Elsevier Editorial System(tm) for Computational Statistics and Data Analysis

Manuscript Draft

Manuscript Number:

Title: Time series clustering based on forecast densities

Article Type: Research Paper

Section/Category: II. Statistical Methodology for Data Analysis

Keywords: Bootstrap; Cluster analysis; L^{2}-distance; Kyoto Protocol; Nonparametric density estimation; Prediction

Corresponding Author: Dr. Ana Justel, PhD

Corresponding Author's Institution: Universidad Autonoma de Madrid

First Author: Andres M Alonso, PhD

Order of Authors: Andres M Alonso, PhD; Jose R Berrendero, PhD; Adolfo Hernandez, PhD; Ana Justel, PhD

Time series clustering based on forecast densities

A.M. Alonso*, J.R. Berrendero**, A. Hernández*** and A. Justel**[1]

*Departamento de Estadística, Universidad Carlos III de Madrid, Spain

**Departamento de Matemáticas, Universidad Autónoma de Madrid, Spain

***Department of Mathematical Sciences, University of Exeter, U.K.

Abstract

A new clustering method for time series is proposed, based on the full probability density of the forecasts. First, a resampling method combined with a nonparametric kernel estimator provides estimates of the forecast densities. A measure of discrepancy is then defined between these estimates and the resulting dissimilarity matrix is used to carry out the required cluster analysis. Applications of this method to both simulated and real life data sets are discussed.

Keywords and phrases: bootstrap, cluster analysis, L^2-distance, Kyoto Protocol, nonparametric density estimation, prediction.

1 Introduction

Time series clustering problems arise when we observe a sample of time series and we want to group them into different categories or clusters. This is an important area of research for different disciplines. In seismology, Kakizawa, Shumway and Taniguchi (1998) apply cluster techniques in order to establish similarities among or differences between classes of events such as earthquakes and mining explosions. Some published examples of cluster analysis in time series have been based on environmental data, where we have time series from different locations and wish to group locations which

[1]Corresponding author: Ana Justel, Departamento de Matemáticas, Universidad Autónoma de Madrid. Campus de Cantoblanco, 28049 Madrid, Spain. Email: ana.justel@uam.es

Tel: +34-914977634; fax: +34-914974889.

show similar behaviour. See for instance Macchiato *et al.* (1995) for a spatial clustering of daily ambient temperature, or Cowpertwait and Cox (1992) for an application to a rainfall problem. Other examples can be found in medicine, economy, engineering, etc.

Cluster problems have been studied extensively, but because a time series often involves several (may be hundreds or thousands) observations collected over time, its dimensionality generally prohibits computations in the time domain using classical multivariate methods. In addition, these general cluster techniques ignore the autocorrelation structure of the time series. This motivates the need for developing specific clustering methods in the field of time series.

We propose a new approach motivated by the fact that in many practical situations we may not be interested in clustering on the basis of cross-sectional information that is inherently static and ignores the evolution of the series. Neither are we interested in clustering based on the models that generated the observations, but in respect of the forecasts at a specific future time. This can be illustrated with the example in Figure 1. Clustering these three time series based on the models or on the last observed values will obviously produce an entirely different result to the one coming from the study of the forecasts at a specific horizon. This idea introduces an extra consideration to the problem, namely what the purpose of the grouping is. Our technique will be specially appropriate when the real interest is either on the long term series convergence or divergence, or on whether some specified level is going to be reached. These questions often arise in many practical situations, as for instance in any sustainable development problem. In this paper we consider the case of country reductions of CO_2 emissions in order to reach the Kyoto Protocol compromises in the fixed horizon of year 2012 (United Nations, 1997). In these situations it may be appropriate to employ a cluster method relying directly on the properties of the predicted values.

The new cluster procedure is based on the full forecast densities for each one of the observed series in the sample, instead of focusing on the point forecasts. We apply a smoothed sieve bootstrap procedure combined with nonparametric kernel density estimation ideas to approximate the distribution of the predictions. This is done in a general context, non constrained to the habitual Gaussianity hypothesis. Differences between each pair of bootstrap densities provide a dissimilarity matrix which will be used to discuss possible clustering structures.

There are some previous references in the literature that have considered the problem

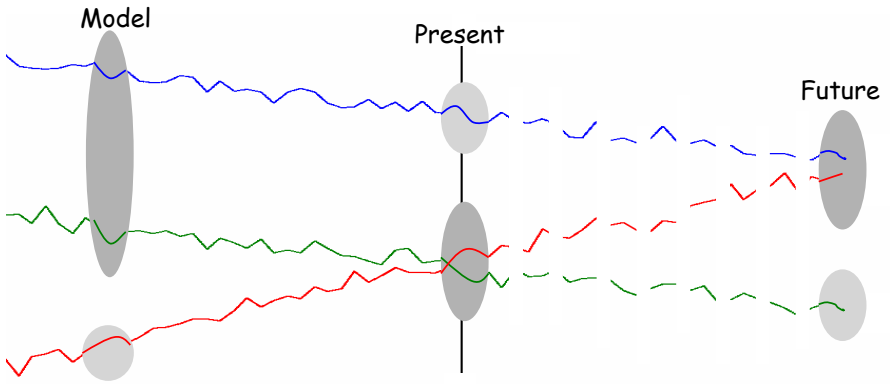

Figure 1: Three different cluster solutions depending on whether modelling, present information or future values are used.

of clustering time series. Among others, we can cite the works from Piccolo (1990) or Maharaj (1996), where time series are first modelled and then a distance is defined in terms of the model parameters. Kakizawa *et al.* (1998) characterize similarities among and differences between multivariate stationary time series in terms of the structure of the covariance or, equivalently, the spectral matrices (the so called *spectral clustering*). Fruhwirth-Schnatter and Kaufmann (2004) propose a Bayesian mixture model with a MCMC algorithm for parameter estimation and data classification. Finally, Pattarin, Paterlini and Minerva (2004) make use of classical multivariate methods after a genetic algorithm has been implemented to reduce the dimensionality of the time series. A survey of clustering and other relevant multivariate techniques for time series, including further references, can be found in Galeano and Peña (2000).

Compared with these proposals, our new procedure presents some interesting advantages. This approach reduces the high dimension of the *3D* problem by converting the *data-cube* structure of p different time series measured in m individuals over T moments of time, into a more tractable *2D* structure of p forecasts obtained for m individuals at fixed time $T + h$ (see Figure 2). Since our predictions will include information from both present and past values of the series, we are not discarding any valuable knowledge. Finally, considering the full forecast densities instead of merely point forecasts potentially allows us to classify into different clusters time series generated by models which are essentially similar, e.g. models which differ only in the variability of the observations or in the distribution of the innovations, but which produce different forecast densities. Moreover, our method will be able to distinguish between situations like the

one illustrated in Figure 3, where both pairs of distributions have the same mean but the full densities offer a better understanding regarding similarities among or differences between them.

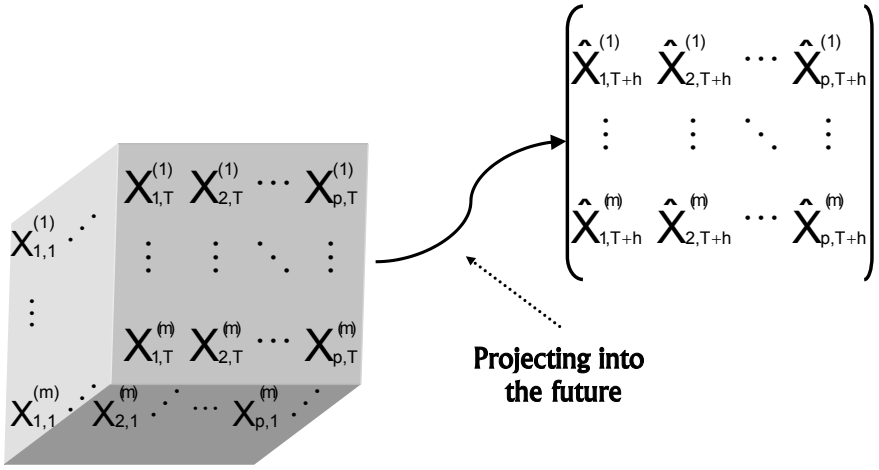

Figure 2: Dimension reduction by using forecast information.

The paper is organized as follows. Section 2 introduces notation and presents the steps of our clustering method. Section 3 is devoted to its application to some simulated time series, with comparisons of its results with those obtained with other techniques. Section 4 analyses historical data of CO_2 emissions in industrialized countries, setting 2012 as the horizon fixed in the Kyoto Protocol. Finally, section 5 presents some conclusions and possible generalized uses of our work.

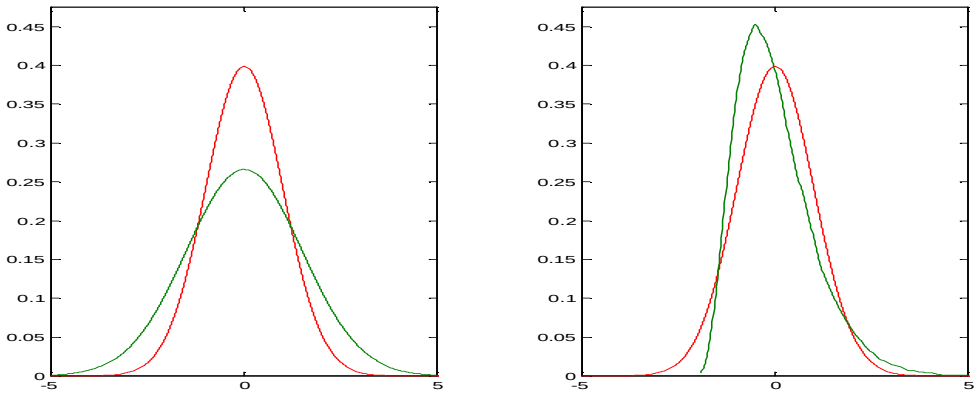

Figure 3: Two examples where point forecast clustering fails.

2 Time series clustering methodology

The definition of an appropriate measure of distance or dissimilarity between observations is of central importance in most applications of cluster analysis. The choice of this distance should take account of the final goals of the clustering procedure so that the distances capture the particular discrepancies between observations that are relevant for our purposes. As discussed in the introduction, our intention is to cluster time series based on their full forecast densities at a specific future time $T+h$. Therefore, a distance between these densities seems appropriate. We have chosen the squared L^2 distance for its computational advantages and its analytical tractability. More specifically, for $i = 1, \ldots, m$, let $\mathbf{X}^{(i)} = (X_1^{(i)}, \ldots, X_T^{(i)})$ be the time series in the sample corresponding to the i-th individual, let $f_{X_{T+h}^{(i)}}$ denote the density function of the forecast $X_{T+h}^{(i)}$, then the selected distance is

$$D_{ij} = \int \left(f_{X_{T+h}^{(i)}} - f_{X_{T+h}^{(j)}}\right)^2 dx, \qquad i, j = 1, \ldots, m. \tag{2.1}$$

As a consequence of the discussion above, implementation of our method for clustering time series $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(m)}$ in respect of their forecasts at a specific horizon requires the following three step methodology.

- Step 1: Computation of predictions.
- Step 2: Computation of the dissimilarity matrix.
- Step 3: Application of classical cluster methods to the dissimilarity matrix.

In practice density forecasts are unknown, so the distances D_{ij} cannot be directly computed and we must approximate them from the data. Assume for simplicity that the model is a zero mean AR(p) process. A classical solution for prediction densities (see for instance Box, Jenkins and Reinsel, 1994) gives the full estimated distribution

$$\mathcal{N}(\widehat{E}[X_{T+h}|X_1, \ldots, X_T], \widehat{\sigma}^2 \sum_{j=0}^{h-1} \widehat{\psi}_j^2),$$

where $\widehat{\psi}_j$ are the estimated MA coefficients and

$$\widehat{E}[X_{T+h}|X_1, \ldots, X_T] = -\widehat{\phi}_1[X_{T+h-1}] - \cdots - \widehat{\phi}_p[X_{T+h-p}],$$

with $[X_t] = X_t$ if $t \leq T$, otherwise $[X_t] = \widehat{E}[X_t|X_{T-p+1}, \ldots, X_T]$.

Two points can be made about the above expressions. On the one hand, they do not take into account the variability of the $\widehat{\phi}$'s and $\widehat{\psi}$'s random variables. On the other, they assume that the distribution of the error process ε is known (a standard Gaussian distribution). These results could be adversely affected by departures from normality. For example, using a Montecarlo study, Thombs and Schucany (1990) have shown the poor performance of this method given a skewed bimodal error distribution.

To overcome these restrictions, our approach uses a smoothed sieve bootstrap procedure, which is a smoothed version of the forecasting procedure proposed by Alonso, Peña and Romo (2002, 2003). The sieve bootstrap is based on residual resampling from an autoregressive approximation of the given process, and has a nice nonparametric property, being model-free within the considered class of linear processes. Thus, the procedure could be applied to a more general class of linear models without specifying a finite dimensional model as in previous bootstrap proposals, e.g., Thombs and Schucany (1990), Cao *et al.* (1997) and Pascual, Romo and Ruiz (2004).

2.1 Generating bootstrap predictions

Let $\{X_t\}_{t\in\mathbb{Z}}$ be a real valued process that admits the representation $\sum_{j=0}^{+\infty}\phi_j B^j Y_t = \varepsilon_t$, where $Y_t = (1-B)^d X_t$, $\phi_0 = 1$ and d is assumed to be known. This AR(∞) representation motivates the sieve bootstrap which proposes to use a sequence of approximating autoregressive models for $\{X_t\}_{t\in\mathbb{Z}}$ with order $p = p(T)$ that increases as a function of the sample size T. Given a sample $(X_1, \ldots, X_T)$, the resampling scheme proceeds as follows:

1. Differentiate the series. Then $Y_t = (1-B)^d X_t$, $t = d+1, \ldots, T$.

2. Select the order $p = p(T)$ for the series $(Y_{d+1}, \ldots, Y_T)$, using any standard criteria. We follow the recommendations of Hurvich and Tsai (1989) and use the AICC criterion, where

$$\text{AICC} = -T' \log\left(\widehat{\sigma}^2\right) + 2\left(p+1\right)T'/\left(T'-p-2\right),$$

   $T' = T - d$, and $\widehat{\sigma}^2$ is the estimated residual variance.

3. Estimate the autoregressive coefficients by least squares, $\widehat{\boldsymbol{\phi}} = (\widehat{\phi}_1, \ldots, \widehat{\phi}_p)'$.

4. Compute the least square residuals, $\widehat{\varepsilon}_t = \sum_{j=0}^{p} \widehat{\phi}_j B^j Y_t$, and the kernel density estimate of the centered residuals, $\widehat{f}_{\tilde{\varepsilon}}$, where $\tilde{\varepsilon}_t = \widehat{\varepsilon}_t - \widehat{\varepsilon}^{(\cdot)}$ and $\widehat{\varepsilon}^{(\cdot)}$ is the mean of the $\widehat{\varepsilon}_t$.

5. Draw a bootstrap-resample ε_t^* of i.i.d. observations from $\widehat{f}_{\tilde{\varepsilon}}$.

6. Define the bootstrap series X_t^* by the recursion

$$\sum_{j=0}^{p} \widehat{\phi}_j B^j (1-B)^d X_t^* = \varepsilon_t^*,$$

then $(Y_{d+1}^*, \ldots, Y_T^*)$ is the bootstrap differentiated series, $Y_t^* = (1-B)^d X_t^*$, and $\widehat{\boldsymbol{\phi}}^*$ their least square estimates of the autoregressive coefficients.

7. Compute bootstrap prediction-paths by the recursion

$$\sum_{j=0}^{p} \widehat{\phi}_j^* B^j (1-B)^d X_t^* = \varepsilon_t^*,$$

for $t = T+1, T+2, \ldots, T+h$, where $h > 0$, and $X_t^* = X_t$, for $t \leq T$.

This method allows us to obtain B copies of the h-step-ahead predicted values, where the horizon h is selected by the user. An outline of the sieve bootstrap procedure is displayed in Figure 4.

Note that obtaining a resample from $\widehat{f}_{\tilde{\varepsilon}}$ is analogous to taking a resample from the empirical distribution of the centered residuals, $\widehat{F}_{\tilde{\varepsilon}}$, and adding some scaled Gaussian noise $g\xi_t$, where g is the bandwidth used in kernel estimation and the ξ_t are i.i.d. standard normal variables.

2.2 Computation of the dissimilarity matrix and application of the clustering method

The bootstrap sample $(X_{T+h}^{*1}, X_{T+h}^{*2}, \ldots, X_{T+h}^{*B})$ is used to approximate the unknown distribution of X_{T+h} given the observed sample. In particular, we will apply kernel estimation techniques, as described in Silverman (1986), to obtain $\widehat{f}_{X_{T+h}^{(i)*}}(x)$, the h-step-ahead kernel density estimator at point x for the i-th time series. We can now estimate the squared L^2 distance D_{ij} defined in equation (2.1) using

$$\widehat{D}_{ij}^* = \int \left(\widehat{f}_{X_{T+h}^{(i)*}}(x) - \widehat{f}_{X_{T+h}^{(j)*}}(x) \right)^2 dx. \tag{2.2}$$

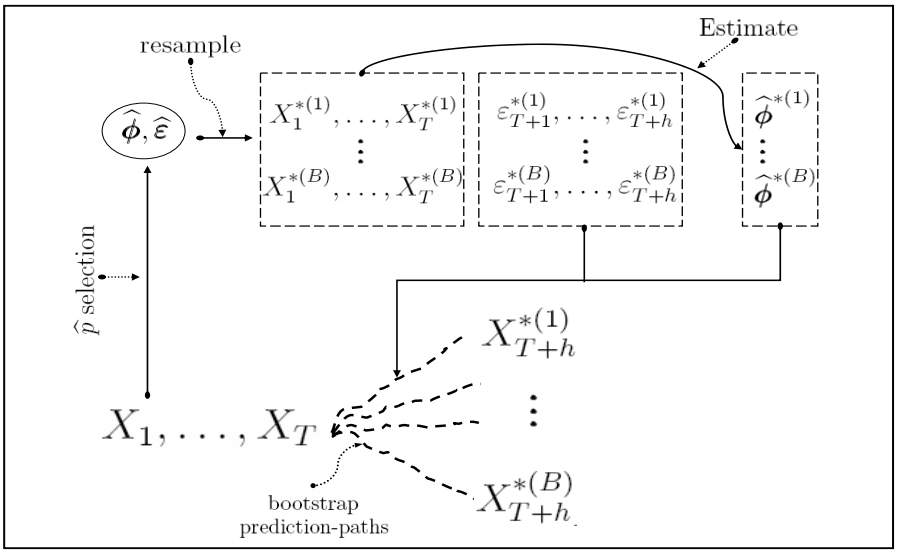

Figure 4: Outline of the sieve bootstrap procedure.

Notice that

$$\widehat{D}_{ij}^{*} = \int \widehat{f}^2_{X_{T+h}^{(i)*}}(x)dx + \int \widehat{f}^2_{X_{T+h}^{(j)*}}(x)dx - 2\int \widehat{f}_{X_{T+h}^{(i)*}} \widehat{f}_{X_{T+h}^{(j)*}} dx. \tag{2.3}$$

Therefore, to compute the kernel estimates we use the bandwidth proposed by Sheather, Hettmansperger and Donald (1994) which is specially designed to estimate the functionals $\int f^2(x)dx$.

In the Appendix we prove the consistency of $\widehat{D}_{ij}^{*}$ as an estimator of D_{ij}. This result follows from the validity of the smoothed sieve bootstrap and some standard results from nonparametric density estimation theory.

Of course, other measures between densities can be used, e.g. L^1 distance or Kullback-Leibler discrepancy, but we choose the L^2 distance for both computational advantages (expression (2.3) can be explicitly calculated) and theoretical considerations (we can use previous bootstrap literature to establish the bootstrap validity).

The dissimilarity matrix $\widehat{D}^{*} = (\widehat{D}_{ij}^{*})$, for $i, j = 1, \ldots, m$, is now used to carry out the cluster analysis. There are many different cluster algorithms that can be used in this situation, but it is worth pointing out that we are restricted to those operations on a dissimilarity matrix and not on the classical multivariate scenario of a $m \times T$ data matrix. This fact excludes, for instance, standard algorithms like k-means. Once the matrix of $m \times m$ dissimilarities has been computed, some widely used clustering methods proceed in an agglomerative manner, that is, starting with each single observation in a different cluster and joining clusters together until there is only one cluster containing

all the observations. In our experiments, an agglomerative hierarchical method with nearest distance (single linkage) as grouping criteria has provided reasonable results. For details, see for instance Everitt *et al.* (2001).

3 Simulation study

A simulation study is carried out in order to evaluate how our method performs compared to the classical Box-Jenkins (BJ) methodology. From the description of the three steps in previous sections, it is clear that the performance of our method relies heavily on how the dissimilarity matrix obtained in step 2 reflects accurately the differences between each pair of series. In order to study this, we generate two series from the same autoregressive model,

$$X_t = 0.75X_{t-1} - 0.5X_{t-2} + \varepsilon_t, \qquad t = 1, \ldots, 100.$$

We then compare the distances between the forecast densities, given by equation (2.1), computed in three different ways: first, according to our procedure, $\widehat{D}^*$; second, from the analytical expression under normality assumptions, $\widehat{D}^{BJ}$; and third, using Monte Carlo forecasts instead of bootstrap predictions in the estimate (2.3), $\widehat{D}^{MC}$. We consider three distributions for the innovations, $\mathcal{N}(0,1)$, Student-t with 3 degrees of freedom and the centered exponential $\mathcal{E}xp(1) - 1$. This allows us to discuss results in situations where both kurtosis or skewness problems are present. We consider three forecast horizons, from the short term $h = 1$ step ahead, to the long term $h = 10$, with an intermediate period of $h = 3$. For each case, we replicate the experiment 1000 times.

Figure 5 shows the results for the standard Gaussian innovations. Each diagram-row corresponds to a different horizon, h =1, 3, 10, and contains the boxplot representations of the differences $D - \widehat{D}^{MC}$ (left), $D - \widehat{D}^{BJ}$ (center) and $D - \widehat{D}^*$ (right). Notice that the Gaussian case is the only one where we can calculate analytically the exact value of D.

Since Monte Carlo results are based on the true generating model (model's structure, innovations distribution and parameters), we consider this estimate of D as a benchmark in the study. However, the differences $D - \widehat{D}^{MC}$ are not always zero because of the Monte Carlo variability from the number generator and, in this particular case, the nonparametric estimation of the integrals in (2.3). In the case of Gaussian innovations, the Box-Jenkins' approach uses the true model structure and innovations distribution,

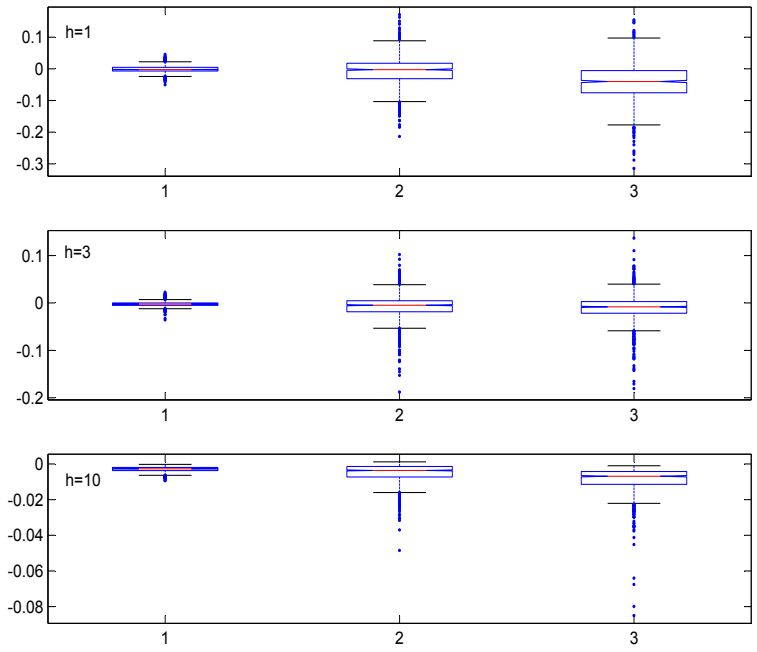

Figure 5: Simulation results for Gaussian innovations. Boxplots of the differences $D - \widehat{D}^{MC}$ (left), $D - \widehat{D}^{BJ}$ (center) and $D - \widehat{D}^{*}$ (right) for prediction horizons $h = 1$, 3, 10.

but not the true value of the parameters which need to be estimated from the data. The differences in symmetry observed in the boxplots for the three horizons are explained by the fact that, although both series are generated from the same stationary model, the forecasts depend on the last observed value which is different for both series. This effect is specially important in short term forecast and gradually disappears in the long term, where distances D must approximate to zero.

As expected, given the normality assumptions, Box-Jenkins' method gets better results than the bootstrap approach, although the benefits of its use tend to decrease for bigger horizons.

Figure 6 shows results for Student-t innovations with 3 degrees of freedom. Because now the true value D is not known, each diagram-row represents differences $\widehat{D}^{MC} - \widehat{D}^{BJ}$ (left) and $\widehat{D}^{MC} - \widehat{D}^{*}$ (right). Bootstrap results are better, specially in the short term $h = 1$, where Box-Jenkins' method gives biased estimates. The effect is less obvious as h increases.

Finally, Figure 7 shows results for centered exponential $\mathcal{E}xp(1) - 1$ innovations. Again, the true value D is not known, therefore each diagram-row represents differences $\widehat{D}^{MC} - \widehat{D}^{BJ}$ (left) and $\widehat{D}^{MC} - \widehat{D}^{*}$ (right). Bias of the Box-Jenkins' methodology is

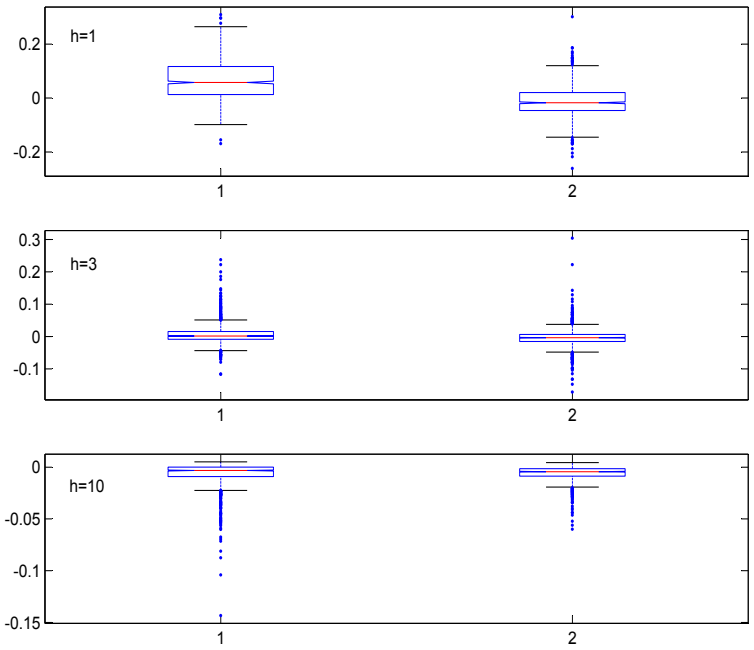

Figure 6: Simulation results for Student-t innovations. Boxplots of the differences $\widehat{D}^{MC} - \widehat{D}^{BJ}$ (left) and $\widehat{D}^{MC} - \widehat{D}^{*}$ (right) for prediction horizons $h = 1, 3, 10$.

more obvious here than in the previous case, again specially for horizon h =1. Here, the Box-Jenkins estimates exceed the Monte Carlo estimates in more than 0.2 in 75% of replications. This effect is less marked as h increases.

As a summary, our method outperforms Box-Jenkins' methodology for non Gaussian innovations, specially for short term forecasts. As expected, Box-Jenkins' estimates give better results if a Gaussian model is fixed.

4 Clustering of countries facing the Kyoto agreements for 2012

We apply the method outlined in this paper to data related to CO_2 emissions in different industrialized countries. The horizon we use is that established at Kyoto where, after lengthy discussions, an agreement was reached to reduce emissions by 2012. It makes sense to group countries in this framework in order to create classes of countries with common interests which can share experiences, politics, etc. in their way to achieve their target values.

Figure 8 shows values of CO_2 emissions (in metric tons per capita) in 24 industri-

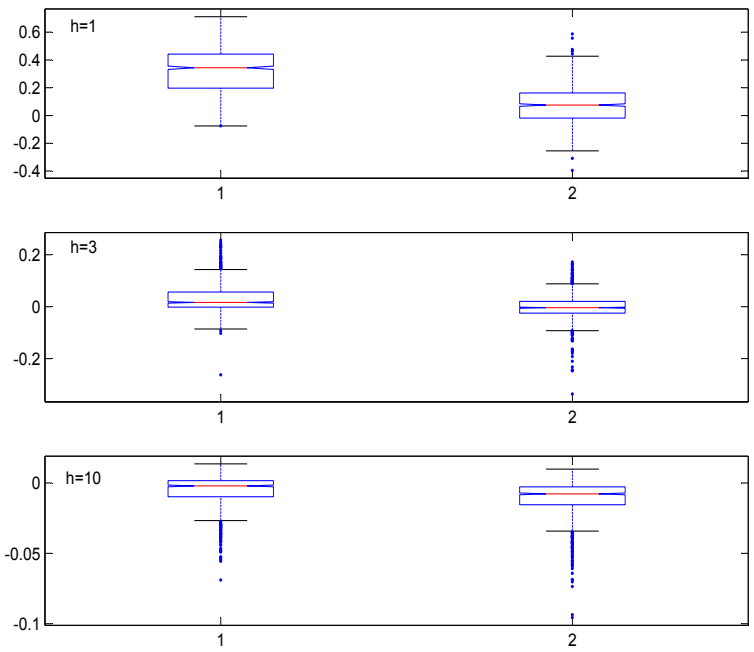

Figure 7: Simulation results for exponential innovations. Boxplots of the differences $\widehat{D}^{MC} - \widehat{D}^{BJ}$ (left) and $\widehat{D}^{MC} - \widehat{D}^{*}$ (right) for prediction horizons $h = 1, 3, 10$.

alized countries. The data covers from 1960 to 1999. It is worth pointing out that Kyoto summit took place in 1997, so this data probably does not include any significant reaction to that event. Broadly speaking, for most countries there is a tendency for increasing values of CO_2 emissions, with those corresponding to the last ten years lying inside a band between 5 and 15 metric tons per capita. Countries departing from this behaviour are: Luxembourg, with extremely high values but a clear decreasing tendency; USA, with values around a (high) constant in the last 30 years; Australia, which shows a pronounced increase in the last 20 years; and Canada which, with the previous three mentioned, has values above 15 metric tones per capita in the last few years. Alternatively, China is the country with the lowest values. Not surprisingly, these countries will feature later in our analysis.

First, we apply cluster techniques to the last available observation. Therefore, this is not a proper time series clustering problem, but a univariate classical one. We have included it for comparison purposes. It is difficult to say anything about the whole structure of the grouping, so we center our discussion in the biggest values of CO_2 emissions. From the dendrogram of Figure 9 it can be stated, for instance, that Luxembourg and Australia are grouped together first, joined later by USA and Canada. This is in

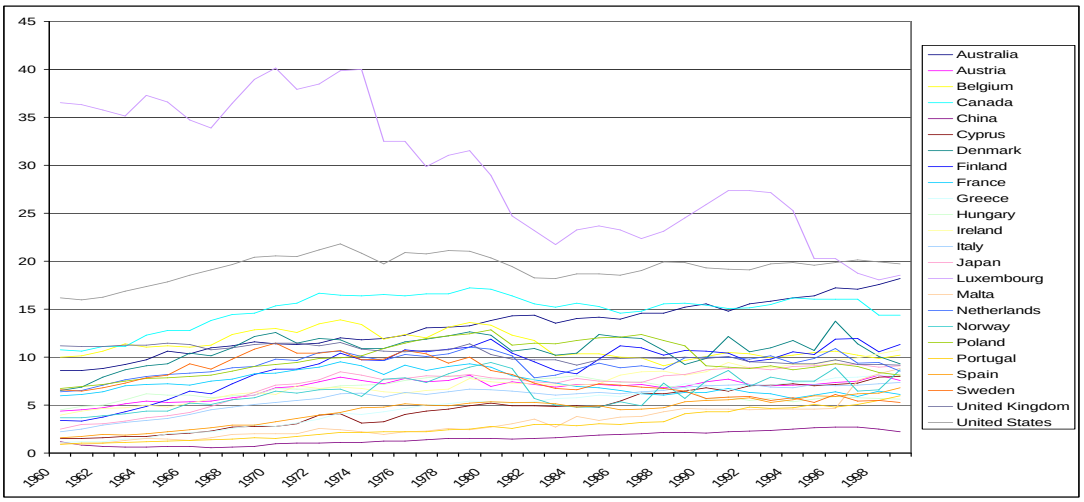

Figure 8: CO2 emissions of 24 industrialized countries in the period 1960–1999.

accord with previous knowledge gained from Figure 8.

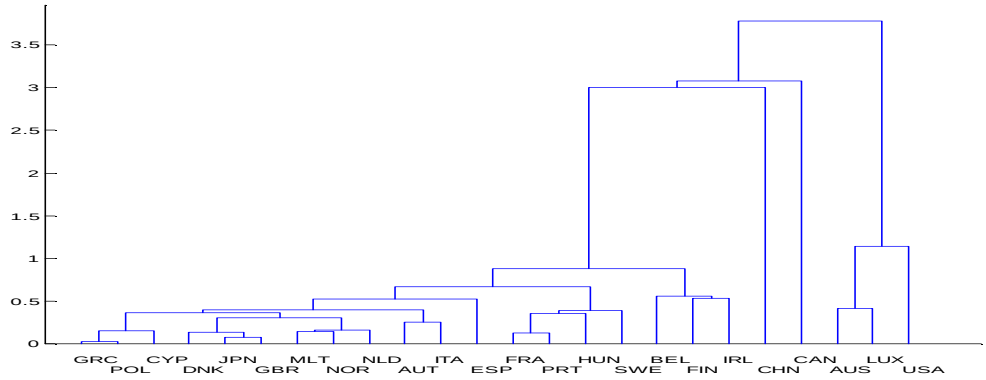

Figure 9: Dendrogram based on absolute differences of the last observation, 1999.

Obviously, consideration of only the last observation implies we do not use all the past history of the time series. Figure 10 shows the dendrogram based on point predictions for 2012, where different results are expected since now the whole history of the process is taken into account. USA and Luxembourg are still together, but now Australia groups with Finland in the first instance. Even Malta joins this group before Canada.

Applying our method, Figure 11 shows the dendrogram based on $\widehat{D}^*$ distances for 2012. Results differ from those obtained in the previous two situations, and new clusters of countries arise. In particular, USA and Luxembourg are no longer grouped together: Luxembourg joins Finland first, and USA is much closer to Australia. This new configuration is clarified when looking at Figure 12, where point forecasts are compared with forecast densities. Although point forecasts for USA and Luxembourg are close, as well

as those from Finland and Australia, if we take into account the full forecast densities, as our method suggests, it turns out that in fact USA is much closer to Australia, and Luxembourg to Finland. This is a situation that could have escaped our consideration had the cluster method relied only on point forecasts.

Additionally, another dendrogram was obtained assuming Gaussian innovations. Results are not equal to Figure 11 but differences are not substantial. This means that the assumption of normality is reasonable for most of the countries in this example and that the observed differences between the dendrograms are due to slight deviations from the Gaussian model. Only for a few countries, the estimated prediction densities exhibit a notable departure to the Gaussian density (e.g., see Finland in Figure 12).

Before concluding with this example, we would like to mention that the number of regular differences, d, for each time series was selected using the program TRAMO (**T**ime series **R**egression with **A**RIMA noise, **M**issing observations and **O**utliers) developed by Gómez and Maravall (1996). TRAMO is a program for modelling time series with missing observations in the presence of possibly several types of outliers. When used automatically, TRAMO identifies the reg-ARIMA model by testing for the log/level transformation and the presence of calendar-type effects, as well as detecting and correcting additive outliers, transitory changes and level shift. A complete description of the methodology behind TRAMO can be found in Gómez and Maravall (1994, 2001). For the CO_2 emission data, the TRAMO program was only used to identify the number of regular differences and to test log/level transformation.

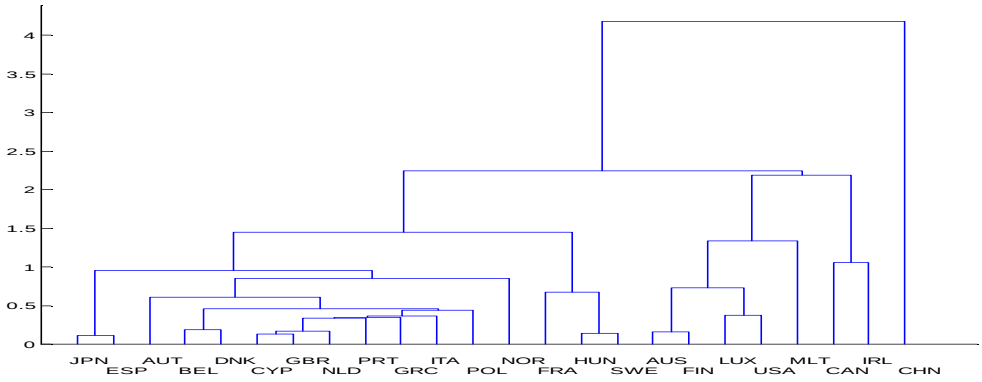

Figure 10: Dendrogram based on absolute differences of point forecasts for 2012.

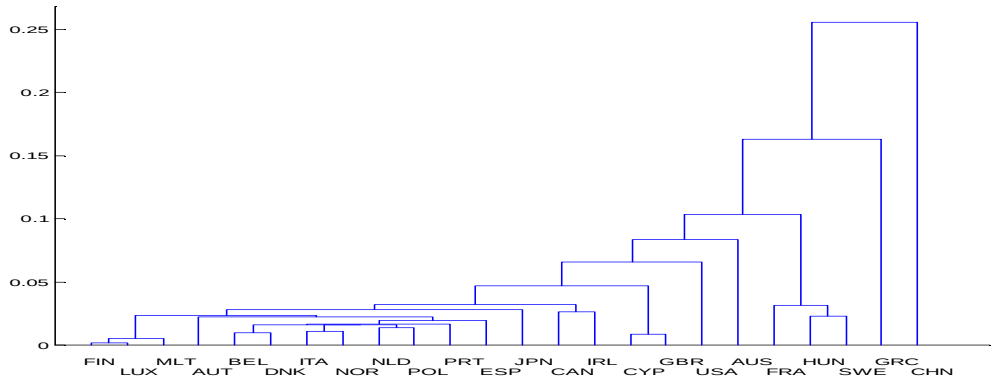

Figure 11: Dendrogram based on $\widehat{D}^*$ distance of prediction densities for year 2012.

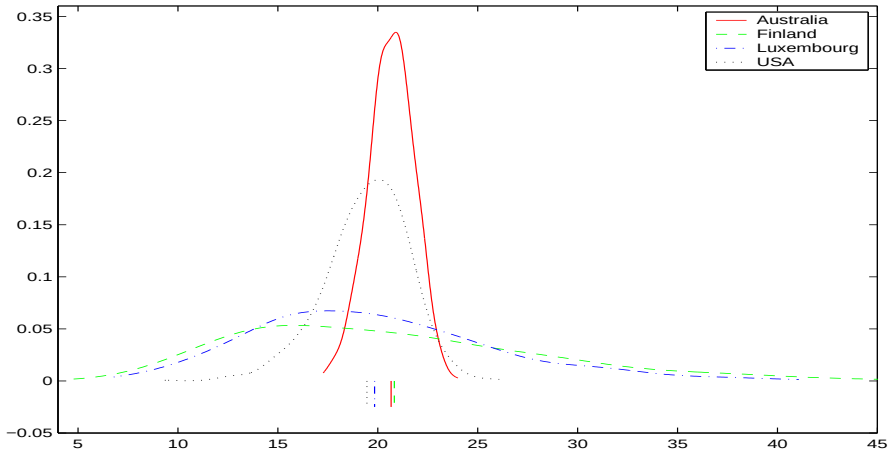

Figure 12: Prediction densities for selected countries.

5 Conclusions and Extensions

We have studied the problem of clustering time series for a general class of linear models and proposed a procedure based on the sieve bootstrap to calculate discrepancies between forecast densities. We have compared its performance with other methods under a simulation study and also applied the methodology in an interesting environmental problem, namely issues concerning CO_2 emissions following the Kyoto Protocol.

Possible extensions of this work include its modification for finding subgroups in each cluster using their observed trend, as well as the generalization to the case where each element of the sample is in fact a multivariate time series.

Acknowledgements

This research has been supported by Comunidad de Madrid (Spain) grant 06/0050/2003 and CICYT (Spain) grant MTM2004-00098. The first author acknowledges the support of a "Juan de La Cierva" grant.

Appendix

In this appendix we show the consistency of $\hat{D}^*_{ij}$ [see equation (2.2)] as an estimate of D_{ij} [see equation (2.1)]. For two stationary processes $\{X_t\}_{t\in\mathbb{Z}}$ and $\{Y_t\}_{t\in\mathbb{Z}}$ we need to show

$$\left\| \hat{f}_{X^*_{T+h}|\boldsymbol{X}^T_{-\infty}} - \hat{f}_{Y^*_{T+h}|\boldsymbol{Y}^T_{-\infty}} \right\| \to \left\| f_{X_{T+h}|\boldsymbol{X}^T_{-\infty}} - f_{Y_{T+h}|\boldsymbol{Y}^T_{-\infty}} \right\|, \quad \text{a.s.},$$

where $\|\cdot\|$ stands for the L^2 distance, $\boldsymbol{X}^T_{-\infty} = (\ldots, X_{T-1}, X_T)$ and $\boldsymbol{Y}^T_{-\infty} = (\ldots, Y_{T-1}, Y_T)$.

Using the triangular inequality and applying a standard result for the L^2 consistency of kernel estimates, it suffices to show

$$\left\| f_{X^*_{T+h}|\boldsymbol{X}^T_{-\infty}} - f_{Y^*_{T+h}|\boldsymbol{Y}^T_{-\infty}} \right\| \to \left\| f_{X_{T+h}|\boldsymbol{X}^T_{-\infty}} - f_{Y_{T+h}|\boldsymbol{Y}^T_{-\infty}} \right\|, \quad \text{a.s..}$$

Using again the triangular inequality, we just need to prove

$$\left\| f_{X_{T+h}|\boldsymbol{X}^T_{-\infty}} - f_{X^*_{T+h}|\boldsymbol{X}^T_{-\infty}} \right\| \to 0, \text{ and } \left\| f_{Y_{T+h}|\boldsymbol{Y}^T_{-\infty}} - f_{Y^*_{T+h}|\boldsymbol{Y}^T_{-\infty}} \right\| \to 0, \text{ a.s.},$$

that is, we must prove the L^2 consistency of the smoothed sieve bootstrap procedure. We now consider the precise assumptions a stationary process $\{X_t\}_{t\in\mathbb{Z}}$ should meet in order to prove the required consistency:

> *Assumption A1:* $X_t - \mu_X = \sum_{j=0}^{+\infty} \psi_j \varepsilon_{t-j}$, $\psi_0 = 1$ $(t \in \mathbb{Z})$ with $\Psi(z) = \sum_{j=0}^{+\infty} \psi_j z^j$ bounded away from zero for $|z| \leq 1$, and $\sum_{j=0}^{+\infty} j^\beta |\psi_j| < \infty$ for some $\beta > 1$.
>
> *Assumption A2:* $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ are i.i.d., with $\mathrm{E}[\varepsilon_t] = 0$, $\mathrm{E}[|\varepsilon_t|^s] < \infty$ for some $s \geq 1$, and have a twice continuously differentiable density $f_\varepsilon(\cdot)$. Moreover, there is a constant k such that $\int_{-\infty}^{\infty} (f_\varepsilon(x) - f_\varepsilon(x+c))^2 dx \leq kc^2 + o(c^2)$, $\forall c \in \mathbb{R}$.

Additionally, we impose the following assumptions about both the order of the autoregressive approximation, $p(n)$, and the nonparametric estimator of f_ε:

Assumption B: $p = p(n) \to \infty$, $p(n) = o((n/\log(n))^{1/(2\beta+2)})$ for the same β as in *A1*, where n is the sample size.

Assumption C1: The kernel, K, used to estimate f_ε satisfies that: K is a density function with $\int_{-\infty}^{\infty} xK(x)dx = 0$, $\int_{-\infty}^{\infty} x^2K(x)dx \neq 0$, there is a constant k (not necessarily equal to that of *A2*) such that $\int_{-\infty}^{\infty}(K(x) - K(x+c))^2dx \leq kc^2 + o(c^2)$, $\forall c \in \mathbb{R}$ and $\int_{-\infty}^{\infty} |x|^s K(x)dx < \infty$ for the same s as in *A2*.

Assumption C2: The bandwidth satisfies:

$$h = h(n) = o(1),\ h(n)^{-1} = o(n) \text{ as } n \to \infty.$$

$h^{-3}p^{-2\beta} = o(1)$ for the same β as in *A1* and n is the sample size.

Theorem 1 *Suppose that assumptions A1, B and C2 with $\beta > 1$, and A2 and C1 with $s = 1$ hold. Then,*

$$\left\| f_{X_{T+h}|\boldsymbol{X}_{-\infty}^T} - f_{X_{T+h}^*|\boldsymbol{X}_{-\infty}^T} \right\| \to 0, \text{ in probability},$$

where $\boldsymbol{X}_{-\infty}^T = (\ldots, X_{T-1}, X_T)$.

Proof: We can write X_{T+h} and X_{T+h}^* as:

$$X_{T+h} = -\sum_{j=1}^{+\infty} \phi_j X_{T+h-j} + \varepsilon_{T+h}$$

and

$$X_{T+h}^* = -\sum_{j=1}^{+\infty} \widehat{\phi}_{j,n} X_{T+h-j}^* + \varepsilon_{T+h}^*,$$

where $\widehat{\phi}_{j,n}$ denote the estimates of ϕ_j with a sample of size n, $\widehat{\phi}_{j,n} = 0$ for $j > p(n)$, and $X_t^* = X_t$ for $t \leq T$. For simplicity of notation we prove the theorem for $h = 1$. Notice that here we use the notation of Thombs and Schucany (1990) which leaves the last observation X_T fixed and a sample of size n is written as $(X_{T-n+1}, \ldots, X_T)$.

Notice that conditioning X_{T+1} and X_{T+1}^* on $\boldsymbol{X}_{-\infty}^T$ we obtain

$$f_{X_{T+1}|\boldsymbol{X}_{-\infty}^T}(x) = f_{\varepsilon_{T+1}}(x + \sum\nolimits_{j=1}^{+\infty} \phi_j X_{T+h-j})$$

and

$$f_{X_{T+1}^*|\boldsymbol{X}_{-\infty}^T}(x) = f_{\varepsilon_{T+1}^*}(x + \sum\nolimits_{j=1}^{+\infty} \widehat{\phi}_{j,n} X_{T+h-j}).$$

So, we can concentrate our attention in the following expression:

$$\begin{aligned}
&\left\| f_{\varepsilon^*_{T+1}}(x + \textstyle\sum_{j=1}^{+\infty} \widehat{\phi}_{j,n} X_{T+h-j}) \; - \; f_{\varepsilon_{T+1}}(x + \textstyle\sum_{j=1}^{+\infty} \phi_j X_{T+h-j}) \right\| \\
&\qquad \leq \left\| f_{\varepsilon^*_{T+1}}(x + \textstyle\sum_{j=1}^{+\infty} \widehat{\phi}_{j,n} X_{T+h-j}) \; - \; f_{\varepsilon_{T+1}}(x + \textstyle\sum_{j=1}^{+\infty} \widehat{\phi}_{j,n} X_{T+h-j}) \right\| + \\
&\qquad\qquad + \left\| f_{\varepsilon_{T+1}}(x + \textstyle\sum_{j=1}^{+\infty} \widehat{\phi}_{j,n} X_{T+h-j}) \; - \; f_{\varepsilon_{T+1}}(x + \textstyle\sum_{j=1}^{+\infty} \phi_j X_{T+h-j}) \right\| \\
&\qquad = S_1 + S_2.
\end{aligned}$$

For S_1 we apply a change of variable and write $S_1 = \left\| f_{\varepsilon^*_{T+1}} - f_{\varepsilon_{T+1}} \right\|$. Moreover, notice that in smoothed sieve bootstrap the density of bootstrap innovations is, by construction, the kernel density estimate of the estimated innovations, i.e., $f_{\varepsilon^*_{T+1}} = \widehat{f}_{\widehat{\varepsilon}_{T+1}}$. Then,

$$S_1 = \left\| \widehat{f}_{\widehat{\varepsilon}_{T+1}} - f_{\varepsilon_{T+1}} \right\| \leq \left\| \widehat{f}_{\widehat{\varepsilon}_{T+1}} - \widehat{f}_{\varepsilon_{T+1}} \right\| \; + \; \left\| \widehat{f}_{\varepsilon_{T+1}} - f_{\varepsilon_{T+1}} \right\| = I_1 + I_2.$$

The convergence of I_2 is a standard consistency result for kernel density estimators which follows from Hypothesis *C1* and the first part of *C2* and *A2*.

For I_1 we have

$$\begin{aligned}
I_1 \; &= \left\| \tfrac{1}{nh} \textstyle\sum_{t=p+1}^{n} K\left(\tfrac{x-\widehat{\varepsilon}_t}{h}\right) - \tfrac{1}{nh} \textstyle\sum_{t=p+1}^{n} K\left(\tfrac{x-\varepsilon_t}{h}\right) \right\| \\
&= \textstyle\int_{-\infty}^{\infty} \tfrac{1}{n^2h^2} \textstyle\sum_{t=p+1}^{n} \textstyle\sum_{s=p+1}^{n} \left(K\left(\tfrac{x-\widehat{\varepsilon}_t}{h}\right) - K\left(\tfrac{x-\varepsilon_t}{h}\right)\right)\left(K\left(\tfrac{x-\widehat{\varepsilon}_s}{h}\right) - \left(\tfrac{x-\varepsilon_s}{h}\right)\right) dx \\
&\leq \textstyle\int_{-\infty}^{\infty} \tfrac{1}{nh} \textstyle\sum_{t=p+1}^{n} \left(K(y) - K\left(y + \tfrac{\widehat{\varepsilon}_t-\varepsilon_t}{h}\right)\right)^2 dx \\
&\leq \tfrac{1}{nh} \textstyle\sum_{t=p+1}^{n} \left(k\left(\tfrac{\widehat{\varepsilon}_t-\varepsilon_t}{h}\right)^2 + o\left(\left(\tfrac{\widehat{\varepsilon}_t-\varepsilon_t}{h}\right)^2\right)\right),
\end{aligned} \tag{5.4}$$

where the last inequality follows from Hypothesis *C1*. Now, we use the expression (5.6) of Bickel and Bühlmann (1999) that relates each estimated innovation, $\widehat{\varepsilon}_t$, to the corresponding innovation, ε_t:

$$\widehat{\varepsilon}_t = \varepsilon_t + Q_{t,n} + R_{t,n} - (\bar{X} - \mu) \sum_{j=0}^{\infty} \phi_j,$$

where $Q_{t,n} = \sum_{j=0}^{p} (\widehat{\phi}_{j,n} - \phi_{j,n})(X_{t-j} - \bar{X})$, $R_{t,n} = \sum_{j=0}^{\infty} (\phi_{j,n} - \phi_j)(X_{t-j} - \bar{X})$ and $(\phi_{1,n}, \phi_{2,n}, \ldots, \phi_{p,n})$ is the solution of the theoretical Yule–Walker equations.

So,

$$\begin{aligned}
\textstyle\sum_{t=p+1}^{n} (\widehat{\varepsilon}_t - \varepsilon_t)^2 \; &= \textstyle\sum_{t=p+1}^{n} (Q_{t,n} + R_{t,n} - (\bar{X} - \mu) \textstyle\sum_{j=0}^{\infty} \phi_j)^2 \\
&\leq 2 \textstyle\sum_{t=p+1}^{n} (Q_{t,n}^2 + R_{t,n}^2 + (\bar{X} - \mu)^2 (\textstyle\sum_{j=0}^{\infty} \phi_j)^2) = J1 + J2 + J3.
\end{aligned}$$

Under our assumptions *A1* and *A2*, we can use Theorem 3.3 of Phillips and Solo (1992) to obtain:

$$J3 = O_{a.s.}\left(n(n/\log\log n)^{-1}\right) = O_{a.s.}\left(\log\log n\right). \tag{5.5}$$

Under our assumptions *A1* and *A2*, we can use the extended Baxter inequality (cf. Hannan and Deistler, 1988) as in expression (5.8) of Bickel and Bühlmann (1999), to obtain:

$$\mathrm{E}[J2] \leq nc \left(\sum_{j=p+1}^{\infty} \phi_j \right)^2 = o(np^{-2\beta}). \tag{5.6}$$

Under our assumptions *A1* and *A2*, we can use Theorem 2.1 of Hannan and Kavalieris (1986) to obtain:

$$\begin{aligned} J1 &\leq \textstyle\sum_{t=p+1}^{n} \left(\max_{0\leq j\leq p(n)} |\widehat{\phi}_{j,n} - \phi_{j,n}| \sum_{j=0}^{p} |X_{t-j} - \bar{X}| \right)^2 \\ &\leq \max_{0\leq j\leq p(n)} |\widehat{\phi}_{j,n} - \phi_{j,n}|^2 \sum_{t=p+1}^{n} \left(\sum_{j=0}^{p} |X_{t-j} - \bar{X}| \right)^2 \\ &\leq \max_{0\leq j\leq p(n)} |\widehat{\phi}_{j,n} - \phi_{j,n}|^2 \sum_{t=p+1}^{n} p \sum_{j=0}^{p} (X_{t-j} - \bar{X})^2 \\ &\leq \max_{0\leq j\leq p(n)} |\widehat{\phi}_{j,n} - \phi_{j,n}|^2 \, p^2 \sum_{t=1}^{n} (X_t - \bar{X})^2 \\ &= O_{a.s.}\left((n/\log n)^{-1}\right) p^2 O_P(n) = O_P(p^2 \log n). \end{aligned} \tag{5.7}$$

Now, using (5.5)-(5.7) in (5.4), we obtain:

$$I1 = \frac{1}{nh^3} \left(O_P\left(p^2 \log n\right) + o_P(np^{-2\beta}) + O_{a.s.}\left(\log\log n\right) \right),$$

which is $o_P(1)$ using assumption B with $\beta > 1$ and assumption *C2*.

Also, for S_2 we apply a change of variable and write:

$$S_2 = \int_{-\infty}^{\infty} \left(f_{\varepsilon_{T+1}}\left(x + \textstyle\sum_{j=1}^{\infty} (\widehat{\phi}_{j,n} - \phi_j) X_{T+1-j}\right) \; - f_{\varepsilon_{T+1}}(x) \right)^2 dx,$$

and using the last part of assumption *A2*:

$$S_2 \leq k \left(\sum_{j=1}^{\infty} (\widehat{\phi}_{j,n} - \phi_j) X_{T+1-j} \right)^2 + o\left(\left(\sum_{j=1}^{\infty} (\widehat{\phi}_{j,n} - \phi_j) X_{T+1-j} \right)^2 \right).$$

This squared term can be handle in a similar way as $J1$ and $J2$ since

$$\sum_{j=1}^{\infty} (\widehat{\phi}_{j,n} - \phi_j) X_{T+1-j} = \sum_{j=1}^{p} (\widehat{\phi}_{j,n} - \phi_{j,n}) X_{T+1-j} \; + \; \sum_{j=1}^{\infty} (\phi_{j,n} - \phi_j) X_{T+1-j},$$

where $(\phi_{1,n}, \phi_{2,n}, \ldots, \phi_{p,n})$ is the solution of the theoretical Yule–Walker equations, and $\widehat{\phi}_{j,n}$ and $\phi_{j,n}$ are equal to zero if $j > p$. So,

$$S_2 = O_{a.s.}\left((n/\log n)^{-1}\right) O_P(p^2) + O_P\left(p^{-2\beta}\right),$$

and application of assumption B with $\beta > 1$ concludes the proof. ∎

References

Alonso, A.M., Peña, D. and Romo, J. (2002), "Forecasting time series with sieve bootstrap", *Journal of Statistical Planning and Inference*, **100**, 1–11.

Alonso, A.M., Peña, D. and Romo, J. (2003), "On sieve bootstrap prediction intervals", *Statistics & Probability Letters,* **65**, 13–20.

Box, G., Jenkins, G. M. and Reinsel, G.(1994), *Time Series Analysis: Forecasting and Control*, 3rd Edn. Prentice Hall.

Bickel, P.J. and Bühlmann, P. (1999), "A new mixing notion and functional central limits theorems for a sieve bootstrap in time series", *Bernoulli*, **5**, 413–446.

Cao, R., Febrero-Bande, M., González-Manteiga, W., Prada-Sánchez, J.M. and García-Jurado, I. (1997), "Saving computer time in constructing consistent bootstrap prediction intervals for autoregressive processes", *Communications in Statistics - Simulation and Computation*, **26**, 961–978.

Cowpertwait, P.S.P., and Cox, T.F. (1992), "Clustering population means under heterogeneity of variance with an application to a rainfall time series problem", *The Statistician,* **41**, 113–121.

Everitt, B.S., Landau, S. and Leese, M. (2001), *Cluster Analysis*, 4th Edn. London: Arnold.

Fruhwirth-Schnatter, S. and Kaufmann, S. (2004) "Model-based clustering of multiple time series", CEPR Discussion Paper No. 4650.

Galeano, P. and Peña, D. (2000), "Multivariate analysis in vector time series", *Resenhas*, **4**, 383–403.

Gómez, V. and Maravall, A. (1994). "Estimation, prediction, and interpolation for nonstationary series with the Kalman filter", *Journal of the American Statistical Association*, **89**, 611–624.

Gómez, V. and Maravall, A. (1996). "Programs TRAMO (Time Series Regression with ARIMA noise, Missing observations and Outliers) and SEATS (Signal Extraction in ARIMA Time Series). Instruction for the user", Working Paper 9628, Bank of Spain, Madrid.

Gómez, V. and Maravall, A. (2001). "Automatic modeling methods for univariate series", In *A Course in Time Series* (D. Peña, G.C. Tiao and R.S. Tsay eds.), ch. 7, John Wiley & Sons, New York.

Hannan, E.J. and Deistler, M. (1988), *The Statistical Theory of Linear Systems*, John Wiley & Sons Inc: New York.

Hannan, E.J. and Kavalieris, L. (1986), "Regression, autoregression models", *Journal of Time Series Analysis*, **7**, 27–49.

Hurvich, C.M. and Tsai, C.-L. (1989), "Regression and time series model selection in small samples", *Biometrika*, **76**, 297–307.

Kakizawa, Y., Shumway, R.H. and Taniguchi, M. (1998), "Discrimination and clustering for multivariate time series", *Journal of the American Statistical Association*, **93**, 328–340.

Maharaj, E.A. (1996). "A significance test for classifying ARMA models", *Journal of Statistical Computation and Simulation*, **54**, 305–331.

Macchiato, M. F., La Rotonda, L., Lapenna, V. and Ragosta, M. (1995), "Time modelling and spatial clustering of daily ambient temperature: an application in Southern Italy", *Environmetrics*, **6**, 31–53.

Pascual, L., Romo, J. and Ruiz, E. (2004), "Bootstrap predictive inference for ARIMA processes", *Journal of Time Series Analysis*, **25**, 449–465.

Pattarin, F., Paterlini, S. and Minerva, T. (2004). "Clustering financial time series: an application to mutual funds style analysis", *Computational Statistics and Data Analysis*, **47**, 353–372.

Phillips, P.C.B. and Solo, V. (1992), "Asymptotics for linear processes", *The Annals of Statistics*, **20**, 971–1001.

Piccolo, D. (1990), "A distance measure for classifying ARIMA models", *Journal of Time Series Analysis*, **11**, 153–164.

Sheather, S.J., Hettmansperger, T.P. and Donald, M.R. (1994). "Data-based bandwidth selection for kernel estimators of the integral of $f^2(x)$", *Scandinavian Journal of Statistics*, **21**, 265–275.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Thombs, L.A., Schucany, W.R. (1990), "Bootstrap prediction intervals for autoregression", *Journal of the American Statistical Association*, **85**, 486–492.

United Nations (1997), *Kyoto Protocol to the United Nations Framework Convention on Climate Change*. New York.

Cover Letter

Dr. Ana Justel
Departamento de Matemáticas
Universidad Autónoma de Madrid
28049 Madrid, SPAIN
E-mail: ana.justel@uam.es
Phone: (34) 914977634
Fax: (34) 914974889

April 14, 2006

Professor Dr. Erricos J. Kontoghiorghes
Co–Editor of Computational Statistics and Data Analysis
School of Computer Science and Information Systems
University of London

Dear Prof. Dr. Kontoghiorghes,

Thanks for your comments on our paper *Time series clustering based on forecast densities* (ref. CSDA-05123e). Please find enclosed the revised version in which we have addressed all comments from the Associated Editor.

Sincerely,

Ana Justel