# A Descriptive Method to Evaluate the Number of Regimes in a Switching Autoregressive Model

**Madalina Olteanu**

SAMOS-MATISSE, Universite Paris I

90 Rue de Tolbiac, 75013 Paris, France

**madalina.olteanu@univ-paris1.fr**

**Abstract -** *This paper proposes a descriptive method for an open problem in time series analysis : determining the number of regimes in a switching autoregressive model. We will translate this problem into a classification one and define a criterion for hierarchically clustering different model fittings. Finally, the method will be tested on simulated examples and real-life data.*

**Key words - switching autoregressive models, hierarchical clustering, Ward distance, SOM**

## 1 Introduction

In the past few years, several nonlinear autoregressive models have been proposed for time series analysis. Some of these models are based on the idea that the process is characterized not by a unique autoregression, but by the fact that two or more regimes are driving the series behaviour. In each regime, an autoregressive function is fitted. We are interested in the case where the autoregressive functions are linear in every regime. The most classical examples are TAR (Threshold Autoregressive) models introduced by Tong (1978) with regimes switching according to the magnitude of a threshold variable, the smoothed version of TAR models (STAR), or the more recent Markov switching autoregressive models, first used by Hamilton (1989) to model the U.S. Gross National Product.

Estimating the parameters of these models is usually done by maximizing the likelihood function, but under a very strong hypothesis: a fixed number of regimes. Choosing the "true" number of regimes is still an open problem, as this is equivalent to testing with lack of identifiability under the null hypothesis. This leads to a degenerated Fisher information matrix and thus the chi-square theory and the likelihood ratio tests fail to apply. An empirical method to detect this kind of non-linearity using Kohonen maps and hierarchical clustering of linear regressions is given below. The second section describes the method and places it among the existing literature on this subject. In the third part, we give examples on simulated and real-life data and, finally, a conclusion will follow.

## 2 The Method

The problem of finding the "true" number of regimes can be rewritten as a classification problem by using a sliding window as follows. Suppose that we have observed the values of a time series $\{y_t\}_{t=\overline{1,T}}$ and we decide to fit an autoregressive model. Once the order of the model has been determined (with an AIC criterion, for example), we can consider the data set of dimension $(T-p)$x$(p+1)$, $\{y_t, y_{t-1}, ...y_{t-p}\}_{t=\overline{p+1,T}}$. Looking for the number of regimes is actually equivalent to looking for the number of regression lines (or hyperplanes) which will best fit the data.

The idea is simple and is based on the possibility of finding patterns in data which will identify the regression hyperplanes. Given the data set in Figure 1, fitting one regression line to the data is clearly not the good choice. If we now suppose that we managed to cluster the data into two groups and we perform a regression within each of these groups, we get two lines which seem to describe the sample better. This is confirmed by the "within-squared error", which is equal to the sum of squared residuals if there is only one regime and, if there are several, to the total sum of squared residuals within each group.

Now, let us remark that a classification that would find a good separation and, implicitly, the regression lines which best fit the data will be strongly connected to the underlying model from which the observations were sampled. Although the number of existing methods for

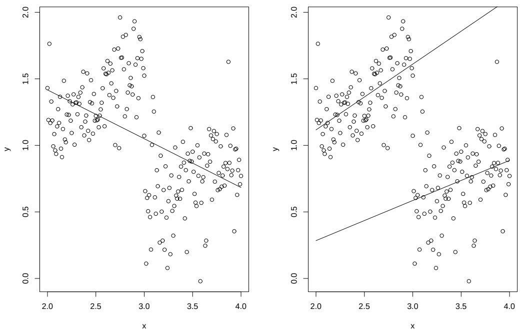

Figure 1: Fitting clustered data

clustering is very large, hundreds of algorithms being available, most of them emphasize the characteristics of the sample instead of the model generating it. Still, the idea of classifying data using conditions referring rather to the model underlying them is not new in the literature. The concept of probabilistic clustering is present in some clustering methods, whether we speak of fixed-partition methods such as regression-type clustering, principal component clustering, projection pursuit or convex support, the mixture models or the high-density clusters.

The method we propose here is close to the regression-type clustering, the concept behind it being the same : identify the hyperplanes which will best fit the data in the sense of a sum of squared-residual criterion or a sum of norms of the orthogonal projections of the points onto the subspace. Introduced by Charles [1977] and developed afterwards by Spath [1979], De Sarbo [1988], Lou, Jiang and Keng[1993], this type of clustering assumes, for a fixed number of classes, that each of them is characterized by a specific regression line. The algorithm is then an extension of $k$-means, the centroids being replaced by the regression hyperplanes. If a “good” number of classes is needed, one should perform this modified $k$-means for a number of classes going from 1 to a sufficiently large $N$ and pick the configuration which minimizes the chosen criterion.

In our case, we would like to start with some “good” initial clusters which will then be classified hierarchically using some squared-error criterion and we would expect to have an important break in the increasing values of this criterion, once we pass from the true number of regimes to a smaller one. A “good” cluster should contain observations belonging to the same regime and, at the same time, have enough points to estimate a regression line.

## 2.1 Initial clustering

For the initial clustering, self-organizing maps were used. Introduced by Kohonen (see, for instance Kohonen [1997]), the algorithm performs a clustering job and also a nonparametric regression at the same time and in a natural manner. Besides, contrary to $k$-means or other unsupervised classification methods, it has the advantage of preserving the topology of the data. Thus, not only will similar observations be in the same cluster, but close clusters on the map will contain similar data in the initial $p+1$-dimensional space. This will be helpful in the hierarchical classification afterwards as it will fasten the algorithm by considering only the neighbouring classes.

One problem arises once we get the clusters and try to fit a regression within each one : are there enough points in every cluster? No cluster will be allowed to have less points than the number of lags or regressor variables. To achieve this, either we eliminate from the analysis those which do not verify this condition, or we force the points to move to a different cluster, by assigning them to the closest sufficiently large cluster. Since very few observations are concerned with this problem and in order to shorten the computing time, the first approach was preferred.

The assumption we make here is that the property of homogeneity of self-organizing maps manages to create clusters in which observations belong to one regime. This could be justified by the fact that the variables used for the classification contain information concerning the regime of the observation and similar profiles will belong to the same cluster. Indeed, in the simulated examples where the different regimes are known, we will see that the map clusters are generally homogeneous from this point of view. This property will no longer apply if the regression hyperplanes are too close and the noise is important.

## 2.2 Hierarchical classification

As we actually need to compare different data fits, which is also equivalent to different numbers of hyperplanes, we need to adapt a hierarchical classification to our case (let us first make the convention to call “clusters” the result of the Kohonen map and “classes” two or more “clusters” joined together by the hierarchical method). We will choose a new “distance” between classes by developing a squared-error criterion.

A very popular method used in classification is to minimize a within-class variation criterion, the variation within a class being defined as the sum of squared distances from the individuals to the barycentre. In hierarchical classification, this principle was adapted by Ward and the algorithm consists in joining together the individuals which minimize the increase of the within-class variation. Our idea was to build an algorithm similar to Ward’s, but, as our interest is to estimate the number of hyperplanes characterizing the data, we replace the barycentres by regression lines and the within-class variation becomes the within sum of squared errors. Obviously, in this case, the between-class variation cannot be defined.

For a fixed number of classes $k$, the within sum of squared errors is defined as

$$SSE_{w,k} = \sum_{l=1}^{k} SSE_{C_l},$$

where $SSE_{C_l} = \sum_{t \in C_l} (y_t - \hat{y}_t)^2$ is the sum of squared residuals and $\hat{y}_t$ is the predicted value of $y_t$ by the linear regression of order $p$ fitted in class $C_l$, $l = \overline{1,k}$. Now, in the frame of hierarchical classification, when passing from $k$ to $k-1$ classes, if classes $i$ and $j$ are grouped, the within sum of squared errors becomes :

$$SSE_{w,k-1}^{i,j} = \sum_{l=1, l \neq i, l \neq j}^{k} SSE_{C_l} + SSE_{C_i \cup C_j}$$

Following the same principle as Ward’s, we want to minimize the increase in the within inertia, which in this case is defined by the within sum of squared errors. This is equivalent to finding $i$ and $j$ which minimize

$$\Delta S^{i,j}_{w,k,k-1} = SSE^{i,j}_{w,k-1} - SSE_{w,k} = SSE_{C_i \cup C_j} - SSE_{C_i} - SSE_{C_j}$$

### 2.2.1 The within sum of squared-error criterion

Now, let us take a closer look at the criterion we have chosen to consider and justify our choice. Newt, we will see, by giving an explicit expression of it as a function of the data and the residuals, that the increase in the within variation is close to zero when two classes with the same underlying model are grouped together and that there is a significant jump when the opposite situation occurs.

Suppose that the case where classes $C_i$ and $C_j$ are grouped is being investigated and we wonder if they are from the same "regime". We will consider the following notations :

- $Y_i = \{Y_t\}_{t \in C_i} \in \mathbb{R}^{n_i}$, $Y_j = \{Y_t\}_{t \in C_j} \in \mathbb{R}^{n_j}$, where $n_i$ and $n_j$ are the cardinalities of $C_i$ and, respectively, $C_j$, are the observed values of the explained variable.

- $X_i = \{1, Y_{t-1}, ..., Y_{t-p}\}_{t \in C_i} \in \mathbb{R}^{n_i \times (p+1)}$, $X_j = \{1, Y_{t-1}, ..., Y_{t-p}\}_{t \in C_j} \in \mathbb{R}^{n_j \times (p+1)}$ are the regressors or the explaining variables.

The linear regressions fitted in classes $C_i$ and $C_j$ can be written as :

$$Y_i = X_i \cdot \beta_i + u_i = X_i \cdot \hat{\beta}_i + e_i$$
$$Y_j = X_j \cdot \beta_j + u_j = X_j \cdot \hat{\beta}_j + e_j,$$

where $\beta_i, \beta_j, \hat{\beta}_i, \hat{\beta}_j \in \mathbb{R}^{p+1}$, $\hat{\beta}_i$ and $\hat{\beta}_j$ are the least-square estimates of $\beta_i$ and $\beta_j$, $u_i, e_i \in \mathbb{R}^{n_i}$ and $u_j, e_j \in \mathbb{R}^{n_j}$ are the error and, respectively, the residual vectors, $u_i \sim N(0, \sigma_i^2 I_{n_i})$ and $u_j \sim N(0, \sigma_j^2 I_{n_j})$. Now, the linear regression fitted in the joint class $C_i \cup C_j$ is written as :

$$Y = X \cdot \beta + u = X \cdot \hat{\beta} + e,$$

where $Y = \begin{bmatrix} Y_i \\ Y_j \end{bmatrix} \in \mathbb{R}^{n_i + n_j}$, $X = \begin{bmatrix} X_i \\ X_j \end{bmatrix} \in \mathbb{R}^{(n_i+n_j) \times (p+1)}$, $\beta, \hat{\beta} \in \mathbb{R}^{p+1}$, $\hat{\beta}$ is the least-square estimate of $\beta$, $u, e \in \mathbb{R}^{n_i+n_j}$ are the error and the residuals vector and $u \sim N(0,\Omega)$, $\Omega = \left( \begin{array}{cc} \sigma_i^2 I_{n_i} & 0 \\ 0 & \sigma_j^2 I_{n_j} \end{array} \right)$. Then, we can compute $\Delta S_{w,k,k-1}^{i,j}$ as :

$$\Delta S_{w,k,k-1}^{i,j} = SSE_{C_i \cup C_j} - SSE_{C_i} - SSE_{C_j} = e^T e - \left( e_i^T e_i + e_j^T e_j \right)$$

**Remark 1 :**

Using this form, Toyoda (1974) proved that $\Delta S_{w,k,k-1}^{i,j}$ is approximately distributed as $\sigma^2 \chi^2 \left(p+1\right)$, where $\sigma^2$ is any well-chosen weighted average of $\sigma_i^2$ and $\sigma_j^2$ and $p$ is the number of lags considered.

$$\Delta S_{w,k,k-1}^{i,j} = SSE_{C_i \cup C_j} - SSE_{C_i} - SSE_{C_j} = \left\| Y - X \cdot \hat{\beta} \right\|^2 - \left\| Y_i - X_i \cdot \hat{\beta}_i \right\|^2 - \left\| Y_j - X_j \cdot \hat{\beta}_j \right\|^2 \quad (1)$$

But

$$Y - X \cdot \hat{\beta} = \left[ \begin{array}{c} Y_i - X_i \cdot \hat{\beta} \\ Y_j - X_j \cdot \hat{\beta} \end{array} \right] = \left[ \begin{array}{c} Y_i - X_i \cdot \hat{\beta}_i \\ Y_j - X_j \cdot \hat{\beta}_j \end{array} \right] + \left[ \begin{array}{c} X_i \cdot \hat{\beta}_i - X_i \cdot \hat{\beta} \\ X_j \cdot \hat{\beta}_j - X_j \cdot \hat{\beta} \end{array} \right] \quad (2)$$

and thus

$$\left\| Y - X \cdot \hat{\beta} \right\|^2 = \left\| \begin{array}{c} Y_i - X_i \cdot \hat{\beta} \\ Y_j - X_j \cdot \hat{\beta} \end{array} \right\|^2 = \left\| \begin{array}{c} Y_i - X_i \cdot \hat{\beta}_i \\ Y_j - X_j \cdot \hat{\beta}_j \end{array} \right\|^2 + \left\| \begin{array}{c} X_i \cdot \hat{\beta}_i - X_i \cdot \hat{\beta} \\ X_j \cdot \hat{\beta}_j - X_j \cdot \hat{\beta} \end{array} \right\|^2$$

since it can easily be seen that the cross-product on the right term in (2) is zero, we get :

$$\left\| Y - X \cdot \hat{\beta} \right\|^2 = \left\| Y_i - X_i \cdot \hat{\beta}_i \right\|^2 + \left\| Y_j - X_j \cdot \hat{\beta}_j \right\|^2 + \left\| X_i \cdot \hat{\beta}_i - X_i \cdot \hat{\beta} \right\|^2 + \left\| X_j \cdot \hat{\beta}_j - X_j \cdot \hat{\beta} \right\|^2 \quad (3)$$

and by replacing (3) in (1) :

$$\Delta S_{w,k,k-1}^{i,j} = \left\| X_i \cdot \hat{\beta}_i - X_i \cdot \hat{\beta} \right\|^2 + \left\| X_j \cdot \hat{\beta}_j - X_j \cdot \hat{\beta} \right\|^2 =$$

$$= (\hat{\beta}_i - \hat{\beta})^T X_i^T X_i (\hat{\beta}_i - \hat{\beta}) + (\hat{\beta}_j - \hat{\beta})^T X_j^T X_j (\hat{\beta}_j - \hat{\beta}) \quad (4)$$

Using the form of the least-square estimators $\hat{\beta}_i$, $\hat{\beta}_j$ and $\hat{\beta}$

$$\hat{\beta}_i - \hat{\beta} = \beta_i - \beta + \left\{ \left[ \left( X_i^T X_i \right)^{-1} X_i^T, 0 \right] - \left( X_i^T X_i + X_j^T X_j \right)^{-1} \left[ X_i^T, X_j^T \right] \right\} \cdot \begin{bmatrix} e_i \\ e_j \end{bmatrix}$$

$$\hat{\beta}_j - \hat{\beta} = \beta_j - \beta + \left\{ \left[ 0, \left( X_j^T X_j \right)^{-1} X_j^T \right] - \left( X_i^T X_i + X_j^T X_j \right)^{-1} \left[ X_i^T, X_j^T \right] \right\} \cdot \begin{bmatrix} e_i \\ e_j \end{bmatrix}$$

and (4) becomes

$$\Delta S_{w,k,k-1}^{i,j} = (\beta_i - \beta)^T X_i^T X_i (\beta_i - \beta) + (\beta_j - \beta)^T X_j^T X_j (\beta_j - \beta) +$$

$$+ (\beta_i - \beta)^T \left\{ X_i^T e_i - X_i^T X_i \left( X_i^T X_i + X_j^T X_j \right)^{-1} \left( X_i^T e_i + X_j^T e_j \right) \right\} +$$

$$+ \left\{ e_i^T X_i - \left( e_i^T X_i + e_j^T X_j \right) \left( X_i^T X_i + X_j^T X_j \right)^{-1} X_i^T X_i \right\} (\beta_i - \beta) +$$

$$+ (\beta_j - \beta)^T \left\{ X_j^T e_j - X_j^T X_j \left( X_i^T X_i + X_j^T X_j \right)^{-1} \left( X_i^T e_i + X_j^T e_j \right) \right\} +$$

$$+ \left\{ e_j^T X_j - \left( e_i^T X_i + e_j^T X_j \right) \left( X_i^T X_i + X_j^T X_j \right)^{-1} X_j^T X_j \right\} (\beta_j - \beta) +$$

$$+ e_i^T X_i \left( X_i^T X_i \right)^{-1} X_i^T e_i + e_j^T X_j \left( X_j^T X_j \right)^{-1} X_j^T e_j -$$

$$- \left( e_i^T X_i + e_j^T X_j \right) \left( X_i^T X_i + X_j^T X_j \right)^{-1} \left( X_i^T e_i + X_j^T e_j \right)$$

If classes $i$ and $j$ come from the same regime, that is $\beta_i = \beta_j = \beta$, the increase in the within sum of squared errors is only

$$\Delta S_{w,k,k-1}^{i,j} = e_i^T X_i \left( X_i^T X_i \right)^{-1} X_i^T e_i + e_j^T X_j \left( X_j^T X_j \right)^{-1} X_j^T e_j -$$

$$- \left( e_i^T X_i + e_j^T X_j \right) \left( X_i^T X_i + X_j^T X_j \right)^{-1} \left( X_i^T e_i + X_j^T e_j \right)$$

This quantity is very close to zero if the classes contain enough points. Thus, together with Remark 1, we obtain that if the joint classes are from the same regime, the within sum of squared errors should be close to zero and if the classes are from different regimes, the larger the difference between the parameters of the two regimes, the larger the increase in the within sum of squared errors should be.

## 2.3 The algorithm

Now, we can write the steps of the algorithm which, at the same time, classifies the data and models the dependencies within each class.

**Step 1 :** Decide upon the explanatory variables and choose the time lag $p$ using some information criterion

**Step 2 :** Build the data set to be used in the analysis by a sliding window of size $p$

**Step 3 :** Choose a lattice form for the Kohonen map - we have restricted ourselves to the rectangular grid case - and a convenient dimension and perform the self-organizing map algorithm. As we have mentioned earlier, the dimension of the map should be chosen as a compromise between the size of the sample and the number of regressors, that is to say one should have enough points in a cluster to estimate the regression hyperplane, but not too many, in order to avoid mixing regimes in one cluster. Of course, there is no theoretical answer to this problem and avoiding mixing becomes impossible when the regression hyperplanes are very close, but that leads to another interesting question: when they are close, does it make sense in practice to consider that the underlying model has two regimes? As a practical rule, the maps used in this paper were squared rectangles with $M^2$ neurons such that, in average, each cluster contain about five times more observations than the number of explanatory variables.

**Step 4 :** Classify hierarchically the resulting $M^2$ clusters

For $k$ going from $M^2$ to 2 :

**Step 4.1** Compute the $k$ regression hyperplanes within each class

**Step 4.2** Find $(i_0, j_0)$ which minimize $\Delta S^{i,j}_{w,k,k-1}$

**Step 4.3** Join together classes $i_0$ and $j_0$, put $k = k - 1$ and go to step 4.1.

At the last step all points are joined together and there is a single regression line. The next thing to do is draw the dendrogram and look how the within sum of squared errors increases over the classification. As mentioned earlier, one might expect an important break when the number of classes is smaller than the real number of regimes.

## 3 Examples and Results

The method was tested on several nonlinear autoregressive regime-switching models. Historically, the first introduced were the threshold models (we will not speak here about the other variants of these models, smoothed, etc.), followed by the Markov switching and next we will consider both examples. Concerning the software for the self-organizing maps we have used the tools developed at the SAMOS laboratory by P. Letremy(2000) in SAS language, while the hierarchical classification was written in R.

### 3.1 TAR Models

The example is a TAR of order two and the coefficients were taken from the paper of Gonzalo & Pitarkis(2002).

$$y_t = \begin{cases} -3 + 0.5y_{t-1} - 0.9y_{t-2} + \varepsilon_t & , y_{t-2} \leq 1.5 \\ 2 + 0.3y_{t-1} + 0.2y_{t-2} + \varepsilon_t & , y_{t-2} > 1.5 \end{cases},$$

where $\{y_t\}$ is the observed series and $\varepsilon_t$ are i.i.d.-standard gaussian.

Three samples containing 200, 400 and 800 points, respectively, were simulated. Let us examine the 200-point sample. The dimension of the self-organizing map was fixed equal to 5x5 and $\{y_t, y_{t-1}, y_{t-2}\}$ were the variables used for clustering. The results are displayed in Figure 2. The left part of the graphic contains the individuals, as clustered by the algorithm on the grid. We have chosen to show the corresponding curves for each data point in the sample - namely $\{y_t, y_{t-1}, y_{t-2}\}$- as a confirmation for the use of self-organizing maps : the clusters are homogeneous, similar profiles are joined together on the map and, moreover, due to the topology preservation, neighbouring clusters contain data with similar curves which are thus supposed to belong to the same underlying model.

On the right part, crossing the map with a boolean variable which distinguishes whether $y_{t-2}$ is above or below the threshold value and representing the pie charts in each cluster allows to see that all the points from one cluster belong to the same regime and that there is no crossing between regimes on the map, they are well separated topographically.

Then, when the hierarchical clustering algorithm is run on the twenty-five clusters, the squared-error criterion increases as in Figure 3 and suggests that a good choice would be a two-regime model. The hierarchical classification also provides good estimates for the parameters of the model when choosing two regimes as shown in Table 1.

The threshold value was not estimated and we will see that the decision on the type of the model (TAR, Markov switching) cannot be made on this basis only. For the 400 and the 800 samples, the results were very similar and in both cases the hierarchical algorithm suggested a two-regime model.

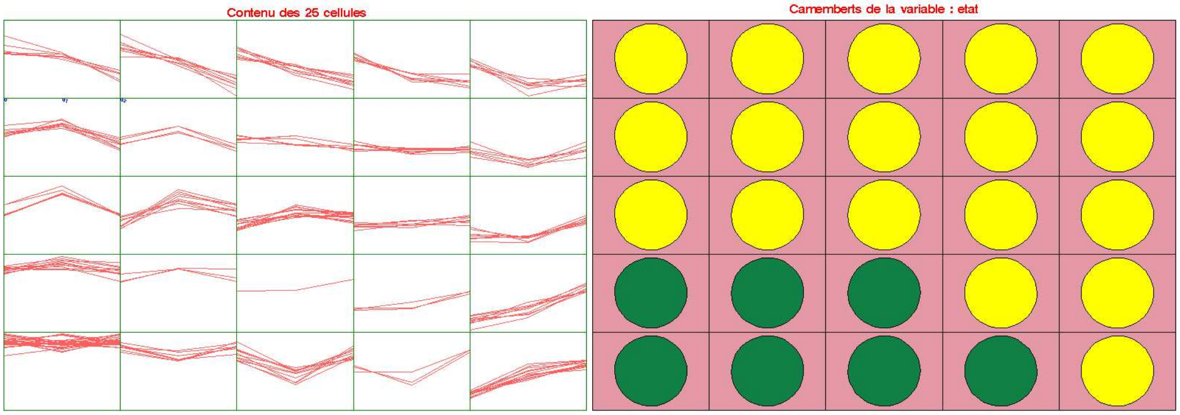

Figure 2: Initial clustering with a Kohonen map for TAR model

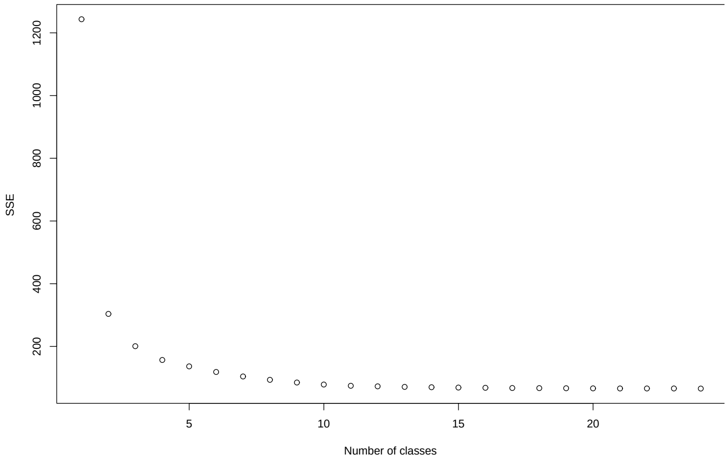

Figure 3: Squared-error criterion for TAR model

### 3.2 A Two-Regime Markov Switching Model

For this example, let us first define a two-regime autoregressive Markov-switching process of order $p$. If $\{y_t\}_{t\in\mathbb{N}}$ is the observed time series, there also exits a two-state Markov chain $\{x_t\}_{t\in\mathbb{N}}$ with transition probability matrix $A = \begin{pmatrix} p & 1-p \\ 1-q & q \end{pmatrix}$ which controls the evolution in time of $y_t$ as follows :

- If $x_t$ is in the first state, $y_t$ can be modelled by a regression function $f_1\left(y_{t-1},...,y_{t-p}\right)$

|  | Regime 1 |  |  | Regime 2 |  |  |
|---|---|---|---|---|---|---|
|  | *Intercept* | $y_{t-1}$ | $y_{t-2}$ | *Intercept* | $y_{t-1}$ | $y_{t-2}$ |
| value | -2.82 | 0.59 | -0.89 | 1.4 | 0.37 | 0.32 |
| t-value | -19.15 | 13.09 | 15.89 | 8.24 | 5.96 | 4.79 |

Table 1: Coefficients for the TAR model

and an independent gaussian noise of variance $\sigma_1$

- If $x_t$ is in the second state, $y_t$ can be modelled by a regression function $f_2\left(y_{t-1},...,y_{t-p}\right)$ and an independent gaussian noise of variance $\sigma_2$

In one line, this can be written as :

$$y_t = f_{x_t}\left(y_{t-1},...,y_{t-p}\right) + \sigma_{x_t}\varepsilon_t$$

The data used here were simulated with the parameters below (a globally stationary process was chosen):

$$\left\{\begin{array}{l} f_1\left(y_{t-1},y_{t-2}\right) = 0.2 + 0.5y_{t-1} + 0.1y_{t-2} \\ f_2\left(y_{t-1},y_{t-2}\right) = 0.3 + 0.9y_{t-1} - 0.1y_{t-2} \end{array}\right. , \left\{\begin{array}{l} \sigma_1 = 0.03 \\ \sigma_2 = 0.02 \end{array}\right. \text{ and } A = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}$$

As in the previous example, three samples of 200, 400 and 800 points were considered. We will only list the results for the 400 sample and remark that the outputs for the other two cases were very similar. The initial clustering was performed using a 6x6 Kohonen map and $\{y_t, y_{t-1}, y_{t-2}\}$ as variables. Figure 4 shows that the map is well-organized, the clusters are homogeneous and when crossing with the variable giving the regime, there is a good separation in the initial clusters, although there are some small overlappings of the two regimes in four of them. Then, from the hierarchical classification of these clusters, again we get a huge jump when passing from two classes to one, as shown in Figure 5. The estimated coefficients in each of the two classes are shown in Table 2 (we will also note that the two classes are homogeneous, the percentage of explained variance being larger than 92% in each of them).

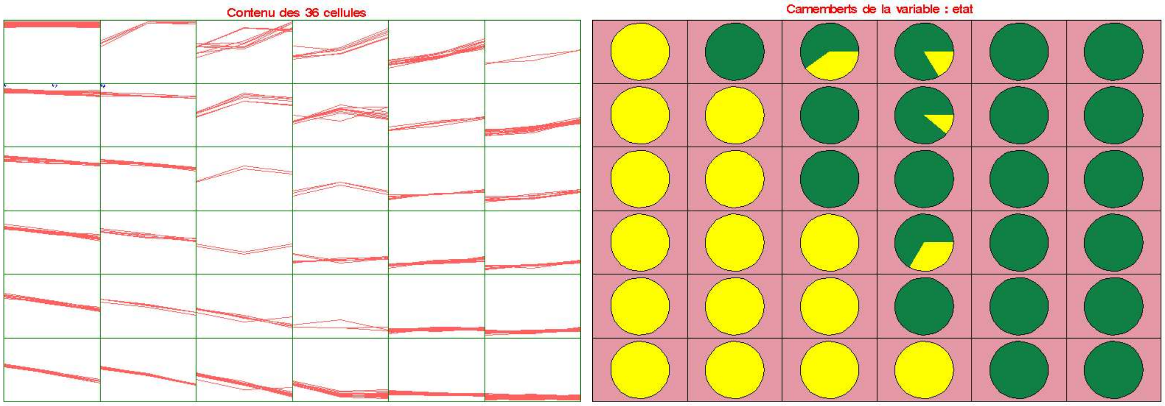

Figure 4: Initial clustering with a Kohonen map for two-regime Markov

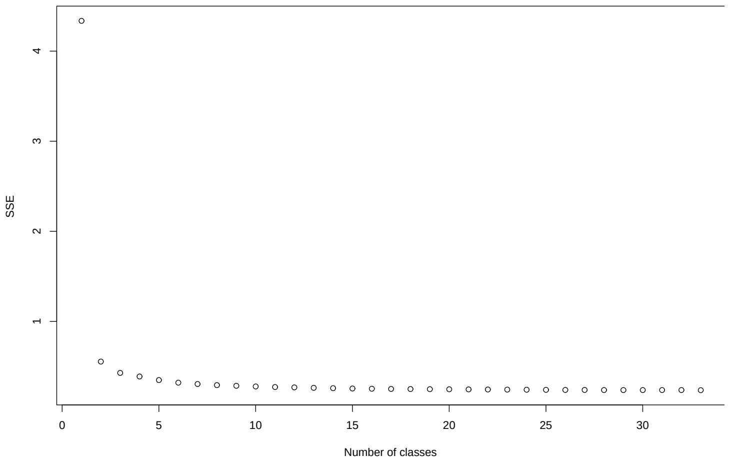

Figure 5: Squared-error criterion for two-regime Markov

| | Regime 1 | | | Regime 2 | | |
|---|---|---|---|---|---|---|
| | $Intercept$ | $y_{t-1}$ | $y_{t-2}$ | $Intercept$ | $y_{t-1}$ | $y_{t-2}$ |
| value | 0.3 | 0.88 | -0.09 | 0.18 | 0.39 | 0.24 |
| t-value | 41.56 | 32.89 | -3.64 | 20.91 | 12.07 | 8.87 |

Table 2: Coefficients for the two-regime Markov model

### 3.3 A Three-Regime Markov Switching Model

Now let us see what happens if we add a new regime to the model, which will moreover be explosive and drive the process into a nonstationary one. The following example was considered :

$y_t = f_{x_t}(y_{t-1}, y_{t-2}) + \sigma_{x_t}\varepsilon_t$ , $f_{x_t}(y_{t-1}, y_{t-2}) \in \{f_1, f_2, f_3\}$, $\sigma_{x_t} \in \{\sigma_1, \sigma_2, \sigma_3\}$, $\varepsilon_t$ is i.i.d. standard gaussian and

$$\begin{cases} f_1(y_{t-1}, y_{t-2}) = 0.2 + 0.5y_{t-1} + 0.1y_{t-2} \\ f_2(y_{t-1}, y_{t-2}) = 0.3 + 0.9y_{t-1} - 0.1y_{t-2} \\ f_3(y_{t-1}, y_{t-2}) = 0.5 + 1.2y_{t-1} + 0.5y_{t-2} \end{cases}, \begin{cases} \sigma_1 = 0.03 \\ \sigma_2 = 0.02 \\ \sigma_1 = 0.03 \end{cases} \text{and } A = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.6 & 0.2 & 0.2 \end{pmatrix}$$

The following results are from a 400-point sample, with an initial 8x8 map. By crossing the map with the regime variable, there is a relatively good separation, although we can notice that the first two regimes seem to come closer together compared to the third one.

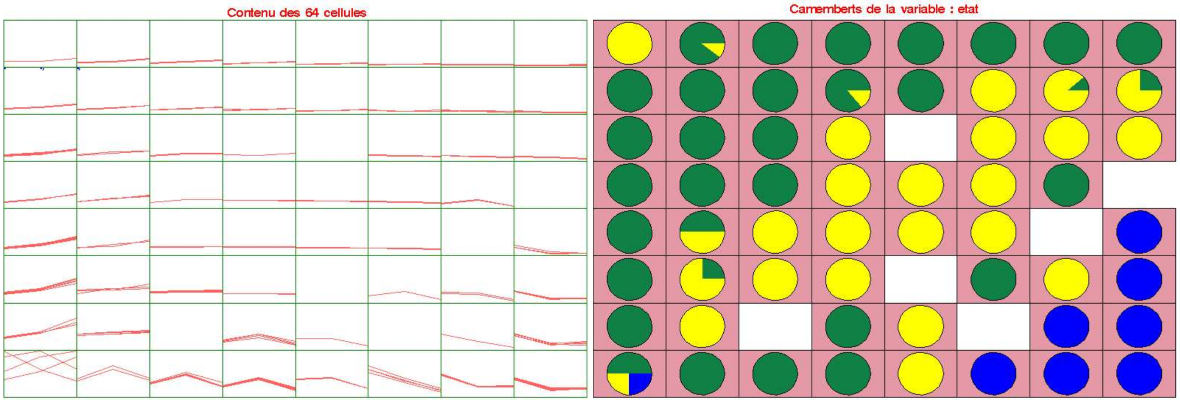

Figure 6: Initial clustering with a Kohonen map

Here, a first conclusion would be that there are four regimes. But let us take a closer look at the hierarchical classification. One of the four final classes contains only one cluster, one cell of the map. Moreover, this cell (bottom-left) is isolated from the rest of the map and it contains only four observations with very high values. If we project the data on a two or three-dimensional space, the same four observations are far from the rest. The algorithm has identified a small class of outliers which are considered as a separate regime and from this point of view the method is close to Ward's which is also sensitive to this kind of observations. We cannot continue with the examples without making an important remark. We have seen that this method of identifying the number of regression lines works quite good in the examples above. Concerning the parameter estimation and the choice of the model (TAR or

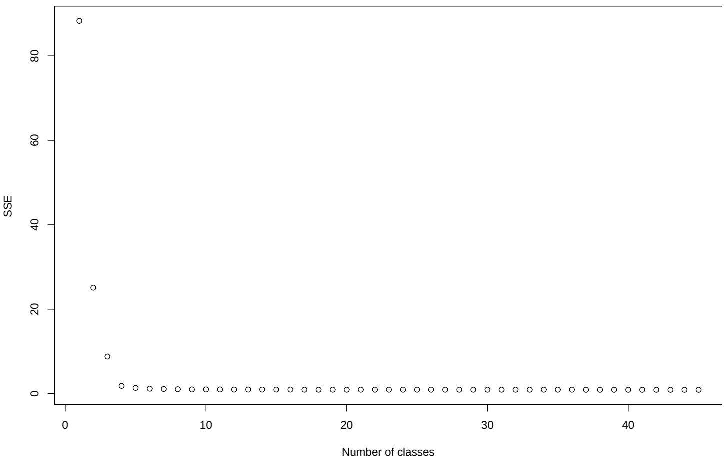

Figure 7: Squared-error criterion for three-regimes Markov

Markov switching?, for instance), the hierarchical classification only provides the estimates for the regression lines, a likelihood approach should be used instead, once we fixed the number of regimes, to estimate the rest of the parameters : threshold value, transition matrix etc. As for the second question, no theoretical result is available yet, econometricians preferring to rely on other criteria (economic, social etc).

## 3.4 What about Real Life Data?

The results of the simulated examples being encouraging, we decided to run the algorithm on real data sets. Three examples were chosen, the first two are the benchmarks Old Faithful Geyser Data and Santa Fe Competition Laser Data, and the third is the U.S. GNP (Gross National Product) series, used by Hamilton to introduce the switching Markov models.

### 3.4.1 Old Faithful Data

The first set of data is the classical Old Faithful Geyser in Yellowstone National Park, consisting of 299 pairs of measurements referring to the waiting time between two successive eruptions, $w_t$, and the duration of the subsequent eruption, $d_t$. The data were collected between August 1st and August 15th, 1985 and the two variables are recorded in minutes.

Several studies of this sample are available, most authors trying to assess either the clustering of the data, either the dependency between successive events. A literature overview, as well as an analysis using time series while "a priori" assuming the existence of two patterns of dependency, can be found in Azzalini and Bowman (1990) paper. Although there are no autoregressors in this case, the problem is the same : finding the number of clusters and fitting a regression within each one.

Our approach would be to detect the clustering of the data and, at the same time, model the dependency within each class. The idea of addressing clustering and dependency at the same time was also used by Hennig (2000) with regression fixed point clusters. Here, the duration of the eruption was modelled as a linear function of the waiting time before the eruption. When plotting the duration against the waiting time, one can see that there are at least two classes of points, depending on the waiting time. Let us make one last remark on the data, which is that, due to inexact observations during the night, there are 53 points with duration=4 (long eruption) and 20 with duration=2 (short eruption). The medium eruptions (duration=3) appear only once.

For the Kohonen clustering the data set $\{w_t, d_t\}_{t=\overline{1,299}}$ was considered (no time lags were introduced). The map was chosen to be a 6x6 grid and 1000 iterations were performed. Afterwards, a hierarchical classification minimizing the within sum of squared-error criterion was applied to the 33 "valid" clusters (one cluster was void and two others contained only one point).

In Figure 8, the within sum of squared errors as well as the difference $\Delta SSE_{k,k-1}$ are plotted. The heterogeneity of the data is obvious and one can see that at least two classes should be considered. On the contrary, passing from three classes to two is less obvious and we shall see next that there is overlapping of the classes and that the regression coefficients are very close.

Once we get the hierarchical classification, we go back to the self-organizing map to see how the classes spread over the grid. On the first graph in Figure 9, the tree-cut at two classes is presented : the first class in the right upper corner in light-grey circles and the second

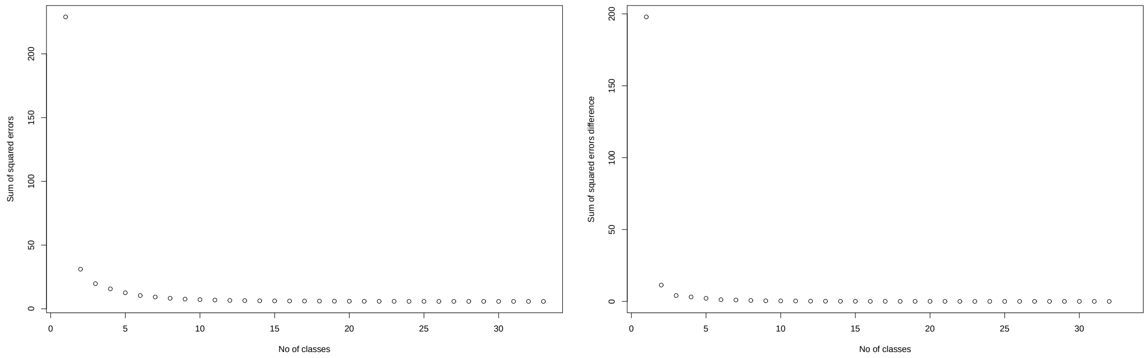

Figure 8: Squared-error criterion for the Old Faithful Data

beyond the diagonal of the grid in dark-grey circles. Clusters 7 and 28 do not contain enough points and were not considered in the classification.

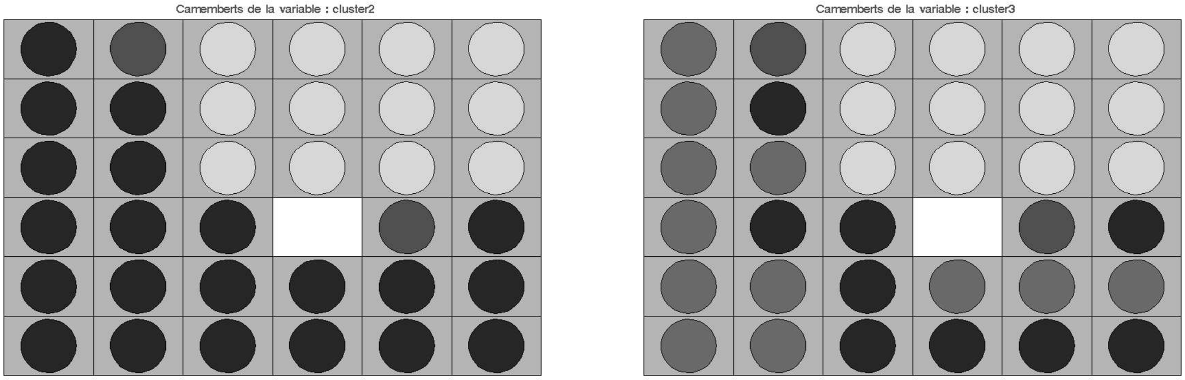

Figure 9: The self-organizing map for the Old Faithful Data

While the 2-classe model is meaningful, in the 3-classe case things seem to be more complicated : the first class is the same, well-isolated in the right upper part of the grid, while classes 2 and 3 are mixed on the grid. This is even more obvious if we plot the duration $d_t$ against the waiting time $w_t$ and identify the classes as shown in Figure 10.

The same situation as in the Kohonen map occurs. In both cases, 2-classe and 3-classe model, a first class (the right upper part of the self-organizing map) is well-isolated from the rest and is concentrated around the duration=2 points. This class corresponds to short-time eruptions preceded by rather long waiting times. The second class on the first graph contains all the

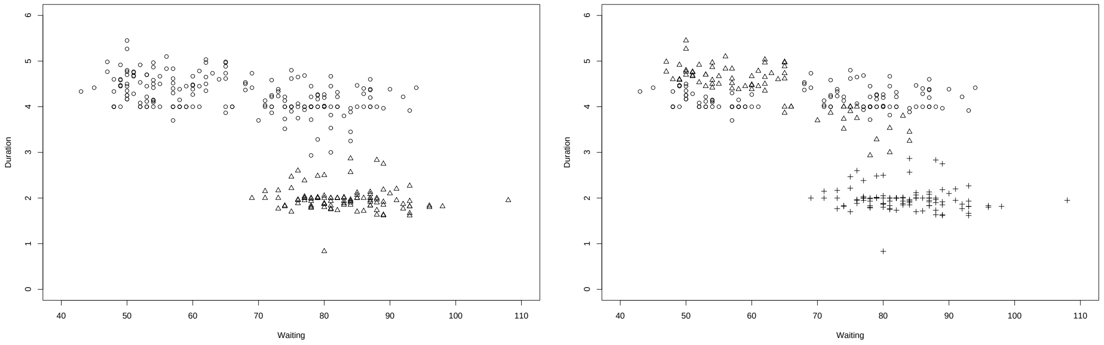

Figure 10: 2-classe and 3-classe of the Old Faithful Data

long-duration points including all duration=4 points. This class is splitted into two when considering the 3-classe case : one sub-class is formed by the duration=4 points and the long time eruptions preceded by long waiting times, while the second contains the data with a moderately decreasing tendency in the duration for increasing waiting times.

Although there is an interpretation for the 3-classe model, the within squared-error criterion used for the classification suggests a 2-classe model, because it corresponds to an important break in the increase of the squared error. Besides, in the discussion of Azzalini and Bowman (1990) geological evidence for the existence of two distinct patterns of eruptions is given and thus our conclusion is enhanced.

### 3.4.2 Santa Fe Competition Laser Series

For the laser series, a highly nonlinear data set, the algorithm selects three classes as shown in Figure 11. In order to compare the results with the existing literature, ten time lags were used. The 10,000 observations were initially clustered on a 9x9 map. Once the number of regimes was fixed, the three regimes were supposed to be the states of a discrete Markov chain and the parameters were estimated using the EM algorithm as described in Rynkiewicz (1999). The prediction results are weaker than those obtained by Weigend (1995) or Rynkiewicz (1999), but the number of estimated parameters is much smaller. But, as the nonlinearities of the series are not entirely explained by a mixture of linear regressions, an adaptation of the method by replacing the linear functions with nonlinear ones could be more interesting for this kind of data.

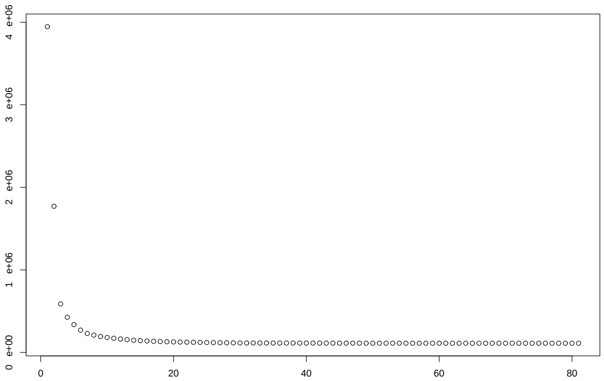

Figure 11: Squared-error criterion for the laser series

### 3.4.3 GNP Series

Concerning the GNP series, Hamilton's approach was based on the assumption that the mean growth rate is subject to occasional, discrete shifts. We have 136 trimestrial observed values of the series at our disposal, from 1952 until 1984. The maximal lag to be considered was determined with the AIC criterion and was fixed at 3. The data was initially clustered with a 4x4 map and considering $\{y_t, y_{t-1}, y_{t-2}, y_{t-3}\}$ as variables. The sixteen clusters were grouped hierarchically by the squared-error criterion and the results are displayed in Figure 12.

The first graph contains the within sum of squared errors plotted against the number of classes and the second its percentage increase. A first break appears when considering six classes instead of seven, but this can be interpreted as being due to possible strongly homogeneous clusters from the same regime which get mixed. There is a second break (13%) when passing from two classes to one but this is less obvious than in previous cases and the decision to model this series by a two-regime model is questionable from our point of view.

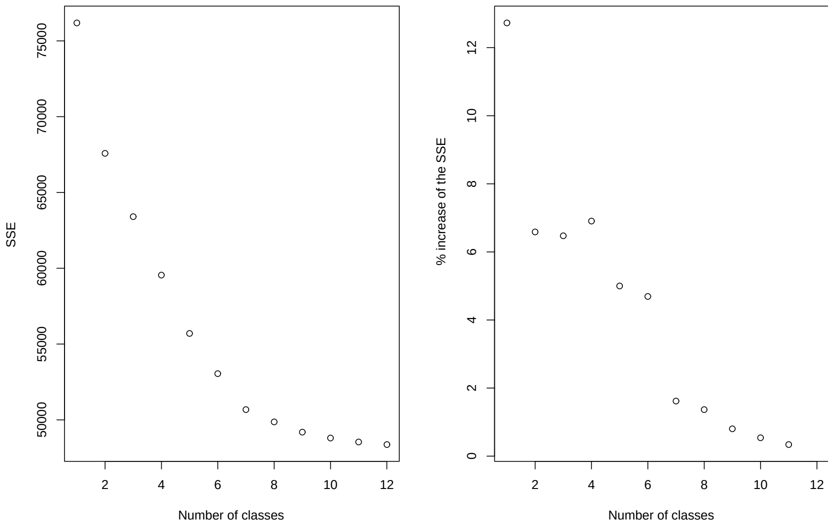

Figure 12: Squared-error criterion for GNP data

# 4 Conclusion and Future Work

We have introduced a descriptive method to assess the presence of regime changes in nonlinear time series analysis. As there is no theoretical answer and no statistical test to solve this problem for the moment, this method may be used, but with precaution. Indeed, self-organizing maps could mix the regimes if the regression hyperplanes are too close and the squared-error criterion seems to be sensitive to outliers.

Thus, several improvements should be made in the future, such as considering a different distance that would take into account the temporal dependency of the data for the initial clustering or looking for a smarter initial clustering that would avoid mixing regimes in the same cluster. Replacing the linear regressors by nonlinear functions could also be a possibility, but then the number of parameters would become larger and the liability of the estimates would decrease.

# References

[1] Azzalini A., Bowman A.W. (1990), A Look at Some Data on the Old Faithful Geyser, *Applied Statistics*, **vol. 39** p. 357-365.

[2] Charles C. (1977), Regression typologique, *Research Report No. 257, Le Chesnay : INRIA*

[3] Chow G.C. (1960), Tests of Equality Between Sets of Coefficients in Two Linear Regressions, *Econometrica*, **vol. 28** p. 591-605.

[4] DeSarbo W.S., Cron W.L. (1988), A maximum likelihood methodology for clusterwise linear regression, *Journal of Classification*, **vol. 5** p. 249-282.

[5] Gonzalo J., Pitarakis J-Y. (2002), Estimation and Model Selection Based Inference in Single and Multiple Treshold Models, *Journal of Econometrics*, **vol. 110** p. 319-352.

[6] Hamilton J.D. (1989), New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle, *Econometrica*, **vol. 57** p. 357-384.

[7] Hennig C. (2000), Regression Fixed Point Clusters : Motivation, Consistency and Simulations, *Preprint 2000-02, Fachbereich Mathematik, Universitat Hamburg*

[8] Kohonen T. (1997), *Self-Organizing Maps*, New-York, Springer-Verlag.

[9] Letremy P. (2000), Notice d'installation et d'utilisation de programmes bases sur l'algorithme de Kohonen et dedies a l'analyse des donnees, *Prepub. Samos 131.*

[10] Lou S., Jiang J., Keng K.(1993), Clustering Objects Generated by Linear Regression Models, *Journal of the Amarican Statistical Association*, **vol. 88** p. 1356-1362.

[11] Rynkiewicz J.(1999), Hybrid HMM/MLP Models for Time Series Prediction, *ESANN'1999 Proceedings*, p. 455-462.

[12] Spath H. (1979), Clusterwise linear regression, *Computing*, **vol. 22** p. 367-373.

[13] Tong H. (1978), On a threshold model, *Pattern Recognition and Signal Processing*, ed C.H. Chen, Amsterdam : Sijhoff&Noordhoff.

[14] Toyoda T. (1974), Use of the Chow Test under Heteroscedasticity, *Econometrica*, **vol. 42** p. 601-608.

[15] Weigend A.S., Mangeas M., Srivastava A.N.(1995), Nonlinear gated experts for time series : discovering regimes and avoiding overfitting, *International Journal of Neural Systems*, **vol. 6** p. 373-399.