Common Trends in European School Populations

P. Sebastiani[1] *(1) and M. Ramoni (2)*

(1) Department of Mathematics and Statistics, University of Massachusetts.
(2) Children's Hospital Informatics Program, Harvard University Medical School.

Keywords: Autoregressive models, model-based clustering, Bayesian model selection, temporal data.

Abstract.

This paper uses a novel Bayesian clustering method to categorize the temporal evolution of the share of population participating in tertiary/higher education in 14 European nations. The method represents time series as autoregressive models and applies an agglomerative clustering procedure to discover the most probable set of clusters describing the essential dynamics of these time series. To increase efficiency, the method uses a distance-based heuristic search strategy. This clustering method partitions the evolution of school population into three groups, thus revealing significant differences among tertiary/higher education in the 14 European nations.

1. Introduction

The time series in Figure 1 describe the evolution of the share of population enrolled in higher education in 14 nations of the European community between 1970 and 1995. Our task is to group the 14 time series on the basis of their similarity in order to detect significant differences among European higher education trends. Data were provided by UNESCO and Eurostat, via the r·cade data bank, available at the URL http://www-rcade.dur.ac.uk, (Unesco, 1997).

The method to solve this problem depends on the meaning we attach to similar time series. Throughout this paper, we will assume that time series are the realization of stochastic processes and two or more time series are similar when the same process generates them. Thus, deciding whether two time series are similar is equivalent to deciding whether they are observations of the same process. Put in this way, the task of grouping of the time series can be solved as a clustering problem: given a batch of time series, we wish to cluster them so that each cluster contains time series generated by the same process. Particularly, we wish to solve this problem without specifying, a priori, the number of clusters. We solve this problem by using a novel Bayesian for method clustering of contributions.

[1] Address for correspondence: Department of Mathematics and Statistics, Lederle Graduate Research Tower, University of Massachusetts, Amherst MA 01003. Email: sebas@math.umass.edu, Telephone: 413 545 0622.

We model the stochastic process generating each time series as an autoregressive model of order p, say AR(*p*), and then we cluster those time series that have a high posterior probability of being generated by the same AR(*p*) model. The distinguished feature of this method is to describe a clustering of time series as a statistical model so that the clustering task can be solved as a Bayesian model selection problem. Thus, the clustering model we look for is the most likely partition of the time series, given the data at hand and prior information about the problem.

In principle, we just need to evaluate the posterior probability of all possible clustering models of time series and select that one with maximum posterior probability. However, the number of clustering models grows exponentially with the number of time series and a heuristic search is needed to make the method feasible. The method we adopt uses a measure of similarity between AR(p) models to drive the search process in a subspace of all possible clustering models. An important feature of this heuristic search is to provide a stopping rule, so that clustering can be done without assuming a given number of clusters as traditional clustering methods do.

The clustering method we use is fully described and evaluated in Sebastiani and Ramoni (2001). In the next section we briefly describe the method and the search algorithm. The analysis of the higher education data set is described in Section 3 and a discussion is in Section 4.

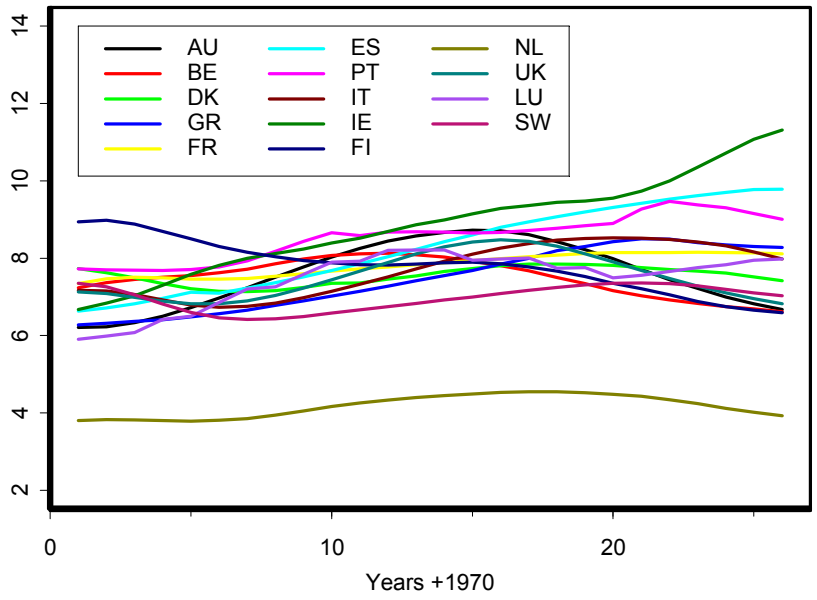

Figure 1. Share of population enrolled in higher education, between 1970 and 1996, in the14 European countries: AU: Austria; BE: Belgium; DK: Denmark; GR: Greece; FR: France; ES: Spain; PT: Portugal; IT: Italy; IE: Ireland; FI: Finland; NL: The Netherland; UK: United Kingdom; Lu: Luxemburg; SW: Sweden.

2. Bayesian clustering by dynamics

The clustering method we describe here has three components: a model for the time series, the posterior probability of a clustering model and a heuristic search strategy. These three elements are described very briefly. More details are provided in Sebastiani and Ramoni (2001).

2.1. Autoregressive models

Let $S = \{y_{-p}, \cdots, y_{-1}, y_1, \cdots, y_t, \cdots, y_n\}$ be a time series of values observed for a continuous variable Y. The series follows an AR(p) model if

$$y \mid \beta = X\beta + \varepsilon$$

where y is the n-dimension vector $y = (y_1, \cdots, y_n)$, X is the $n \times n$ matrix with tth row given by the vector of p observations $y_{t-1}, \cdots, y_{t-p}$, $\beta = (\beta_1, \cdots, \beta_p)$ is a vector of autoregressive coefficients, and ε is a vector of uncorrelated errors. We assume that the errors are normally distributed, with expectation $E(\varepsilon_t) = 0$, and variance $V(\varepsilon_t) = \sigma^2$ for any t. We shall denote by τ the precision, so that $\sigma^2 = 1/\tau$.

The value p is called the order of the autoregression, and specifies the Markov order of the series: namely that $y_t \perp (y_{-p}, \cdots, y_{t-p-1}) \mid (y_{t-1}, \cdots, y_{t-p})$, where we use the symbol $\perp$ to denote stochastic independence. The series follows a stationary process if the roots of the polynomial $f(u) = 1 - \sum_{j=1}^{p} \beta_j u^j$ have moduli greater than unity. When some of the roots have moduli smaller than unity, the process is non-stationary, but typically some transformations of the data are sufficient to achieve stationarity.

The model above describes the evolution of the process around a zero mean. By adding an intercept term β_0, the model can be extended to include a non-zero mean μ, for each y_t, so that $\beta = (\beta_0, \beta_1, \cdots, \beta_p)$ and the matrix X is augmented of a column of ones. The process mean and the autoregressive coefficients are related by: $\mu = \beta_0 / (1 - \sum_{j=1}^{p} \beta_j)$.

We wish to compute Bayesian estimates of the parameters β and τ. To compute the Bayesian estimates of β and τ, we need to up-date their joint prior density $f(\beta, \tau)$ into the posterior density $f(\beta, \tau \mid y)$, by using Bayes' Theorem:

$$f(\beta, \tau \mid y) = \frac{f(\beta, \tau) f(y \mid \beta, \tau)}{f(y)},$$

Where $f(y \mid \beta, \tau)$ is the likelihood function and $f(y)$ is the marginal likelihood, which is computed as

$$f(y) = \int f(\beta, \tau) f(y \mid \beta, \tau) d\beta d\tau .$$

For a given autoregressive order p, we compute the likelihood function, conditional on the first p values of the time series, as

$$f(y \mid \beta, \tau) = \sqrt{\frac{\tau^n}{(2\pi)^n}} \exp\left(-\frac{\tau (y - X\beta)^T (y - X\beta)}{2} \right).$$

We assume as prior density for β and τ the improper prior $f(\boldsymbol{\beta, \tau}) \propto X\tau^{-2}$, with $\tau > 0$ (see Jeffreys, 1946). Suppose the matrix X is of full rank, and let $\hat{\beta}$ and RSS denote respectively the ordinary least squares estimate of β and the residual sum of squares respectively:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$RSS = y^T (I_n - X(X^T X)^{-1} X^T) y$$

where I_n is the identity matrix. Then, one can show that the marginal likelihood is

$$f(y) = \frac{\left(\frac{RSS}{2} \right)^{q+2-n} \Gamma\left(\frac{n-q-2}{2} \right)}{(2\pi)^{n-q} \det(X^T X)^{1/2}}$$

where q is the dimension of the vector β. Furthermore, the posterior distribution of β and τ, is normal-gamma, with

$$\beta \mid y, \tau \sim N(\hat{\beta}, (\tau (X^T X))^{-1})$$

$$\tau \mid y \sim Gamma\left(\frac{RSS}{2}, \frac{n-q-2}{2} \right),$$

where *Gamma*(*a,b*) is a gamma distribution with expected value *a/b* and variance a/b^2.

Both distributions are proper whenever as the matrix X is of full rank, and $n > q + 2$. The Bayesian posterior point estimates of β and τ are $\propto$ and $(n - q - 2) / RSS$.

2.2. Clustering

Suppose we have a batch of time series $S = \{S_1, \cdots, S_m\}$, which are generated by an unknown number of stationary AR(p) models with a common autoregressive order p, and different autoregressive coefficients. We wish to cluster the time series in S according to their dynamics. Our goal is two-fold:

- To find the set of clusters that gives the best partition of the data;
- To assign each time series to one and only one cluster.

Contrary to common practice, we do not want to specify, a priori, a preset number of clusters.

Formally, the clustering method regards a partition as an unobserved discrete variable C with states $C_1, \cdots, C_c$. Each state C_k of the variable C labels, in the same way, the time series generated by the same AR(p) model and, hence, it represents a cluster of time series. The number c of states of the variable C is unknown, but it is bounded above by the total number of time series in the data set S. The clustering algorithm tries to re-label those time series that are likely to have been generated by the same AR(p) model and thus merges the initial states $C_1, \cdots, C_m$ of the variable C into a subset $C_1, \cdots, C_c$, with $c < m$.

The specification of the number c of states of the variable C and the assignment of one of its states to each time series S_i define a statistical model M_c. This allows us to regard the clustering task as a Bayesian model selection problem, in which the model we seek is the most probable way of re-labeling time series, given the data. If $P(M_c)$ is the prior probability of each model M_c, by Bayes' Theorem its posterior probability is $P(M_c \mid S) \propto P(M_c) f(S \mid M_c)$, where $f(S \mid M_c)$ is the marginal likelihood, now written as explicit function of the clustering model. A model-based Bayesian solution to the clustering problem consists of selecting the clustering model with maximum posterior probability. It is shown in Sebastiani and Ramoni (2001) that, under some assumptions on the sample space, the adoption of a particular parameterization for the model M_c and the specification of an improper-uniform prior lead to a simple, closed-form expression for the marginal likelihood $f(S \mid M_c)$.

Conditional on the model M_c and, hence, on a specification of the number of states of the variable C and of the labeling of the original time series, we suppose that the marginal distribution of the variable C is multinomial, with cell probabilities $\theta_k = P(C = C_k \mid \theta)$. Furthermore, we suppose that, conditional on $C = C_k$, the batch of m_k time series $\{S_{kj}\}$ assigned to cluster C_k are independent of the batch of time series $\{S_{lj}\}$ assigned to any other cluster C_l, and that the time series generated by the same AR(p) model in cluster C_k are mutually independent. We denote by β_k the vector of auto-regression coefficients and by τ_k the precision of the AR(p) model generating the time series in cluster C_k. We suppose that each of these series can be represented as

$$y_{kj} \mid \beta_k, \tau_k = X_{kj}\beta_k + \varepsilon_{kj}.$$

The index k indicates cluster membership, and ε_{kj} is a vector of uncorrelated errors, which we assume to be normally distributed, with $E(\varepsilon_{kjt}) = 0$ and $V(\varepsilon_{kjt}) = \tau_k^{-1}$, for any t. The fact that series assigned to the same cluster C_k are characterized by the same vector of auto-regression coefficients β_k, and by the same variance $\sigma^2{}_k = \tau_k^{-1}$, allows us to represent the whole batch of series $\{S_{kj}\}$ in cluster C_k as

$$y_k \mid \beta_k, \tau_k = X_k\beta_k + \varepsilon_k$$

where the vector y_k and the matrix X_k are defined as

$$y_k = \begin{pmatrix} y_{k1} \\ \vdots \\ y_{km_k} \end{pmatrix}$$

$$X_k = \begin{pmatrix} X_{k1} \\ \vdots \\ X_{km_k} \end{pmatrix}$$

Let β denote the set of parameter vectors $\beta = (\beta_k)$, where each β_k is a random vector, and let τ denote the set of parameters $\tau = (\tau_k)$, for $k = 1, \cdots, c$. Then, by the independence of series assigned to different clusters, the overall likelihood function is

$$f(y \mid \theta, \beta, \tau) = \prod_{k=1}^{c} \theta_k^{m_k} f(y_k \mid X_k, \beta_k, \tau_k)$$

where m_k is the number of time series that are assigned to cluster C_k. Here, the overall likelihood is conditional on the set of $c(p+2)$ values upon which the likelihood function of each series is conditioned.

We take as our prior distribution for θ a Dirichlet $D(\alpha_1, \cdots, \alpha_c)$, and assign the improper prior with density $f(\beta, \tau) \propto \prod_k \tau_k^{-2}$ to β and τ. Then, using standard results on Dirichlet integration, it is easy to show that the marginal likelihood is

$$f(y \mid M_c) = \int f(y \mid \theta, \beta, \tau) f(\theta, \beta, \tau) d\theta d\beta d\tau$$
$$= \frac{\Gamma(\alpha)}{\Gamma(\alpha + m)} \prod_{k=1}^{c} \frac{\Gamma(\alpha_k + m_k)}{\Gamma(\alpha_k)} \frac{(RSS_k / 2)^{q+2-n_k} \Gamma((n_k - q - 2)/2)}{(2\pi)^{(n_k - q)/2} \det(X^T{}_k X_k)^{1/2}}$$

where $\alpha = \sum_k \alpha_k$ is the overall cluster prior precision, n_k is the dimension of the vector y_k, and $RSS_k = y^T{}_k (I_n - X_k (X^T{}_k X_k)^{-1} X^T{}_k) y_k$ is the residual sum of squares in cluster C_k. The marginal likelihood is well defined as long as each matrix X_k is of full rank.

Once the most likely partition has been selected *a posteriori*, each cluster C_k is associated with theparameters β_k, which model the auto-regression equation, and the precision τ_k. The posterior distribution of $\beta_k \mid \tau_k, y_k$ is $N(\hat{\beta}_k, (\tau_k (X^T{}_k X_k))^{-1})$ with $\hat{\beta}_k = (X^T{}_k X_k)^{-1} X^T{}_k y_k$, while the posterior distribution of $\tau_k \mid y_k$ is $Gamma((RSS_k / 2), (n_k - q - 2)/2)$. The marginal posterior distribution of the auto-regression coefficients $\beta_k \mid y_k$ is a non-central Student's *t*, with expectation $\hat{\beta}_k$, which provides a point-estimate of β_k. The estimate of the within cluster precision τ_k is $((n_k - q - 2) / RSS_k)$. The probability that $C = C_k$ is estimated by $\hat{\theta}_k = (\alpha_k + m_k)/(\alpha + m)$.

In our application, we use a symmetric prior distribution for the parameter vector θ, with a common prior precision α. The initial m hyper-parameters α_k are set equal to α / m and, when two time series are assigned to the same cluster C_k, their hyper-parameters are summed up. Thus, the hyper-parameters of a cluster merging m_k time series will be $m_k(\alpha / m)$. In this way, the specification of the prior hyper-parameters requires only the global prior precision α, which measures the confidence in the prior model. The current implementation of the algorithm assumes that the series follow stationary autoregressive models of a given order *p*, and then checks that the stationarity conditions are met at the end of the clustering process.

2.3. Search

In principle, the clustering method described in the previous section requires one to compute the posterior probability of each clustering model and then choose the clustering model with maximum posterior probability. Since the number of possible partitions grows exponentially with the number of series, a heuristic method is required to make the search feasible.

Our method uses a measure of similarity between AR(*p*) models to efficiently guide the search process in a subset of all possible clustering models. Since all AR(*p*) models have the same order, this similarity measure is an estimate of the symmetric Kullback-Liebler divergence (Jeffreys, 1946) between marginal posterior distributions of the auto-regressive coefficients $\beta_k \mid y_k$ associated with the clusters. The estimate is given by computing the symmetric Kullback-Liebler

divergence for every pair of parameters β_k, β_j, assuming a normal distribution conditional on the within-cluster precisions τ_k, τ_j. The precisions are then replaced by their posterior estimates.

Initially, the algorithm transforms the time series in S into a set of m AR(p) models, using the procedure described in the previous section, and computes the set of $m(m-1)/2$ pair-wise distances between posterior distributions of the parameters. Then, the algorithm sorts the generated distances, labels in the same way the two closest AR(p) models and evaluates whether the resulting clustering model M_c, in which the two closest AR(p) models are assigned to the same cluster, is more probable than the model M_s in which they are distinct. If the probability $P(M_c \mid y)$ is larger than $P(M_s \mid y)$, the algorithm updates the set of series by replacing the two series with the cluster resulting from their merging. Consequently, the algorithm updates the set of ordered distances by removing all the ordered pairs involving the merged time series, and by adding the distances between the parameters of the new AR(p) model and the remaining models in the set. The procedure is then iterated on the new set. If the probability $P(M_c \mid y)$ is not larger than $P(M_s \mid y)$, the algorithm tries to merge the second best, the third best, and so on, until the set of pairs is empty and, in this case, returns the most probable partition found thus far. The rationale behind this heuristic is that merging closest AR(p) models first should speed up the search for clustering models with large posterior probability. Empirical evaluations of the methods on simulated data appear to support this intuition (see Sebastiani and Ramoni, 2001).

3. Analysis

We apply the clustering algorithm described in section 2 to the analysis of the fourteen time series reporting the temporal evolution of the share of the population engaged in tertiary/higher education in 14 European countries depicted in Figure 1. Since the average length of a university degree across European nations is three-four years, we applied the clustering algorithm under the assumption that all time series were generated by stationary AR(3) models with a non-zero mean. We assumed $\alpha = 1$, the improper prior with density $f(\beta, \tau) \propto \prod_k \tau_k^{-2}$, and uniform prior on all clustering models. Stationarity of the autoregressive models was checked at the end of the clustering process. Figure 2, 3 and 4 show the three clusters of time series found by the algorithm.

Cluster C_1 groups the evolutions of the proportion of the population enrolled in higher education institutes in Portugal and Luxembourg, see Figure 2. The estimates of the auto-regression coefficients are $\hat{\beta}_0 \cong 0.657$, $\hat{\beta}_1 \cong 1.133$, $\hat{\beta}_2 \cong 0.044$ and $\hat{\beta}_3 \cong -0.254$. Thus, the model is stationary --- the roots of the polynomial $f(u)$ are -2.38, $1.28 \pm 0.11i$ --- with a mean $\hat{\mu} \cong 8.532\%$. Note that the time series describing the evolution of school population in Luxemburg has a slight increasing trend during the 1970s, and then becomes stationary, with a mean slightly above 8%.

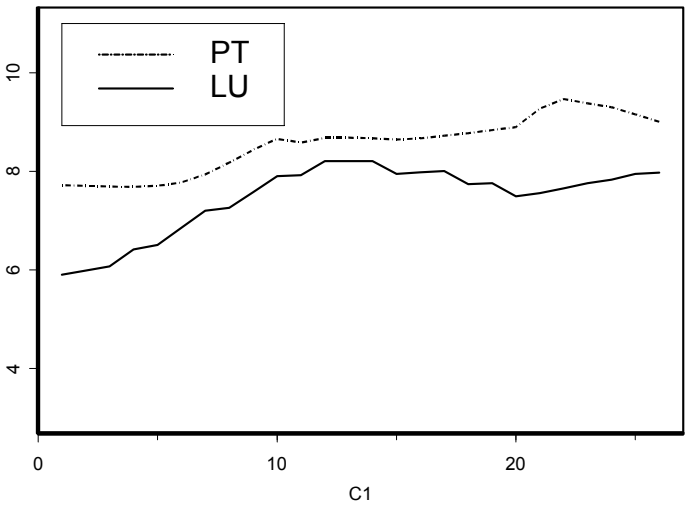

Figure 2. Cluster C_1 groups the evolution of school population in Portugal and Luxemburg.

The evolutions of the proportion of the population enrolled in higher education institutes in Austria, Denmark, Greece, Spain and Ireland are merged into cluster C_2 in Figure 3. The estimates of the auto-regression coefficients are $\hat{\beta}_0 \cong 0.074$, $\hat{\beta}_1 \cong 2.085$, $\hat{\beta}_2 \cong -1.233$ and $\hat{\beta}_3 \cong 0.138$, with a mean $\hat{\mu} \cong 7.4$. The AR(3) model is stationary, with roots of the polynomial $f(u)$ equal to 6.09 and 1.02±0.1*i*.

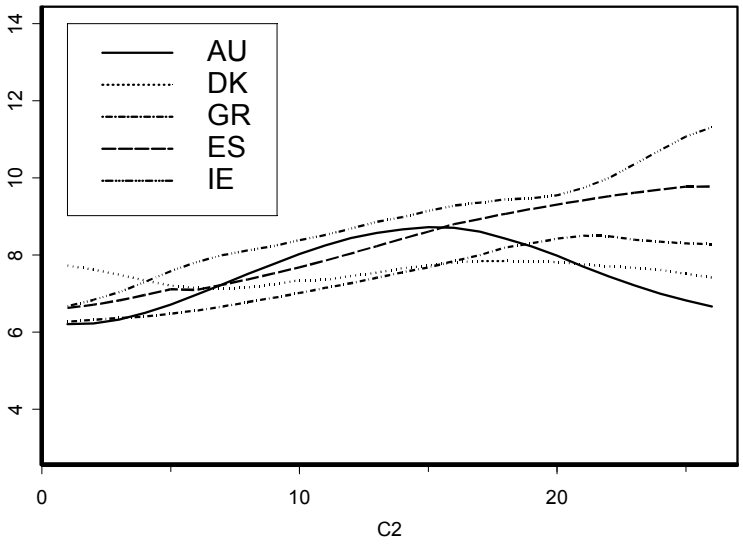

Figure 3 Cluster of time series describing the evolution of school population in Austria, Denmark, Greece, Spain and Ireland.

Of the series assigned to this cluster, those describing the evolution of school population in Austria, Denmark and Greece are evidently stationary, while the time series describing the evolution of school population in Spain and, particularly, Ireland exhibit some trend. The assignment of the two series to this cluster could indicate that the increasing trend is only temporary, and that the proportion of the population enrolled in higher education institutes becomes stable during the 1990s.

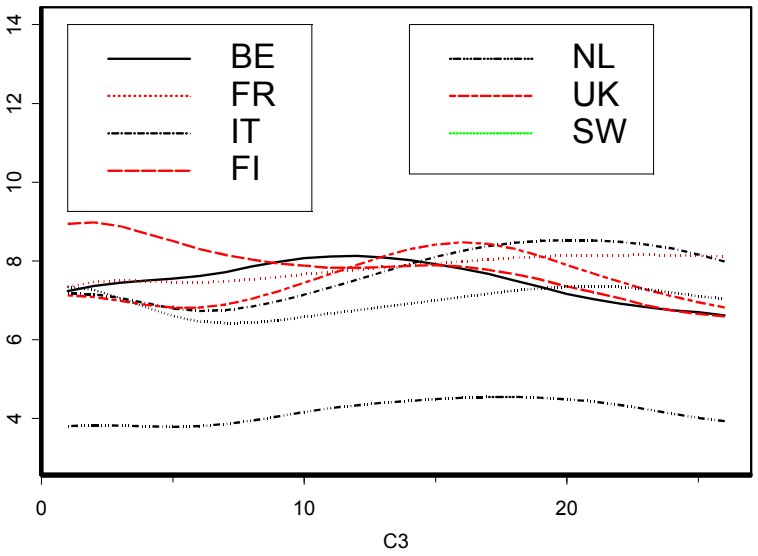

Figure 4 Cluster merging the evolution of school population in Belgium, France, Italy, Finland, The Netherlands, the United Kingdom and Sweden.

Cluster C_3 in Figure 4 groups the evolutions of the proportion of the population enrolled in higher education institutes in Belgium, France, Italy, The Netherlands, Finland, United Kingdom and Sweden. The estimates of the auto-regression coefficients are $\hat{\beta}_0 \cong 0.015$, $\hat{\beta}_1 \cong 2.593$, $\hat{\beta}_2 \cong -2.283$, and $\hat{\beta}_3 \cong 0.688$, thus defining a stationary auto-regression equation, with roots of the polynomial $f(u)$ equal to 1.023 and 1. $\pm 0.32i$. The mean of the process is $\hat{\mu} \cong 7.5$.

This cluster groups the European nations that have been consistently stronger from an economic point of view in the past thirty years. All these nations have a solid higher education tradition, and university curricula lasting, on the average, four years. All series assigned to this cluster are increasing up to the 1980s, and then decrease. This fact would be consistent with the large demand for highly skilled labors and for higher education created by the pace of economic development in Europe in the 1960s. The contraction of the population together with the economic recession in the 1980s, could be responsible for the decrease of the proportion of population enrolled in higher education in the late 1980s and the 1990s.

The means of the processes generating the time series assigned to clusters C_2 and C_3 are essentially the same. However, the autoregressive equation for cluster C_3 describes a more stable process around the mean, with smaller fluctuations. Thus, the results would suggest a more stable higher education enrollment in Italy, France, The Netherlands, United Kingdom, Belgium, Finland and Sweden, compared to Austria, Denmark, Greece, Spain and Ireland.

The fact that the time series describing the evolution of the population in higher education of The Netherlands is assigned to the third cluster is slightly disappointing: the dynamic of this series is similar to that of the other series in the cluster, but this series has a different mean. To evaluate the influence of this time series on the results, we run the clustering algorithm excluding the time series of The Netherlands. The algorithm found the same three clusters, thus showing that this series is not "influential".

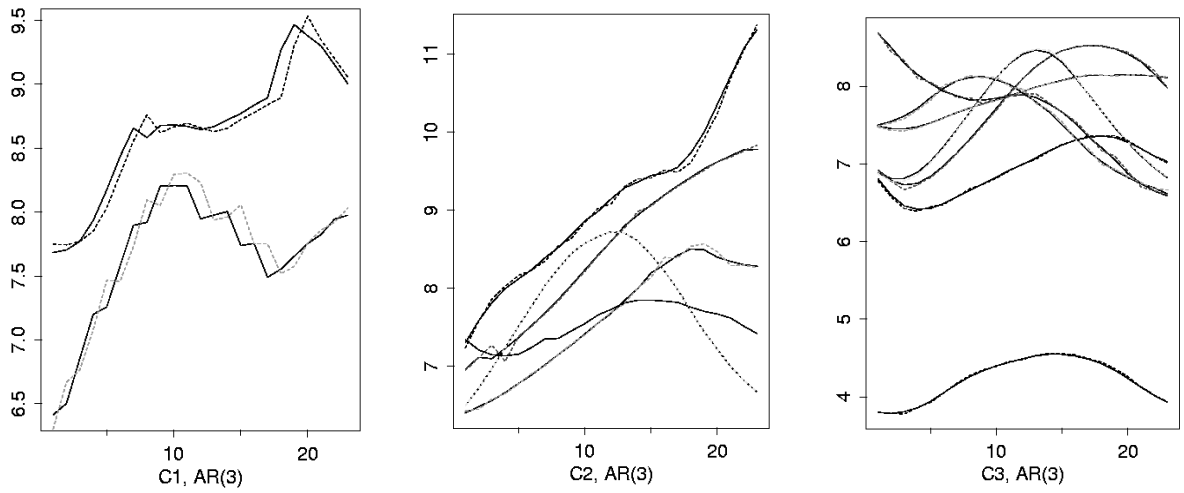

Figure5 .Observed (continuous line) and fitted (dash line) time series in the clusters in Figure 2.

During the analysis we assumed the time series were generated by AR(3) models. Plots of the observed and fitted values within clusters provide an overall assessment of the robustness of the result with respect to this assumption. Figure 5 plots the time series of observed values in the three clusters and values fitted using the AR(3) models associated with each cluster. The close match supports the assumption that AR(3) models are a good approximation of the processes generating the original fourteen series.

Finally, we note that the search algorithm found the three clusters of time series in just eighteen steps. This number is much smaller than the total number of clusters to be considered without the heuristic search. Figure 6 shows the increase of the log-marginal likelihood --- up to a constant --- at each step of the agglomerative search procedure. In the first seven steps, there is a linear increase of the marginal likelihood. Thus, merging the time series that belong to the clusters with nearest

autoregressive coefficients increases the marginal likelihood. In the next eight steps, merging the closest clusters does not always increase the marginal likelihood, so that the merging of the "second best" is evaluated and accepted. This is so until step 15, when the algorithm has merged the fourteen time series into three clusters. At this point, the three possible merging of two clusters at a time are evaluated and, since they all result in a decrease of the marginal likelihood, the algorithm stops and returns the three clusters so found.

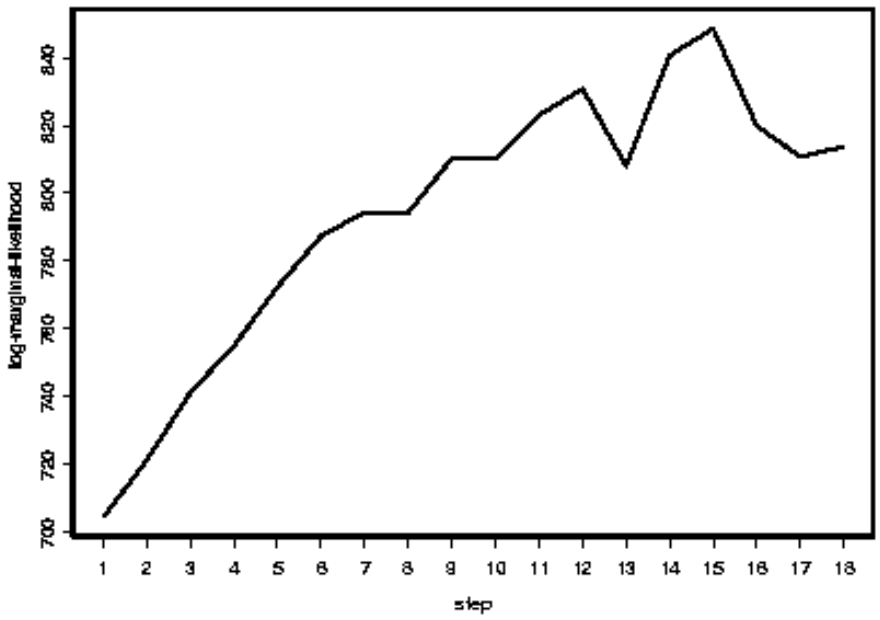

Figure 6.Change of the marginal likelihood, in logarithmic scale, at each step of the agglomerative search procedure.

4. Discussion and related work

Auto-regressive models have received great attention, (see Box and Jenkins, 1976, for a systematic exposition and West and Harrison, 1997, for a Bayesian analysis). Bayesian model-based clustering was originally proposed by Banfield and Raftery, (1993), to cluster static data. Ramoni et al., (2000, 2001) proposed a Bayesian clustering by dynamics algorithm, called BCD, to cluster discrete time series. BCD clusters time series modeled as Markov chains and, contrary to popular methods, finds also the number of clusters. Notwithstanding the, somewhat restrictive, Markov chain assumption, BCD has been applied successfully to cluster robot experiences based on sensory inputs (see Sebastiani et al., 2001), simulated war games (Sebastiani et al., 1999), as well as the behavior of stocks in the financial market and automated learning and generation of Bach's counterpoint.

Unlike BCD, the algorithm used in this paper clusters time series of continuous variables. The different type of data requires different modeling assumptions thus producing an algorithm which is

similar to BCD, in being Bayesian and model-based, but its methodology is novel. The heuristic search used by the clustering method is similar to that implemented in BCD although, here, the search is driven by a distance between posterior distributions of parameters characterizing the AR(p) models of different clusters, while in BCD the search uses the distance between predictive distributions of estimated Markov chains.

The model selection strategy of our algorithm seeks the clustering model with maximum posterior probability. Other choices here would be possible such as selecting the median posterior probability model (Barbieri and Berger, 2000). One would need to compare these different model choices and see whether a similar heuristic search can be developed when the algorithm seeks for the median posterior probability model.

At first glance, modeling time series with auto-regression equations of the same order may appear to be a severe restriction. We have investigated the limitation of this assumption in simulated data (see Sebastiani and Ramoni, 2001) and the emerging result is that the results of our clustering method are robust to misspecification of the autoregressive order.

5. Acknowledgements

This research was supported by Eurostat, under contract EP29105. The authors thank Ed George and an anonymous referee for their invaluable help to improve the paper.

6. References

Banfield, J. D., and Raftery, A. E. (1993), 'Model-based Gaussian and non-Gaussian clustering', *Biometrics*, Vol. 49, pp. 803-821.

Barbieri, M., and Berger, J.O. (2000), 'Optimal predictive variable selection', ISDS Discussion paper, Duke University.

Box, G. E. P., and Jenkins, G. M. (1976, *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, CA.

Jeffreys, H. (1946), 'An invariant form for the prior probability in estimation procedures', *Proceedings of the Royal Society, London, A*, Vol. 186, pp. 453-461.

Ramoni, M., Sebastiani, P., and Cohen, P. (2000), 'Multivariate clustering by dynamics', in *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, San Francisco, CA. Morgan Kaufmann, pp. 633-638.

Ramoni, M., Sebastiani, P., and Cohen, P. (2001), 'Bayesian clustering by dynamics', *Machine Learning*, to appear.

Sebastiani, P., and Ramoni, M. (2000), 'Bayesian model-based clustering of time series' Submitted.

Sebastiani, P., Ramoni, M., and Cohen, P. (2001), 'Bayesian analysis of sensory inputs of a mobile robot', In *Proceedings of the 5th Workshop on Case Studies in Bayesian Statistics*, pp. 379-395.

Sebastiani, P., Ramoni, M., Cohen, P., Warwick, J., and Davis, J. (1999), 'Discovering Dynamics Using Bayesian Clustering', In *Proceedings of the Third International Symposium on Intelligent Data Analysis*. Lecture Notes in Computer Science, Springer, New York, pp. 199-210.

Unesco (1997), 'Schooling population [computer file], Paris:UNESCO (producer), r·cade online service (distributor),' Universities of Durham and Essex.

West, M., and Harrison, J. (1997), '*Bayesian Forecasting and Dynamic Models* (2nd Ed)', Springer, New York, NY.