ISBA 2000, Proceedings, pp. 000–000
ISBA and Eurostat, 2001

BAYESIAN CLUSTERING BY DYNAMICS OF EUROPEAN SCHOOL POPULATION

PAOLA SEBASTIANI AND MARCO RAMONI
Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA and
Children's Hospital Informatics Program, Harvard Medical School, Boston, MA

SUMMARY

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 auto-regressive 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 algorithm uses a distance-based heuristic search strategy.

Keywords: AUTO-REGRESSIVE MODELS; MODEL-BASED CLUSTERING.

1. INTRODUCTION

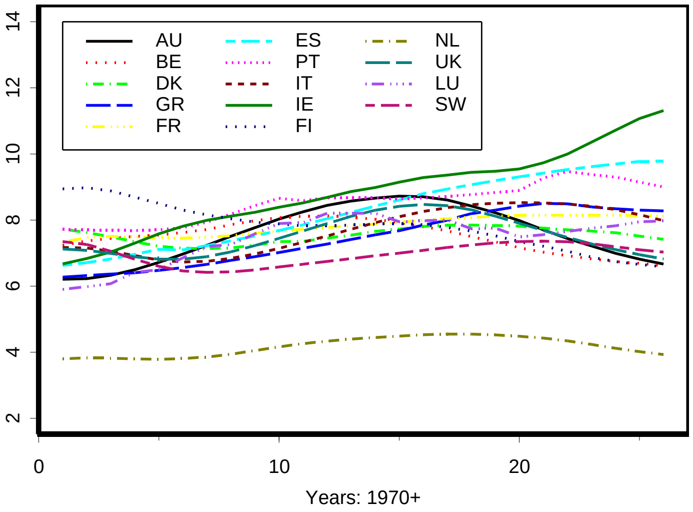
The time series in Figure 1 describe the share of population enrolled in higher education in 14 nations of the European community between 1970 and 1995. Our task is to group the 14 time series on the basis of their similarity in order to detect significant differences among European higher education trends. Data were provided by UNESCO and Eurostat, via the `r·cade` data bank[1], (Unesco, 1997).

The method to solve this problem depends on the meaning attached to the words *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 they are generated by the same process. Thus, deciding whether two time series are similar is equivalent to deciding whether they are observations of the same process. Put in this way, this task can be described 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. The solution we propose is an algorithm for Bayesian clustering by dynamics of continuous time series.

We model the stochastic process generating a time series as an auto-regressive model of order p, say $\mathsf{AR}(p)$, and then we cluster those time series that have a

[1] `http://www-rcade.dur.ac.uk`

Figure 1. *Share of population enrolled in higher education, between 1970 and 1996, in 14 European countries.*



high posterior probability of being generated by the same $\mathsf{AR}(p)$ model. The distinguished feature of our 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. Hence, 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 solution we have developed is to use a measure of similarity between $\mathsf{AR}(p)$ 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 methods do.

The next section describes the method and the search algorithm. The analysis of the higher education data set is describe in Section 3 and a discussion is in Section 4.

2. DESCRIPTION OF THE CLUSTERING ALGORITHM

The clustering method we present here has three components: a model for the time series, the posterior probability of a clustering model and a heuristic search strategy. These elements are considered in turns and they are explained with more details in Sebastiani and Ramoni (2000).

Autoregressive Models Let $S = \{y_{-p}, y_{-p+1}, \dots, y_t, \dots y_n\}$ be a time series of continuous values. The series follows an $\mathsf{AR}(p)$ model if

$$y \,|\, \beta, y_{-p}, \dots, y_{-1} = X\beta + \epsilon$$

where y is the vector $(y_0, y_1, \dots, y_n)^T$; X is the $n \times p$ matrix with tth row given by $(y_{t-1}, \dots, y_{t-p})$; $\beta = \{\beta_1, \dots, \beta_p\}$ is a vector of autoregressive coefficients and ϵ is a vector of uncorrelated errors which we assume normally distributed, with $E(\epsilon_t) = 0$ and $V(\epsilon_t) = \sigma^2 = \tau^{-1}$, for any t. The value p is called the order of the auto-regression, and specifies the Markov order of the series: namely that $y_t \perp (y_{-p}, \dots, y_{t-p-1}) \,|\, (y_{t-p}, \dots, y_{t-1})$ where the symbol $\perp$ denotes 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. The model can be extended to include a non-zero mean μ for each y_t, so that $\beta = \{\beta_0, \beta_1, \dots, \beta_p\}$, $\beta_0 = \mu(1 - \sum_{j=1}^{p} \beta_j)$, and the matrix X is augmented of a column of ones.

Conditional on the first p values, the likelihood function is

$$f(S \,|\, \beta, \tau, y_{-p}, \dots, y_{-1}) = \sqrt{\frac{\tau^n}{(2\pi)^n}} \exp\left(-\frac{\tau(y - X\beta)^T(y - X\beta)}{2}\right). \qquad (1)$$

We need to compute the posterior distributions of β and τ and the marginal likelihood $f(S \,|\, p)$, given a prior distribution for (β, τ) and the conditional likelihood in (1). Given an $\mathsf{AR}(p)$ model specification, we assume as prior density for (β, τ) the improper prior $f(\beta, \tau) = \tau^{-2}$, with $\tau > 0$. Suppose X is of full rank, and let

$$\hat{\beta} = (X^T X)^{-1} X^T y \qquad \mathsf{rss} = y^T(I_n - X(X^T X)^{-1} X^T)y. \qquad (2)$$

Then, one can show that the marginal likelihood is

$$f(S \,|\, p) = \frac{\left(\frac{\mathsf{rss}}{2}\right)^{(q+2-n)/2} \Gamma\left(\frac{n-q-2}{2}\right)}{(2\pi)^{(n-q)/2} \det(X^T X)^{1/2}} \qquad (3)$$

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

$$\beta \,|\, y, \tau \sim N(\hat{\beta}, [\tau(X^T X)]^{-1}) \qquad \tau \,|\, y \sim \text{Gamma}\left(\frac{\mathsf{rss}}{2}, \frac{n - q - 2}{2}\right) \qquad (4)$$

Both distributions are proper as long as X is of full rank, and $n > q + 2$. Bayesian estimates of β and σ^2 are $\hat{\beta}$ and $\mathsf{rss}/(n - q - 2)$.

Clustering Suppose now we have a batch of time series $S = \{S_1, S_2, ..., S_m\}$, which are supposed to be generated by an unknown number of stationary $\mathsf{AR}(p)$ models with a common auto-regressive order p and different auto-regressive coefficients. We want to cluster the time series in S according to their dynamics. The task of the algorithm is two-fold: finding the set of clusters that gives the best partition of the data and assigning each time series S_i 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 $C_1, \ldots, C_c$. Each state C_k of the variable C labels, in the same way, the time series generated by the same $\mathsf{AR}(p)$ model with auto-regression coefficients β_k 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 $\mathsf{AR}(p)$ model and thus merges the initial states $C_1, \ldots, C_m$ of the variable C into a subset $C_1, \ldots, C_c$, with $c \leq m$.

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 $P(M_c \mid S) \propto P(M_c) f(S \mid M_c)$ and we select the clustering model with maximum posterior probability. We show next 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 $\mathsf{AR}(p)$ model in cluster C_k are mutually independent. We denote by β_k the vector of auto-regression coefficients of the $\mathsf{AR}(p)$ model generating the time series in cluster C_k and suppose that each of these series can be represented as

$$Y_{kj} \mid \beta_k, \tau_k = X_{kj} \beta_k + \epsilon_{kj}.$$

The index k indicates cluster membership, and ϵ_{kj} is a vector of uncorrelated errors which we assume normally distributed, with $E(\epsilon_{kjt}) = 0$ and $V(\epsilon_{kjt}) = \tau_k^{-1}$, 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 same variance $\sigma_k^2 = \tau_k^{-1}$

allows us to represent the whole batch of series $\{S_{kj}\}$ in cluster C_k as

$$Y_k \mid \beta_k, \tau_k = X_k \beta_k + \epsilon_k$$

where the vector Y_k and the matrix X_k are defined as

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

Let β denote the set of parameter vectors $\{\beta_k\}$, where each β_k is a random q-vector, and let τ denotes the set of parameters τ_k for $k = 1, \ldots, c$. Then, by the independence of series assigned to different clusters, the overall likelihood function is

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

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 now define a prior distribution for θ as a Dirichlet $D(\alpha_1, ..., \alpha_c)$ and assign improper priors $f(\beta, \tau) = \prod_k \tau_k^{-2}$ to β and τ. Then, using the result in (3), and standard results on Dirichlet integration, it is easy to show that the marginal likelihood $f(S \mid M_c) = \int f(S \mid \theta, \beta, \tau) f(\theta) f(\beta, \tau) d\theta d\beta d\tau$ is

$$f(S \mid M_c) = \frac{\Gamma(\beta)}{\Gamma(\alpha + \sum_k m_k)} \prod_{k=1}^{c} \frac{\Gamma(\alpha_k + m_k)}{\Gamma(\alpha_k)} \frac{\left(\frac{\mathsf{rss}_k}{2}\right)^{(q+2-n_k)/2} \Gamma\left(\frac{n_k - q - 2}{2}\right)}{(2\pi)^{(n_k - q)/2} \det(X_k^T X_k)^{1/2}} \tag{5}$$

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

Once the, *a posteriori*, most likely partition has been selected, each cluster C_k is associated with parameters β_k modeling the auto-regression equation, and precision τ_k. The posterior distribution of $\beta_k \mid \tau_k, y_k$ is $N(\hat{\beta}_k, [\tau_k (X_k^T X_k)]^{-1})$, while the posterior distribution of $\tau_k \mid y_k$ is Gamma $(\mathsf{rss}_k/2, (n_k - q - 2)/2)$, from which the marginal posterior distribution of the auto-regression coefficient $\beta_k \mid y_k$ is a non-central Student's t, with expectation $\hat{\beta}_k$. Thus, $\hat{\beta}_k$ provides a point-estimate of β_k, and $(n_k - q - 2)/\mathsf{rss}_k$ is an estimate of the within-cluster precision. The probability of $C = C_k$ can be estimated as $\hat{p}_k = (\alpha_k + n_k)/(\alpha + \sum_k m_k)$.

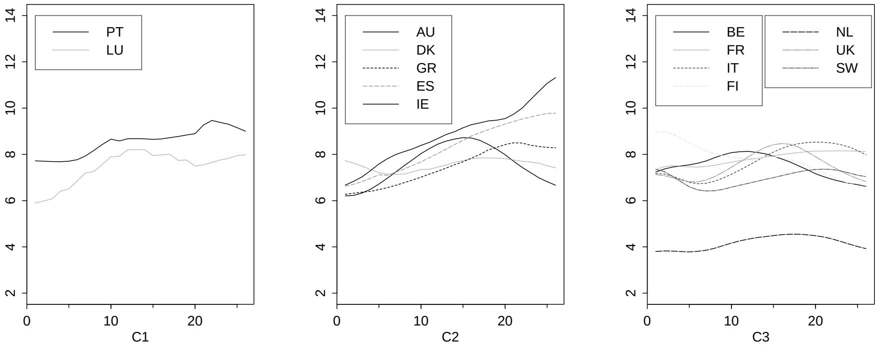
In practical applications, we use symmetric prior distributions for the parameters θ with a common prior precision α. The initial m hyper-parameters α_k are set equal to α/m and, when two time series are assigned to the same cluster C_k, their hyper-parameters are summed up. Thus, the hyper-parameters of a cluster corresponding to the merging of m_k time series will be $m_k \alpha/m$. In this way, the specification of the prior hyper-parameters requires only the prior global 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. A search algorithm to identify the best autoregressive order is described in Sebastiani and Ramoni (2000).

Search In principle, the clustering method described in the previous section requires one to compute the posterior probability of each clustering model and choose that one 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 $\mathsf{AR}(p)$ models in order to efficiently guide the search process in a subset of all possible clustering models. Since all $\mathsf{AR}(p)$ 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 $\beta_k \mid S$ 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 normal distribution conditional on the within-cluster precisions τ_k and τ_j. The precisions are then replaced by their posterior estimates.

Initially, the algorithm transforms the time series in S in a set of m $\mathsf{AR}(p)$ models, using the procedure described in the previous sections, and computes the set of $m(m-1)/2$ pairwise distances between posterior distributions of the parameters. Then, the algorithm sorts the generated distances, labels in the same way the two closest $\mathsf{AR}(p)$ models and evaluates whether the resulting model M_c, in which the two closest $\mathsf{AR}(p)$ models are assigned to the same cluster, is more probable than the model M_s in which they are distinct. If the probability $P(M_c \mid S)$ is larger than $P(M_s \mid S)$, 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 new parameter of the new $\mathsf{AR}(p)$ model and the remaining models in the set and the procedure is iterated on the new set. If the probability $P(M_c \mid S)$ is not larger than $P(M_s \mid S)$, 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 so far. The rationale behind this heuristic is that merging closest $\mathsf{AR}(p)$ models first should speed up the search for clustering models with large posterior probability and empirical evaluations of the methods on simulated data appear to support this intuition.

Figure 2. *Clusters found by the algorithm for the time series in Figure 1.*



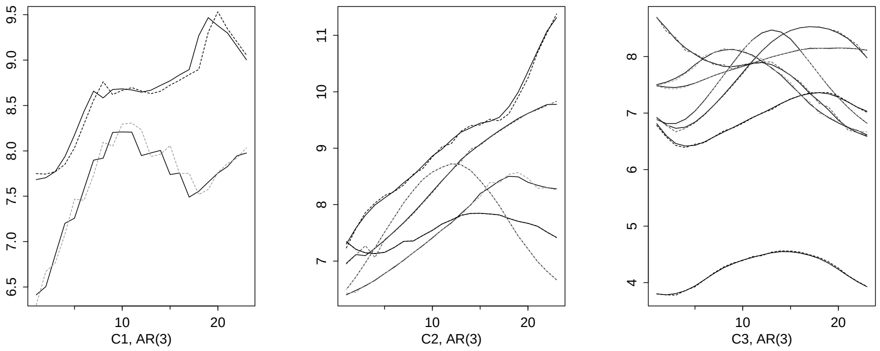
3. ANALYSIS

We apply the clustering algorithm described in the previous section to the analysis of the data set reporting the temporal evolution of the share of the population engaged in tertiary/higher education in 14 European countries. The fourteen time series 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 $\mathsf{AR}(3)$ models with a non-zero mean. We assumed $\alpha = 14$, the improper prior $f(\beta_k, \tau_k) = \tau^{-2}$, and uniform prior on all clustering models. Stationarity of the autoregressive models was checked at the end of the clustering process.

Figure 2 shows the three clusters of time series found by the algorithm. Cluster C_1 groups the trends of Portugal and Luxembourg. Estimates of the auto-regression coefficients are $\hat{\beta}_0 \approx 0.657$, $\hat{\beta}_1 \approx 1.133$, $\hat{\beta}_2 \approx 0.044$ and $\hat{\beta}_3 \approx -0.254$, thus the model is stationary — roots of the polynomial $f(u)$ are -2.38, 1.28±0.11i — with a mean $\hat{\mu} = 8.532$. Cluster C_2 groups the trends of Austria, Denmark, Greece, Spain and Ireland, and the estimates of the auto-regression coefficients are $\hat{\beta}_0 \approx 0.074$, $\hat{\beta}_1 \approx 2.085$, $\hat{\beta}_2 \approx -1.233$ and $\hat{\beta}_3 \approx 0.138$, with a mean $\hat{\mu} = 7.4$. The $\mathsf{AR}(p)$ is stationary, with roots of the polynomial $f(u)$ equal to 6.09 and $1.02 \pm 0.1i$. Cluster C_3 groups the time series of Belgium, France, Italy, The Netherlands, Finland, United Kingdom and Sweden. Estimates of the auto-regression coefficients are $\hat{\beta}_0 \approx 0.015$, $\hat{\beta}_1 \approx 2.593$, $\hat{\beta}_2 \approx -2.283$ and $\hat{\beta}_3 \approx 0.688$, thus defining a stationary autoregression equation with roots of the polynomial $f(u)$ equal to 1.023 and $1.14 \pm 0.32i$, and a mean $\hat{\mu} = 7.5$. These results do not seem to be susrprising. For example the third cluster groups the European nations which have been consistently stronger from an economic point of view in the past thirty years.

The fact that the time series of The Netherlands is assigned to the third

Figure 3. *Observed (continuous line) and fitted (dash line) time series in the clusters in Figure 2.*

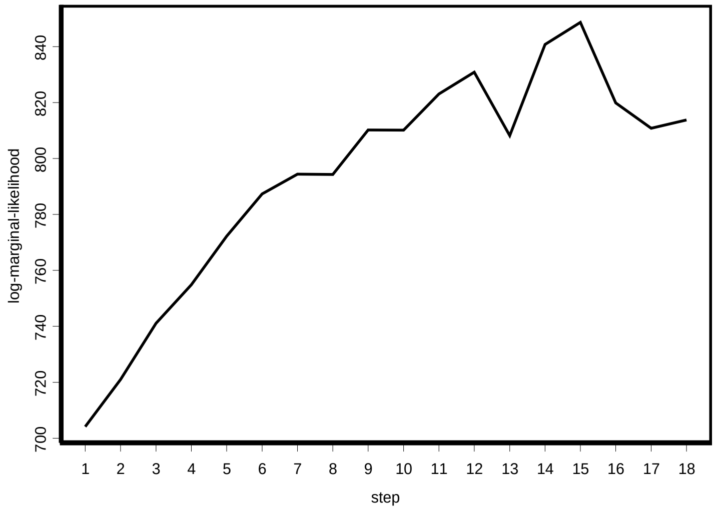


cluster is slightly disappointing: the dynamic of this series is similar to that of the other series in the cluster, but this series has a different mean. To evaluate the influence of this time series on the results, we run the clustering algorithm excluding the time series of The Netherlands. The algorithm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

During the analysis we assumed the time series were generated by AR(3) models. Plots of the observed and fitted values within clusters provides an overall assessment of the robustness of the result with respect to this assumption. Figure 3 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 the AR(3) models are a good approximation of the processes generating the original fourteen series.

Finally, it is interesting to point out that the three clusters of time series were found by the search algorithm in just eighteen steps. This number is much smaller than the total number 2^{14} of clusters to be considered without the heuristic search. Figure 4 shows the increase of the log-marginal likelihood 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 which 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 nearest clusters 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 4. *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 and a systematic exposition is in Box and Jenkins (1976). The Bayesian analysis of AR(p) models is described in West and Harrison (1997). Bayesian model-based clustering was originally proposed by Banfield and Raftery (1993) to cluster static data. Ramoni *et al.* (2000, b) 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 (Sebastiani *et al.*, 1999), simulated war games (Ramoni *et al.*, 2000, b), as well as the behavior of stocks in market and automated learning and generation of Bach's counterpoint. The algorithm was extended by Ramoni *et al.* (2000, a) to cluster multivariate time series.

Contrary to BCD, the algorithm presented in this paper clusters continuous time series. 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 described in this paper 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, for example, selecting the median posterior probability model (Barbieri and Berger, 2000). An open question is to compare these different model choices and to 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 as auto-regression models of the same order may appear to be a severe restriction. We have investigated the limitation of this assumption in simulated data (Sebastiani and Ramoni, 2000) and the emerging result is that the results of our clustering algorithm are robust to mispecification of the autoregressive order, as long as the specified order is higher than the order of the generating processes.

5. ACKNOWLEDGEMENTS

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

REFERENCES

Banfield, J. D., and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, **49**, 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* (2nd Ed). 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*, **186**, 453–461.

Ramoni, M., Sebastiani, P., and Cohen, P. R. (2000, a). Multivariate clustering by dynamics. In *Proceedings of the Seventeeth National Conference on Artificial Intelligence* San Francisco, CA. Morgan Kaufmann.

Ramoni, M., Sebastiani, P., Cohen, P. R., Warwick, J., and Davis, J. (2000, b). Bayesian clustering by dynamics. *Machine Learning.* To appear.

Sebastiani, P., and Ramoni, M. (2000). Bayesian clustering by dynamics of autoregressive models. Technical report, Department of Mathematics, Imperial College. Available from the authors.

Sebastiani, P., Ramoni, M., and Cohen, P. (1999). Bayesian analysis of sensory inputs of a mobile robot. In *Proceedings of the 5th Workshop on Case Studies in Bayesian Statistics.* To appear.

Unesco (1997) Schooling population [computer file], Paris: UNESCO (producer), rcade 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.