) prediction (Eng et al., 2009) and Horizontal Gene Transfer (HGT) (Eng et al., 2011) detection that voluntarily implements a minimum amount of knowledge. The original features of the presented methodology include (i) the use of the CARROTAGE framework, (ii) an automatic area extraction algorithm that captures atypical DNA motifs of various size based on the variation of the state *a posteriori* probability, and (iii) a set of post processing algorithms suitable to the biologic interpretation of these segments. On some points, our data mining method is similar to the work of Bize et al. (Bize et al., 1999) and Nicolas et al. (Nicolas et al., 2002). All the methods use one HMM to model the entire genome. The parameter estimation is done in all cases by the EM algorithm. All the methods look for attributing biological characteristics to the states by analyzing the state output *a posteriori* probability. But our method differs on the following points: we use (i) an HMM2 that has proved interesting capabilities in modelling short sequences, and (ii) depending on the modelled dependencies in the genomic sequence, we can locate either short nucleotides sequences that could be part of SFBS (box1 or box2) or more generally regulation sites for gene expression –Transcriptional Factor Binding sites (TFBS)– or even wider areas potentially acquired by HGT. These sequences are post processed to assess the exact nature of the heterogeneities (SFBS, TFBS or HGT).

4.1.1 Data preparation

In this application, the genome is modelled as an ordered nucleotide sequence whose unknown structure is represented by the state Markov chain. The index t in equation (2) refers to the nucleotide index in the ordered sequence of nucleotides. In a genome sequence, two templates must be considered depending upon the strength of the compositional biases. To incorporate the biased base composition of DNA strands relative to the position of the replication origin when a marked GC skew[3] is observed, as in the case of *Streptococcus thermophilus*, a sequence is constructed *in silico* by concatenating the two leading strands from the origin to the terminus of replication. Its reverse complement is also considered. In contrast, when the genome does not show a marked GC skew, as in *Streptomyces coelicolor*, the 5′ to 3′ sequence of the linear chromosome and its reverse complement are considered. In both cases, these two sequences are used for training purposes and specify two HMM2 named HMM2+ and HMM2-. The best decoding state is identified for both models.

We have also investigated the use of *k-mer* (Delcher et al., 1999) as output symbols instead of nucleotides. A *k-mer* may be viewed as a single nucleotide y_t observed at index t with a specific context $y_{t-k+1},\dots y_{t-1}$ made of $k-1$ nucleotides that have been observed at index $t-k+1,\dots,t-1$. Similarly, a DNA sequence can be viewed as a sequence of overlapping *k-mer* that an HMM analyzes with a consecutive shift of one nucleotide. For example, the seven nucleotide sequence `TAGGCTA` can be viewed as a sequence of seven *3-mer*: `##T - #TA - TAG - AGG - GGC - GCT - CTA`, where # represents an empty context.

4.1.2 *a posteriori* decoding

The mining of irregularities follows the general flow chart given in figure 5. The *a posteriori* probability variations look very different depending on the dependencies that are implemented in the genomic sequence. When modelling the *k-mer* sequence using a `M2-M0 HMM`, the decoding stage locates atypical short DNA segments (see Fig. 6) whereas the modelling of the nucleotide sequence using a `M2-M2 HMM` exhibits wider atypical areas (see Fig.7).

4.1.3 Post processing

The atypical regions extracted by the stochastic models must be processed in order to extract valuable information. A specific suite of algorithms has been designed and tuned in the two applications: TFBS and HGT detections.

4.1.3.1 TFBS retrieval

Our bacterial model is the Gram-positive actinomycete *Streptomyces coelicolor* whose genome is 8.7 Mb long. The streptomycetes are filamentous bacteria that undergo complex morphological and biochemical differentiation, both processes being inextricably interlinked. The purpose of the TFBS application is to retrieve composite motifs *box1-spacer-box2* involved in the *Streptomyces coelicolor* regulation. The two boxes can be part of the intergenic peak motifs (see Fig. 6). The spacer ranges from 3 to 25 and is tuned depending on the type of the investigated TFBS. The basic idea of the mining strategy is to cluster the set of intergenic ipeak motifs located by a `M2-M0 HMM` modelling *3-mer*, select a cluster having a well defined consensus, extend all the sequences belonging to this cluster and look for over-represented motifs by appropriate software (Hoebeke & Schbath, 2006). The consensus of the cluster acts for *box1*, the shorter motifs spaced with appropriate spacer value(s) act for *box2*. Interested readers

[3] the GC skew is a quantitative feature that measures the relative nucleotide proportion of G *versus* C in the DNA strand

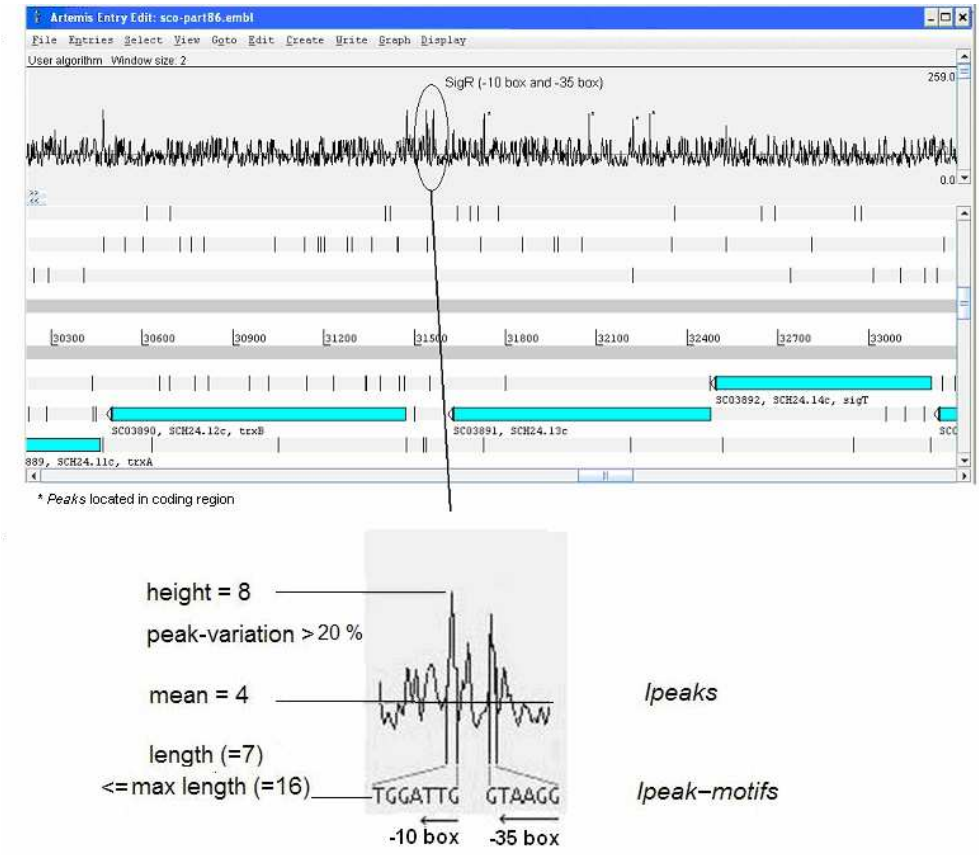


Fig. 6. *A posteriori* probability variation of a `M2-M0 HMM` hidden state as a function of the *3-mer* index in the *Streptomyces coelicolor* genome. The top graph shows the *a posteriori* probability together with the annotated physical sequence (using the EMBL file). As an example, among the intergenic peak motifs, the -35 box (GGAAT) and -10 box (GTT) motifs recognized by the sigma factor SigR are detected. Peak characteristics (peak-variation and length) are marked in the figure. The biological interpretation of the peaks inside the coding regions is not yet fully established (Eng et al., 2009)

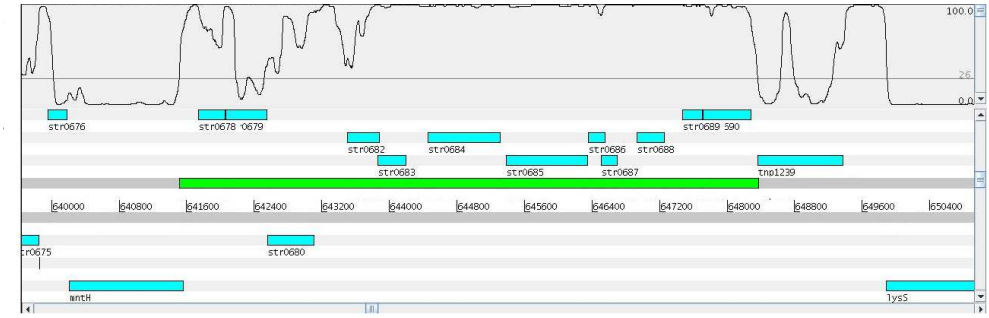


Fig. 7. *A posteriori* probability variation of a `M2-M2 HMM` hidden state as a function of the nucleotide index in the *Streptococcus thermophilus* genome. The additional dependencies in the nucleotide sequence dramatically smooth the state *a posteriori* probability

will find in (Eng et al., 2009) an extensive description of this data mining strategy based on stochastic and combinatorial methods.

4.1.3.2 Horizontal Gene Transfer detection

Our bacterial model is the Gram-positive bacteria *Streptococcus thermophilus* which is a lactic acid bacteria carrying a 1.8 Mb genome and having a considerable economic importance. It is used as starter for the manufacturing of yogurts and cheeses. *Streptococcus thermophilus* is assumed to have derived very recently at the evolutionary time-scale (3,000-30,000 years back: the beginning of the pastoral epoch) from a commensal ancestor which is closely related to the contemporary oral bacterium *Streptococcus salivarius* to adapt to its only known ecological niche: the milk. HGT deeply shaped the genome and played a major role in adaptation to its new ecological niche.

In this application, we have observed that the `M2-M2 HMM` modelling nucleotides performs better than `M1-M2 HMM` as implemented in SHOW software [4] (Nicolas et al., 2002) and `M2-M0 HMM` modelling *3-mer* (see section 4.1.3.1).

After tuning the HMM topology, the decoding state that captures the highest heterogeneities is selected by considering the distances between all states according to the Kullback-Leibler distance. The state which is the most far away from the others is selected. On this state, the variations of the *a posteriori* probability as a function of the index in the nucleotide sequence are analyzed. The positions having *a posteriori* probabilities higher than the mean over the whole genome are considered. Regions enriched in these positions through at least 1000 nucleotide length were extracted and named atypical regions. A total of 146 atypical regions were extracted. If a gene were at least half included in these regions then it was considered. A total of 362 genes of 1915 (the whole gene set of the bacterium), called "atypical", were retrieved from these regions. Based on their functional annotation and their sporadic distribution either at the interspecific level (among the other genomes belonging to the same phylum: the Firmicutes) or at the intraspecific level (among a collection of 47 strains of *Streptococcus thermophilus*), a HGT origin can be predicted for a large proportion (about two thirds) (Eng, 2010).

[4] http://genome.jouy.inra.fr/ssb/SHOW/

4.2 Mining agricultural landscapes

In agricultural landscapes, land-use (LU) categories are heterogeneously distributed among different agricultural fields managed by farmers. At a first glance, the landscape spatial organization and its temporal evolution seem both random. Nevertheless, they reveal the presence of logical processes and driving forces related to the soil, climate, cropping system, and economical pressure. The mosaic of fields together with their land-use can be seen as a noisy picture generated by these different processes.

Recent studies (Le Ber et al., 2006; Castellazzi et al., 2008) have shown that the ordered sequences of LU in each field can be adequately modelled by a high order Markov process. The LU at time t depends upon the former LU at previous times: $t-1, t-2 \ldots$ depending on the order of the Markov process. In the space domain, the theory of the random Markov fields is an elegant mathematical way for accounting neighbouring dependencies (Geman & Geman, 1984; Julian, 1986). In this section, we present a data mining method based on CARROTAGE to cluster a landscape into patches based on its pluri annual LU organization. Two medium-size agricultural landscapes will be considered coming from different sources: long-term LU surveys or remotely sensed LU data.

4.2.1 Data preparation

For CARROTAGE, the input corpus of LU data is an array in which the columns represent the LU year by year and the rows represent regularly spaced locations in the studied landscape (e.g. 1 point every 20 m). Data preparation aims at reducing the requirement of the memory resources while putting the data in the appropriate format required by CARROTAGE. The data preparation process must tackle several issues: (i) to regroup into LU categories the different LU when there are too many observations, (ii) to define the elementary observation for the HMM, and (iii) to choose the sampling spatial resolution.

The corpus of spatiotemporal LU data is generally built either from long-term LU surveys or from remotely sensed LU data. Depending on the data source, several differences in the LU database may exist. These differences are mostly 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 (long-term LU field surveys) is illustrated by the Niort Plain case study (Lazrak et al., 2010), and the second (remotely sensed LU) is illustrated by the Yar watershed case study. Principal characteristics of the two case studies are summarized in table 1.

	Case study	
	Niort Plain	**Yar watershed**
Data source	Land-use surveys	Remote sensing
Surface (sq. km)	350	60
Study period	1996 to 2007	1997 to 2008
Number of LU modalities	47	6
Spatial representation	Vector	Raster (converted to vector)
Elementary spatial entities	Elementary plots (polygons)	Pixels (20 x 20 sq. m)
Data base format	ESRI Shapefile	ESRI Shapefile

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

4.2.1.1 The agricultural landscape mosaic

The agricultural landscape can be seen as an assemblage of polygons of variable size where each polygon holds a given LU. When data derives from LU surveys, the polygons are fields bounded by a road, a path or a limit of a neighbouring field. The polygon boundaries can change every year. To take into account this change, the surveyors update each year the boundaries of fields in the GIS database. For remotely sensed images, the polygons are obtained by grouping similar pixels in the same class and are represented in vector format. In the two cases, the list of the polygon boundaries –that change over the time– led to the definition of the elementary polygon –the plot– as the result of the spatial union of previous polygon boundaries (Figure 8). Each plot holds one LU succession during the study period. There are about 20,000 elementary plots in the Niort study area over the 1996 – 2007 period.

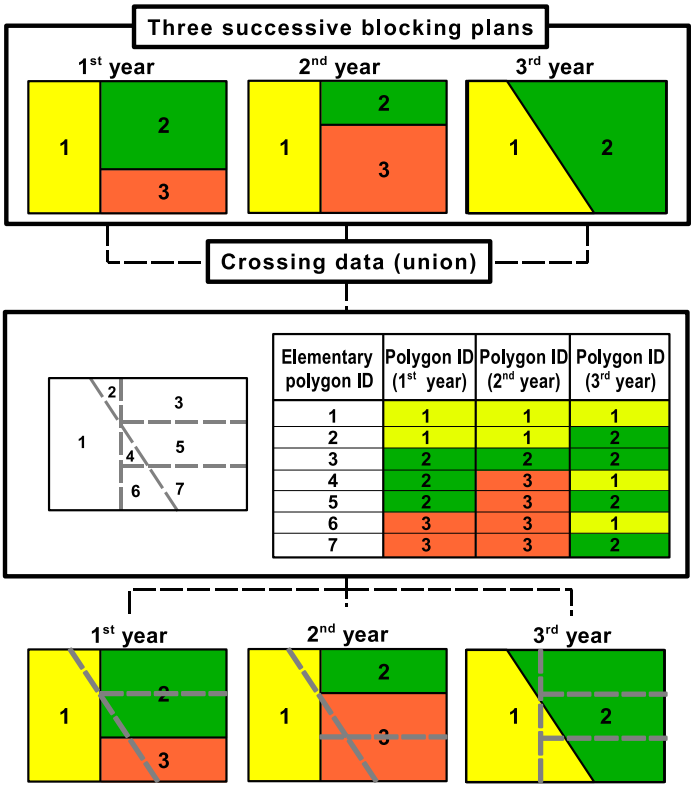


Fig. 8. An example of field boundary evolution over three successive years. The union of field boundaries during this period leads to the definition of seven plots

The corpus of land-use data is next sampled and is represented in a matrix in which the columns are related to the time slots and the rows to the different grid locations.
Following Benmiloud and Pieczynski (Pieczynski, 2003), we have approximated the Markov random field by sampling the 2-D landscape representation using a regular grid and, next, defining a scan by a Hilbert-Peano curve (figure 9). The Markov field is then represented

by a Markov chain. Two successive points in the Markov chain represent two neighbour 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 (MRF) modelling (Benmiloud & Pieczynski, 1995). To take into account the irregular neighbour system, we can also adjust the fractal depth to the mean plot size. The figure 9 illustrates this concept.

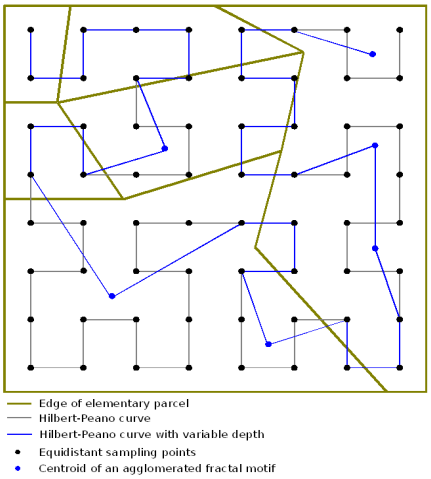


Fig. 9. Variable depth Hilbert-Peano scan to take into account the field size. Two successive merging in the bottom left field yield to the agglomeration of 16 points

4.2.1.2 LU categories definition

When LU derive from LU surveys, there is often a great number of LU modalities which must be reduced by defining LU categories. For the Niort Plain case study, the 47 LU have been grouped with the help of agricultural experts into 10 categories (see Tab. 2) following an approach based on the LU frequency in the spatiotemporal database and the similarity of crop management.
For the Yar watershed case study, only six LU have been distinguished: Urban, Water, Forest, Grassland, Cereal and Maize. There was no need of grouping them into categories.

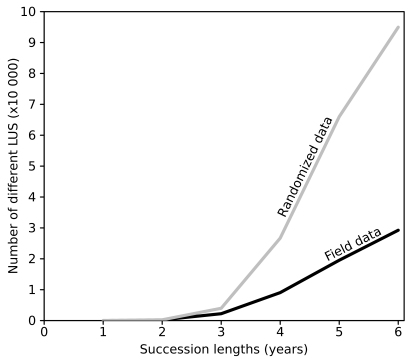
4.2.1.3 Choice of the elementary observation

An elementary observation can range from a LU (such as Cereal in the Yar watershed case study) or a LU category (such as Wheat in the Niort Plain case study) to a LU succession (LUS) spanning several years. For this latter, the length of the LU succession influences the interpretation of the final model. However, the total number of LUS is a power function of the succession length, and memory resources required during the estimation of HMM2 parameters increase dramatically.
To determine the succession length, we compared the diversity of LUS between field-collected data (the Niort Plain) and randomly generated data for different lengths of successions

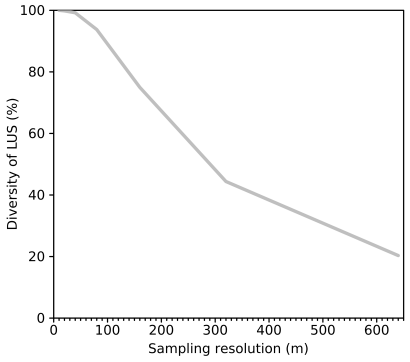
LU category	LU	Frequency	Cumul
Wheat	Wheat, bearded wheat, cereal	0.337	0.337
Sunflower	Sunflower, ryegrass followed by sunflower	0.139	0.476
Rapeseed	Rapeseed	0.124	0.600
Urban	Built area, peri-village, road	0.096	0.696
Grassland	Grassland of various types, alfalfa,...	0.078	0.774
Maize	Maize, ryegrass followed by maize	0.076	0.850
Forest	Forest or hedge, wasteland	0.034	0.884
Winter barley	Winter barley	0.034	0.918
Ryegrass	Ryegrass, ryegrass followed by ryegrass	0.024	0.942
Pea	Pea	0.022	0.964
Others	Spring barley, grape vine, clover, field bean, ryegrass, cereal-legume mixture, garden/market gardening,...	0.036	1.000

Table 2. Composition and average frequencies of adopted LU categories (Lazrak et al., 2010)

(Fig. 10(a)). For this case study, 4-year successions begin to clearly differentiate the landscape from a random landscape in which the LU are randomly allocated in the plots. Therefore, 4-year successions appear to be the shortest HMM2 elementary observation symbol suitable for modelling LUS within the Niort Plain landscape. The choice for the elementary observation can also be set by domain specialists based on previous works (Le Ber et al., 2006; Mignolet et al., 2007). This was the case for the Yar watershed where we chose to model the agricultural dynamics through 3-year LUS.



(a) Compared diversity of LUS between field-collected data and 10 random generated data sets for different succession lengths



(b) Information loss in terms of LUS diversity in relation to sampling resolutions for 4-year LUS

Fig. 10. Relations between LUS diversity and sampling rates

4.2.1.4 Choice of the spatial resolution

For medium-size and large landscapes, a high-resolution sampling generates a large amount of data. With such amount, only rough models can be tested. On the other hand, with a coarse resolution sampling, small fields are omitted. In order to have an objective criterion for choosing the optimal spatial resolution, we can estimate information loss in terms of LUS diversity for increasingly coarse resolution samplings. Figure 10(b) shows the obtained curve for the Niort Plain case study. The tested resolutions were: 10, 20, 40, 80, 160, 320 and 640 m. Irregularity in sampling intervals is dictated by an algorithmic constraint: the resolution must be proportional to a power of 2. The most precise resolution is considered as the reference (100%). As a compromise, we chose the 80 m x 80 m resolution that led to a corpus 64 times smaller than the original one, with only a loss of 6% in information diversity.
For the Yar watershed landscape, which has a surface roughly 7 times smaller than the Niort Plain landscape and has few LU modalities, we were not constrained by the corpus size. Thus, we chose a 20 m x 20 m resolution which was the original resolution of satellite images used to identify the LU.

4.2.2 *a posteriori* decoding

We propose to build a time spatial analysis through spatial analysis of crop dynamics. This data mining method is a time x space analysis where a temporal analysis is performed in order to identify temporal regularities before locating these regularities in the landscape by means of a hierarchical HMM2 (HHMM2). The HHMM2 allows segmenting the landscape into patches, each of them being characterized by a temporal HMM2.

4.2.2.1 Mining temporal regularities

Depending on the investigated temporal regularities, we can either use a linear HMM2 or a multi-column ergodic HMM2 (Fig. 12). Linear models allow segmenting the study period into homogeneous sub-periods in terms of LUS distributions (see Figure 11).

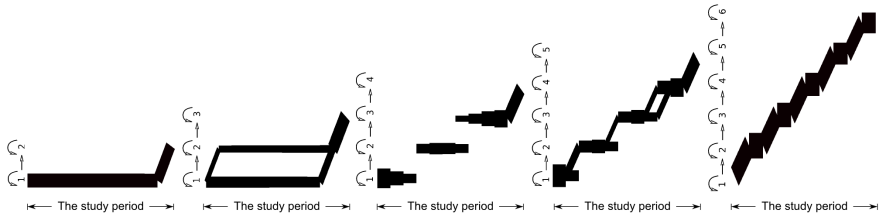


Fig. 11. Seeking the best temporal segmentation of the Yar watershed study period by using 5 growing state number linear HMM2. The line width is proportional to the *a posteriori* transition probability (Eq. 6). The 6 state HMM2 segments the study period into 6 non-overlapping periods

Multi-column ergodic models (Mari & Le Ber, 2006; Le Ber et al., 2006) (Fig. 12) have been designed for measuring the probability of a succession of 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. Figure 12 shows the topology of a `HMM2` that has two kinds of states: Dirac states associated to the most frequent land-use categories

(wheat, sunflower, 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. Therefore this model generalises both hidden Markov models and Markov models.

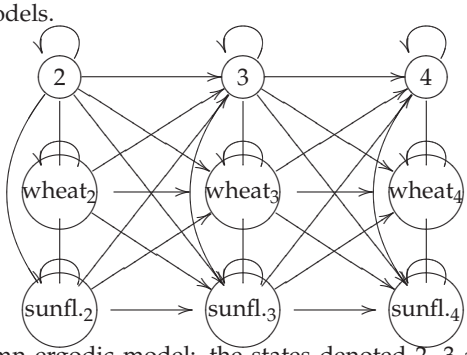


Fig. 12. Multiple column ergodic model: the states denoted 2, 3 and 4 are associated to a distribution of land-use categories, as opposite to the *Dirac* states denoted with a specific land-use category. The number of columns determines the number of time intervals (periods). A connection without arrow means a two directional connection

The model generation follows the same flowchart given in figure 5. When it is needed, the *Dirac states* can be initialized by some search patterns for capturing one or many particular observations.

Agronomists interpret the resulting diagrams to find the LU dynamics. Figure 13 shows a quasi steady agricultural system. The crop rotations involve Rapeseed, Sunflower and Wheat. In order to determine the exact rotations (2-year or 3-year), it is necessary to envisage the modelling of 4-year LUS (Lazrak et al., 2010). Note the monoculture of Wheat that starts in 2004.

4.2.2.2 Spatial clustering based on HMM2

We model the spatial structure of the landscape by a MRF whose sites are random LUS. The dynamics of these LUS are modelled by a temporal HMM2. This leads to the definition of a hierarchical HMM2 (Figure 14) where a master HMM2 approximates the MRF. Then, the probability of LUS is given by a temporal HMM2 as fully described in (Fine et al., 1998; Mari & Le Ber, 2006; Lazrak et al., 2010). This hierarchical HMM is used to segment the landscape into patches, each of them being characterized by a temporal HMM2. At each index l in the Hilbert-Peano curve, we look for the best *a posteriori* state in the HHMM2 (Maximum Posterior Mode algorithm). The state labels, together with the geographic coordinates of the indices l, determine a clustered image of the landscape that can be coded within an ESRI shapefile. An example of this segmentation for the Yar watershed case study is given in Figure 15.

4.2.3 Post Processing

For the Yar watershed case study, we have performed preliminary temporal segmentation tests with linear models having an increasing number of states (Figure 11). This led us to use a 6-state HMM2 to segment the study period into 6 sub-periods characterized by different `pdf`. Plotting together the 6 sub-periods gives a global view on the LU dynamics (Figure 15).

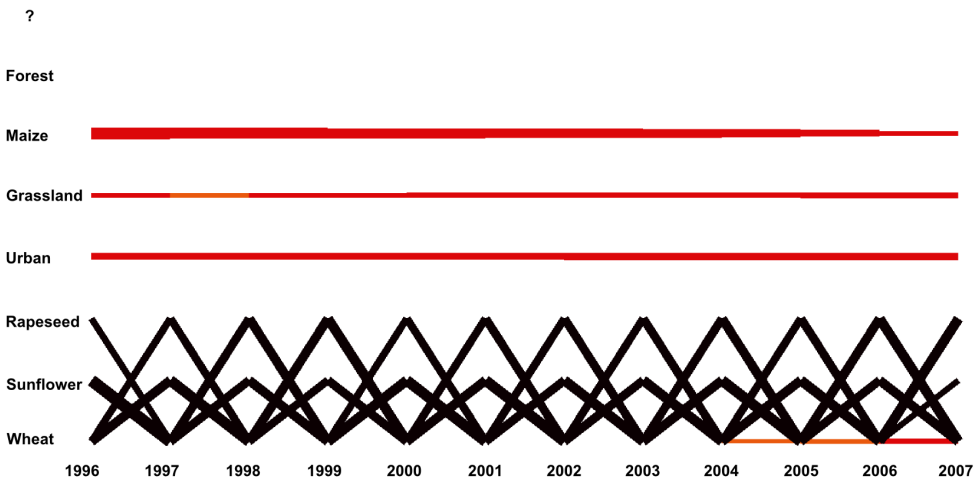


Fig. 13. Markov diagram showing transitions between LU categories in the Niort Plain. The x-axis represents the study period. The y-axis stands for the states of the ergodic one-column HMM2 used for data mining. Each state represents one LU category. The state '?' is the *container* state associated to a `pdf`. Diagonal transitions stand for inter-annual LU changes. Horizontal transitions indicate inter-annual stability. For simplicity, only transitions whose frequencies are greater than 5 % are displayed. The line width reflects the *a posteriori* probability of the transition assuming the observation of the 12-year LU categories (Eq. 6)

In figure 15, the Yar watershed is represented by a mosaic of patches of LU evolutions. These patches 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 LU 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 shows very likely that a change of cropping system was undertaken in the patches belonging to this state.

4.3 Mining hydro-morphological data

In this section we describe the use of `HMM2` for the segmentation of data describing river channels. Actually, a river channel is considered as a continuum and is characterised by its width or depth that is increasing downstream whereas its slope and grain size decrease (Schumm, 1977). The segmentation of this continuum with respect to local characteristics is an important issue in order to better manage the river channels (e.g. protection of plant or animal species, prevention of flood or erosion processes, etc.). Several methods have been proposed to perform such a segmentation. Markov chains Grant et al. (1990) and `HMM1` (Kehagias, 2004) are also been used.

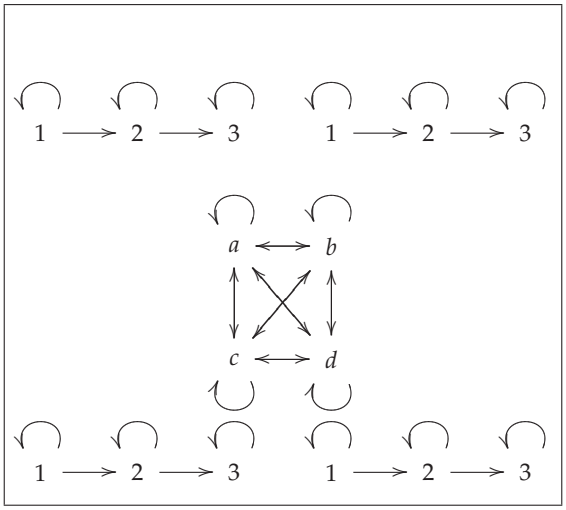


Fig. 14. Example of hierarchical HMM2. Each spatial state a, b, c, d of the master HHMM2 (ergodic model) is a temporal HMM2 (linear model) whose states are 1, 2, 3

4.3.1 Data preparation

The aim is to establish homogeneous units of the river Drome (South-East of France) continuum according to its geomorphological features. First of all, the continuum has been segmented within 406 segments of 250 meters length. Each segment is then described with several variables computed from aerial photographs (years 1980/83 and 1994/96) supplemented with terrain observations. Details about the computing of these variables can be found in (Aubry & Piégay, 2001; Alber & Piégay, 2010; Alber, 2010). In the following, we focus on the variable describing the width of the active channel (i.e. the water channel and shingle banks without vegetation).

4.3.2 *a posteriori* decoding

The stochastic modelling follows the same flow chart given in figure 5. Both linear and ergodic models have been used. The `pdf` associated in the `M2-M0 HMM` are univariate Gaussian $\mathcal{N}(\mu_i, \Sigma_i)$.

$$b_i(O_t) = \mathcal{N}(O_t; \mu_i, \Sigma_i) \tag{7}$$

where O_t is the input vector (the frame) at index t and $\mathcal{N}(O_t; \mu, \Sigma)$ the expression of the likelihood of O_t using a gaussian density with mean μ and variance Σ. The maximum likelihood estimates the mean and covariance are given by the formulas using the definition of P_0 (cf. Equ.3):

$$\overline{\mu_i} = \frac{\sum_t P_0(i,t) O_t}{\sum_t P_0(i,t)} \tag{8}$$

$$\overline{\Sigma_i} = \frac{\sum_t P_0(i,t)(O_t - \mu_i)(O_t - \mu_i)^t}{\sum_t P_0(i,t)} \tag{9}$$

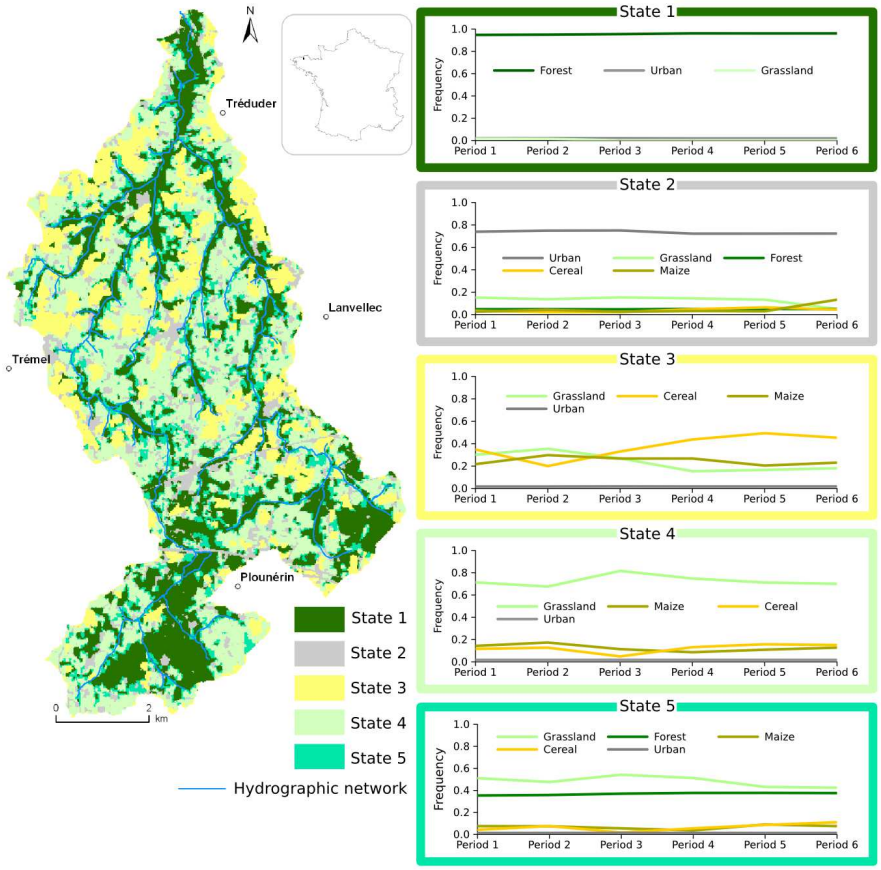


Fig. 15. The Yar watershed seen as patches of LU dynamics. Each map unit stands for a state of the HHMM2 used to achieve the spatial segmentation. Each state is described by a diagram of the LU evolution. The 6 sub-periods are the time slots derived from the temporal segmentation with the 6-state HMM2 describing each state of the HHMM2. Location of the Yar watershed in France is shown by a black spot depicted in the upper middle box

Specific user interfaces have been designed, in order to fit the experts' requirements: the original data are plotted, together with the mean value and the standard deviation of the current (most probable) state.

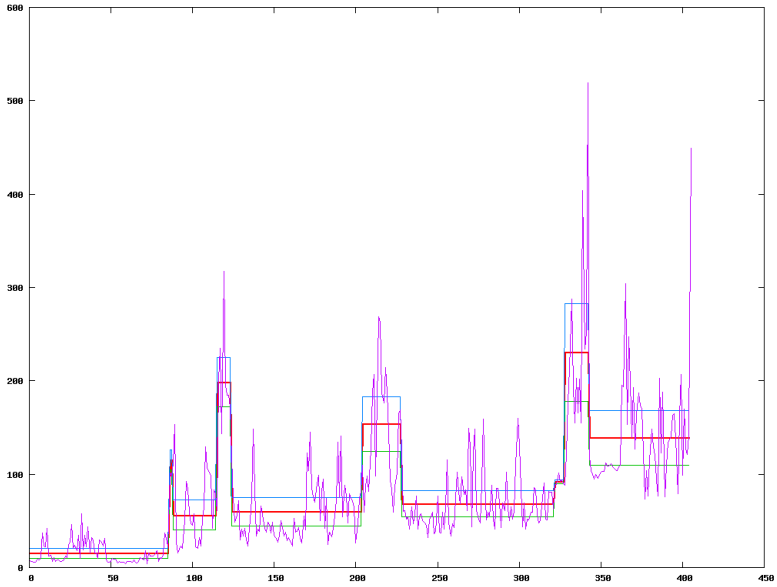


Fig. 16. Clustering the active channel width of the Drome river: linear HMM2 with 10 states

The linear model (Fig. 16) allows to detect a limited number (due to the specified number of states) of high variations, i.e. large and short vs narrow and long sections of the river channel. The ergodic model (Fig. 17) allows to detect an unknown number of small variations and repetitions.

4.3.3 Post processing

The final aim of this study is to build a geomorphical typology based on the river characteristics and to link it to external criteria (e.g. geology, land-use). The clustering is useful to define a relevant scale for this typology. If the typology is limited to the Drome river, the linear HMM allows to detect a set of segments that can be characterised by further variables and used as a basis for the typology. Ten segments for 101.5 kilometres appeared to be a good scale. On the contrary, if a whole network is considered -with several rivers and junctions-, the segmentation performed by the ergodic HMM would be more interesting since it allows to segment the data with less states than the linear model and to reveal similar zones (i.e. belonging to the same state) in the network. The probability transitions between states can also be exploited to reveal similar sequences of states along the network and thus to perform nested segmentations. Furthermore, transition areas appearing as significant mixtures of several states may be dealt with separately or excluded from a typology. Specific algorithms have to be designed and tuned to deal with these last questions.

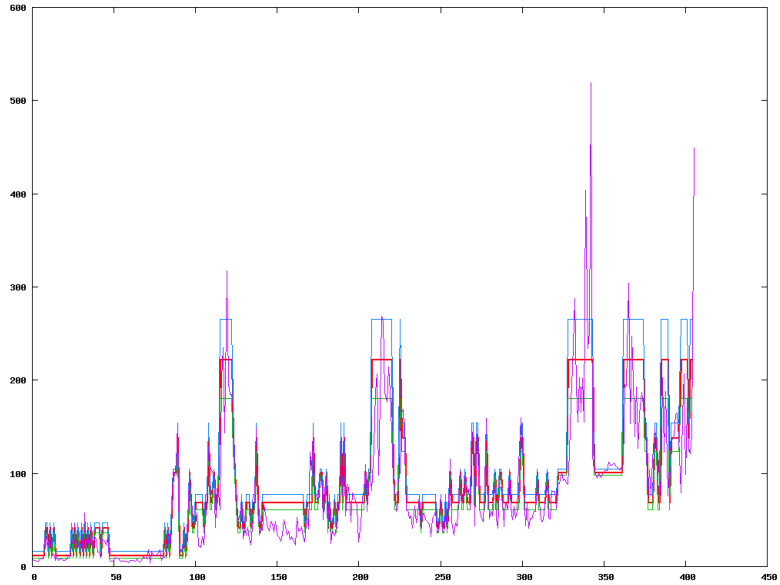


Fig. 17. Clustering the active channel width of the Drome river: ergodic HMM2 with 6 states

5. Conclusions

We have described in this chapter a general methodology to mine temporal and spatial data based on a high order Markov modelling as implemented in CARROTAGE. The data mining is basically a clustering process that voluntary implements a minimum amount of knowledge. The HMM maps the observations into a set of states generated by a Markov chain. The classification is performed, both in time domain and spatial domain, by using the *a posteriori* probability that the stochastic process stays in a particular state, assuming a sequence of observations. We have shown that spatial data may be re-ordered using a fractal curve that preserves the neighbouring information. We adopt a Bayesian point of view and measure the temporal and the spatial variability with the *a posteriori* probability of the mapping. Doing so, we have a coherent processing both in temporal and spatial domain. This approach appeared to be valuable for time space data mining.

In the genomic application, two different HMM (M2-M0 HMM and M2-M2 HMM) have extracted meaningful regularities that are of interest in the area of promoter and HGT detection. The dependencies in the observation sequence smooth dramatically the *a posteriori* probability. We put forward the hypothesis that this smoothing effect is due to the additional normalisation constraints used to transform a 64 bin pdf of *3-mer* into 16 pdf of nucleotides. This smoothing effect allows the extraction of wider regularities in the genome as it has been shown in the HGT application.

In the agronomic application, the hierarchical HMM produces a time space clustering of agricultural landscapes based on the LU temporal evolution that gives to the agronomist a concise view of the current trends. CARROTAGE is an efficient tool for exploring large land use databases and for revealing the temporal and spatial organization of land use, based on crop

sequences (Mari & Le Ber, 2003). Furthermore, this mining strategy can also be used to investigate and visualize the crop sequences of a few specific farms or of a small territory. In a recent work (Schaller et al., 2010) aiming at modelling the agricultural landscape organization at the farm and landscape levels, the stochastic regularities have been combined with farm surveys to validate and explain the individual farmer decision rules. 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 (Mignolet et al., 2004).
In the mining of hydro-morphological data, the `HMM` have given promising results. They could be used to perform nested segmentations and reveal similar zones in the hydrological network. We are carrying out extensive comparisons with other methods in order to assess the gain given by the high order of the Markov chain modelling.
In all these applications, the extraction of regularities has been achieved following the same flowchart that starts by the estimation of a linear `HMM` to get initial seeds for the probabilities and, next, a linear to ergodic transform followed by a new estimation by the forward backward algorithm. Even if the data do not suit the model, the `HMM` can give interesting results allowing the domain specialist to put forward some new hypothesis. Also, we have noticed that the data preparation is a time consuming process that conditions all further steps of the data mining process. Several ways of encoding elementary observations have been tried in all applications during our interactions with the domain specialists.
A much discussed problem is the automatic design of the `HMM` topology. So far, CARROTAGE does not implement any tools to achieve this goal. We plan to improve CARROTAGE by providing it with these tools and assess this new feature in the numerous study cases that we have already encountered. Another new trends in the area of artificial intelligence is the clustering of both numerical and symbolic data. Also, based on their transition probabilities and `pdf`, the `HMM` could be considered as objects that have to be compared and clustered by symbolical methods. The frequent items inside the `pdf` can be analyzed by frequent item set algorithms to achieve a description of the intent of the classes made of the most frequent observations that have been captured in each state in the `HMM`. These issues must be tackled if we want to deal with different levels of description for large datasets.

6. Acknowledgments

Many organizations had provided us with support and data. The genetic data mining work was supported by INRA, the région Lorraine and the ACI IMP-Bio initiative. Hydro-morphological data were provided by H. Piégay and A. Alber, UMR 5600 CNRS, Lyon. The original idea of this particular work arose from discussions with T. Leviandier, ENGEES, Strasbourg. The agronomic work was supported by the ANR-ADD-COPT project, the API-ECOGER project, the région Lorraine and the ANR-BiodivAgrim project. We thank the two CNRS teams: UPR CEBC (Chizé) for their data records obtained from the "Niort Plain database" and UMR COSTEL (Rennes) for the "Yar database".

7. References

Alber, A. (2010). PhD thesis, U. Lyon 2, France. to be published.

Alber, A. & Piégay, H. (2010). Disaggregation-aggregation procedure for characterizing spatial structures of fluvial networks: applications to the Rhône basin (France), *Geomorphology* . In press.

Aubry, P. & Piégay, H. (2001). Pratique de l'analyse de l'autocorrélation spatiale en géomorphologie fluviale : définitions opératoires et tests, *Géographie Physique et Quaternaire* **55**(2): 115–133.

Baker, J. K. (1974). Stochastic Modeling for Automatic Speech Understanding, *in* D. Reddy (ed.), *Speech Recognition*, Academic Press, New York, New-York, pp. 521 – 542.

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

Bize, L., Muri, F., Samson, F., Rodolphe, F., Ehrlich, S. D., Prum, B. & Bessières, P. (1999). Searching Gene Transfers on *Bacillus subtilis* Using Hidden Markov Models, *RECOMB'99*.

Castellazzi, M., Wood, G., Burgess, P., Morris, J., Conrad, K. & Perry, J. (2008). A systematic representation of crop rotations, *Agricultural Systems* **97**: 26–33.

Charniak, E. (1991). Bayesian Network without Tears, *AI magazine* .

Churchill, G. (1989). Stochastic Models for Heterogeneous DNA Sequences, *Bull Math Biol* **51**(1): 79 – 94.

Delcher, A., Kasif, S., Fleischann, R., Peterson, J., White, O. & Salzberg, S. (1999). Alignment of whole genomes, *Nucl. Acids Res.* **27**(11): 2369 – 2376.

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

Eng, C. (2010). *Développement de méthodes de fouille de données fondées sur les modèles de Markov cachés du second ordre pour l'identification d'hétérogénéités dans les génomes bactériens*, PhD thesis, Université Henri Poincaré Nancy 1. http://www.loria.fr/~jfmari/ACI/these_eng.pdf.

Eng, C., Asthana, C., Aigle, B., Hergalant, S., Mari, J.-F. & Leblond, P. (2009). A new data mining approach for the detection of bacterial promoters combining stochastic and combinatorial methods, *Journal of Computational Biology* **16**(9): 1211–1225. http://hal.inria.fr/inria-00419969/en/.

Eng, C., Thibessard, A., Danielsen, M., Rasmussen, T., Mari, J.-F. & Leblond, P. (2011). In silico prediction of horizontal gene transfer in *Streptococcus thermophilus*, *Archives of Microbiology* . in preparation.

Fine, S., Singer, Y. & Tishby, N. (1998). The Hierarchical Hidden Markov Model: Analysis and Applications, *Machine Learning* **32**: 41 – 62.

Forbes, F. & Pieczynski, W. (2009). New Trends in Markov Models and Related Learning to Restore Data, *IEEE International Workshop on Machine Learning for Signal Processing (MSLP)*, IEEE, Grenoble.

Forney, G. (1973). The Viterbi Algorithm, *IEEE Transactions* **61**: 268–278.

Furui, S. (1986). Speaker-independent Isolated Word recognition Using Dynamic Features of Speech Spectrum, *IEEE Transactions on Acoutics, Speech and Signal Processing* .

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

Grant, G., Swanson, F. & Wolman, M. (1990). Pattern and origin of stepped-bed morphology in high-gradient streams, Western Cascades, Oregon, *Geological Society of America Bulletin* **102**: 340–352.

Hoebeke, M. & Schbath, S. (2006). R'mes: Finding exceptional motifs. user guide, *Technical report*, INRA.
URL: *http://genome.jouy.inra.fr/ssb/rmes*

Huang, H., Kao, M., Zhou, X., Liu, J. & Wong, W. (2004). Determination of local statistical significance of patterns in markov sequences with application to promoter element identification, *Journal of Computational Biology* **11**(1).

Jain, A., Murty, M. & Flynn, P. (1999). Data Clustering: A Review, *ACM Computing Surveys* **31**(3): 264 – 322.

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

Kehagias, A. (2004). A hidden Markov model segmentation procedure for hydrological and environmental time series, *Stochastic Environmental Research* **18**: 117–130.

Lazrak, E., Mari, J.-F. & Benoît, M. (2010). Landscape regularity modelling for environmental challenges in agriculture, *Landscape Ecology* **25**(2): 169 – 183. http://hal.inria.fr/inria-00419952/en/.

Le Ber, F., Benoît, M., Schott, C., Mari, J.-F. & Mignolet, C. (2006). Studying Crop Sequences With CarrotAge, a HMM-Based Data Mining Software, *Ecological Modelling* **191**(1): 170 – 185. http://hal.archives-ouvertes.fr/hal-00017169/fr/.

Li, C., Bishas, G., Dale, M. & Dale, P. (2001). *Advances in Intelligent Data Analysis*, Vol. 2189 of *LNCS*, Springer, chapter Building Models of Ecological Dynamics Using HMM Based Temporal Data Clustering – A Preliminary study, pp. 53 – 62.

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

Mari, J.-F. & Le Ber, F. (2003). Temporal and spatial data mining with second-order hidden markov models, *in* M. Nadif, A. Napoli, E. S. Juan & A. Sigayret (eds), *Fourth International Conference on Knowledge Discovery and Discrete Mathematics - Journées de l'informatique Messine - JIM'2003, Metz, France*, IUT de Metz, LITA, INRIA, pp. 247–254.

Mari, J.-F. & Le Ber, F. (2006). Temporal and Spatial Data Mining with Second-Order Hidden Markov Models, *Soft Computing* **10**(5): 406 – 414. http://hal.inria.fr/inria-00000197.

Mari, J.-F. & Schott, R. (2001). *Probabilistic and Statistical Methods in Computer Science*, Kluwer Academic Publishers.

Mignolet, C., Schott, C. & Benoît, M. (2007). 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. http://www.sciencedirect.com/science/article/B6V78-4N3P539-2/2/562034987911fb9545be7fda6dd914a8.

Mignolet, C., Schott, C. & Benoît, M. (2004). 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.

Nicolas, P., Bize, L., Muri, F., Hoebeke, M., Rodolphe, F., Ehrlich, S. D., Prum, B. & Bessières, P. (2002). Mining *Bacillus subtilis* Chromosome Heterogeneities Using Hidden Markov Models, *Nucleic Acids Research* **30**(6): 1418 – 1426.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference*, Morgan Kaufman.

Pieczynski, W. (2003). Markov models in image processing, *Traitement du signal* **20**(3): 255–278.

Rabiner, L. & Juang, B. (1995). *Fundamentals of Speech Recognition*, Prentice Hall.

Schaller, N., Lazrak, E.-G., Martin, P., Mari, J.-F., Aubry, C. & Benoît, M. (2010). Modelling regional land use: articulating the farm and the landscape levels by combining farmers'

decision rules and landscape stochastic regularities, Poster session, European Society of Agronomy. Agropolis2010, Montpellier.
Schumm, S. (1977). *The fluvial system*, Wiley, New York. 338p.
Tou, J. T. & Gonzales, R. (1974). *Pattern Recognition Principles*, Addison-Wesley.
Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*, Wiley.