Nonlinear time series clustering based on forecast densities

José A. Vilar Fernández

Departamento de Matemáticas, Universidade da Coruña, Spain

eijoseba@udc.es

Juan M. Vilar Fernández

Departamento de Matemáticas, Universidade da Coruña, Spain

eijvilar@udc.es

Andrés M. Alonso Fernández

Departamento de Estadística, Universidad Carlos III de Madrid, Spain

amalonso@est-econ.uc3m.es

Abstract

The problem of clustering time series is studied for a general class of nonparametric autoregressive models. The dissimilarity between two time series is based on comparing their full forecast densities at a given horizon. In particular, two functional distances are considered: L^1 and L^2. As the forecast densities are unknown, they are approximated using a bootstrap procedure that mimics the underlying generating processes without assuming any parametric model for the true autoregressive structure of the series. The estimated forecast densities are then used to construct the dissimilarity matrix and hence to perform clustering. Asymptotic properties of the proposed method are provided and an extensive simulation study is carried out. The results show the good behavior of the procedure for a wide variety of nonlinear autoregressive models and its robustness to non-Gaussian innovations. Finally, the proposed methodology is applied to a real dataset involving economic time series.

JEL codes: C22 and C53

1 Introduction

Time series clustering is aimed at classifying the series under study into homogeneous groups in such a way that the within-group-series similarity is minimized and the between-group-series dissimilarity is maximized. This is a central problem in many application fields and hence, time series clustering is nowadays an active research area in different disciplines including signal processing, finance and economics, medicine, seismology, meteorology and pattern recognition, among others. Some illustrative examples of specific applications reported in the current literature are: clustering of ecological dynamics [Li et al.(2001)], comparison of daily hydrological time series [Grimaldi(2004)], clustering of industrialized countries according to historical data of CO_2 emissions [Alonso et al.(2006)], detection of similar immune response behaviors of CD4+ cell number progression in patients affected by immunodeficiency virus (HIV) [Chouakria-Douzal and Nagabhushan(2007)], and classification of models of industrial production series [Corduas and Piccolo(2008)]. An extensive survey of many other application areas is provided in [Liao(2005)].

According to other clustering problems, the metric chosen to assess the similarity/dissimilarity between two data objects plays a crucial role in time series clustering. However, the concept of dissimilarity between two time series is non-trivial. In fact, in addition to conventional metrics such as the Euclidean distance or the Fréchet distance, other dissimilarity criteria specifically designed to deal with time series have been proposed in the literature. For example, for the class of $ARIMA$ invertible models, [Piccolo(1990)] and [Maharaj(1996)] introduced metrics based on the disparity between the corresponding fitted autoregressive expansions. In the frequency domain, dissimilarity measures based on spectral density estimators were considered in [Kakizawa et al.(1998)], [Taniguchi and Kakizawa(2000)] and [Vilar and Pértega(2004)]. [Caiado et al.(2006)] used the Euclidean distance between the logarithms of the normalized periodograms to discriminate between stationary and non-stationary processes. [Chouakria-Douzal and Nagabhushan(2007)] introduced a new dissimilarity index based on an automatic adaptive tuning function to modulate the temporal correlation between the series and the proximity between their raw values. An interesting overview of dissimilarity criteria between time series can be seen in the review by [Liao(2005)] and in [Corduas and Piccolo(2008)] and references therein.

Conceptually most of the dissimilarity criteria proposed for time series clustering lead to the notion of similarity relying on two possible criteria: (i) proximity between raw series data and (ii) proximity between underlying generating processes. In both cases, the classification task becomes inherently static since similarity searching is governed only by the behavior of the series over their periods of observation. [Alonso et al.(2006)] argue that, in many practical situations, the real interest of clustering is the long term behavior, and in particular, on how the forecasts at a specific horizon can be grouped. For instance, the Kyoto Protocol establishes a reduction of CO_2 emissions by countries in the fixed horizon of the year 2012 and, in this context, the classification of different industrialized countries based on their predictions of CO_2 emissions for 2012 is interesting and was investigated by [Alonso et al.(2006)]. Similar problems where the interest is to reach target values at a pre-specified future time frequently arise by studying economic or financial indicators, sustainable development strategies, etc. For this kind of situation, they propose a dissimilarity measure based on comparing the full forecast densities associated to each series in the sample. Note that comparing the forecast densities instead of the point forecasts can help separate into different clusters times series having similar or equal predictions but different underlying generating models (e.g. models that differ only in the innovations distribution). Furthermore, using the forecast densities allows us to take into account the variability of the predictions, that is completely ignored when the comparison is based on the point forecasts. In practice, the forecast densities are unknown and must be approximated from the data. In [Alonso et al.(2006)], this approximation is constructed using a smoothed sieve bootstrap procedure combined with kernel density estimation techniques. Such a procedure requires the assumption that the considered time series admit an $AR(\infty)$ representation because the sieve bootstrap is based on residual resampling from autoregressive approximations of the processes.

In this paper, the clustering procedure proposed by [Alonso et al.(2006)] is extended to cover the case of nonparametric models of arbitrary autoregressions. Our approach does not assume any parametric model for the true autoregressive structure of the series, which is estimated by using kernel smoothing techniques. As a consequence, only nonparametric approximations to the true autoregressive functions are available in this new setup, and hence, the sieve bootstrap is not a valid resampling procedure. In our procedure, the mechanism used to obtain bootstrap predictions is based on mimicking the generating process using a nonparametric estimator of the autoregressive function and a bootstrap resample of the nonparametric residuals. So, we provide a useful device for classifying nonlinear autoregressive time series, including extensively studied parametric models such as the threshold autoregressive (TAR), the exponential autoregressive ($EXPAR$), the smooth-transition autoregressive ($STAR$) and the bilinear, among others (see e.g. [Tong(1990)] and the references therein).

The rest of the paper is organized as follows. In Section 2, we describe the clustering procedure and outline the resampling mechanism proposed to generate the bootstrap predictions. In Section 3, we establish the consistency of two different dissimilarity measures under appropriate conditions, and hence, our clustering procedure asymptotically leads to the cluster solution based on the true generating processes. Section 4 reports the results from an extensive Monte Carlo simulation designed to study the performance of the proposed clustering procedure and to compare it to two other resampling methods, namely the AR-sieve bootstrap scheme and a conditional version of our initial resampling scheme. In Section 5, we illustrate the performance of our method in a real data example involving economic time series. More specifically the dataset is formed by a collection of series representing the monthly industrial production indices for 21 countries. Technical proofs of the results in Section 3 can be found in Vilar et al. (2009).

2 Description of the clustering procedure

Denote by Ξ the class of real valued stationary processes $\{X_t\}_{t\in\mathbb{Z}}$ that admit a general autoregressive representation of the form

$$X_t = m(\boldsymbol{X}_{t-1}) + \varepsilon_t, \tag{2.1}$$

where $\{\varepsilon_t\}$ is an i.i.d. sequence and $\boldsymbol{X}_{t-1}$ is a d-dimensional vector of known lagged variables. The unknown autoregressive function $m(\cdot)$ is assumed to be a smooth function but it is not restricted to any pre-specified parametric model. Hence, both linear and nonlinear autoregressive processes are included in Ξ.

Our concern is to perform a cluster analysis on a set S of s partial realizations from time series belonging to Ξ, i.e. $S = \left\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(s)}\right\}$, where each element $\mathbf{X}^{(i)} = \left(X_1^{(i)}, \ldots, X_T^{(i)}\right)$ is generated from a process satisfying model (2.1). Following the ideas by [Alonso et al.(2006)], the final goal of our clustering is to capture similarities in the forecasts at a specific future time $T + b$. Indeed, the horizon b is not subjectively chosen, but it is clearly determined by the nature of the problem. In other words, the grouping itself is directly motivated by knowing the behavior of the predicted values at the specific horizon $T+b$. Hence, we adopt the criterion of measuring the dissimilarity between two time series in terms of the disparity between their corresponding full forecast densities at $T+b$. In particular, this disparity is evaluated by using two possible distances. First, we consider the squared L^2 functional distance, i.e. the distance between the time series $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(j)}$, $i, j = 1, \ldots, s$, is defined by

$$D_{2,ij} = \int \left(f_{X^{(i)}_{T+b}}(x) - f_{X^{(j)}_{T+b}}(x) \right)^2 dx, \tag{2.2}$$

where $f_{X^{(i)}_{T+b}}(\cdot)$ denotes the density function of the forecast $X^{(i)}_{T+b}$, with $T+b$ the prefixed prediction horizon.

The distance $D_{2,ij}$ was also used by [Alonso et al.(2006)] due to its computational advantages and its analytical tractability. Nevertheless, $D_{2,ij}$ presents a serious drawback when performing cluster analysis. If the sets $\{x : f_{X^{(i)}_{T+b}}(x) > \varepsilon\}$ and $\{x : f_{X^{(j)}_{T+b}}(x) > \varepsilon\}$ are disjoint for a sufficiently small $\varepsilon > 0$, then

$$D_{2,ij} \approx \int f^2_{X^{(i)}_{T+b}}(x) dx + \int f^2_{X^{(j)}_{T+b}}(x) dx,$$

and hence, $D_{2,ij}$ has a poor performance in the clustering task because it removes the effect of the distance between the point forecasts and it is only governed by the shape of the forecast densities. Therefore, $D_{2,ij}$ distance provides good results when the forecasts at $T+b$ are not very separated, but it is not a useful distance measure to perform clustering in other cases. An interesting alternative is to consider the L^1 functional distance given by

$$D_{1,ij} = \int \left| f_{X^{(i)}_{T+b}}(x) - f_{X^{(j)}_{T+b}}(x) \right| dx. \tag{2.3}$$

Note that if $X^{(i)}_{T+b}$ and $X^{(j)}_{T+b}$ are quite distant, or more precisely, their forecast densities have disjoint supports, then,

$$D_{1,ij} = \int f_{X^{(i)}_{T+b}}(x)\, dx + \int f_{X^{(j)}_{T+b}}(x)\, dx = 2,$$

regardless of the shapes of the densities. Therefore, $D_{1,ij}$ allows to correctly identify the most distant series and leads to a reasonable cluster solution. The behavior of both distances to perform time series clustering is examined and discussed in our experiments in Section 4 and in the analysis of the real data example in Section 5.

Direct computation of distances $D_{u,ij}$, $u = 1, 2$, is not feasible in practice because the forecast densities are unknown. To overcome this difficulty, distances $D_{u,ij}$ are consistently approximated by replacing the unknown forecast densities in (2.2) and (2.3) by kernel-type density estimates constructed on the basis of bootstrap predictions. In particular, we have considered a bootstrap procedure based on generating a process

$$X_t^* = \hat{m}_g(\boldsymbol{X}_{t-1}^*) + \varepsilon_t^*, \tag{2.4}$$

where $\hat{m}_g$ is a nonparametric estimator of m and $\{\varepsilon_t^*\}$ is a conditionally i.i.d. resample from the nonparametric residuals. This bootstrap method, called *autoregression bootstrap*, completely mimics the dependence structure of the underlying process. Actually autoregression bootstrap uses an approach similar to that of the residual-based resampling of linear autoregressions, but it takes advantage of being free of the linearity requirement, and hence, it can be applied to our class Ξ of nonparametric models. Consistency of this bootstrap procedure is established by [Franke et al.(2002)]. A detailed description of the steps involved in generating a set of bootstrap predictions is provided below.

Let $(X_1, \ldots, X_T)$ be a partial realization from a process $X(t) \in \Xi$, i.e. $X(t)$ admits the representation given in (2.1). The resampling scheme proceeds as follows.

1. Estimate the autoregressive function $m(\cdot)$ using a modified Nadaraya-Watson estimator with bandwidth g_1.

   In particular, the following truncated Nadaraya-Watson estimator can be considered.

$$\hat{m}_{g_1}(\boldsymbol{x}) = \begin{cases} \check{m}_{g_1}(\boldsymbol{x}) & \text{if } |\check{m}_{g_1}(\boldsymbol{x})| \leq C_m \left|\overline{X}\right| \\ C_m\overline{X} & \text{if } \check{m}_{g_1}(\boldsymbol{x}) > C_m \left|\overline{X}\right| \\ -C_m\overline{X} & \text{if } \check{m}_{g_1}(\boldsymbol{x}) < -C_m \left|\overline{X}\right| \end{cases} \tag{2.5}$$

   where $\overline{X}$ is the sample mean, C_m is a constant and $\check{m}_{g_1}(\boldsymbol{x})$ is given by

$$\begin{aligned} \check{m}_{g_1}(\boldsymbol{x}) &= \frac{\check{\varphi}_{g_1}(\boldsymbol{x})}{\check{f}_{g_1}(\boldsymbol{x})} \\ &= \left(\frac{1}{n}\sum_{t=d+1}^{T} \boldsymbol{K}_{g_1}(\boldsymbol{x}-\boldsymbol{X}_{t-1})X_t\right)\left(\frac{1}{n}\sum_{t=d+1}^{T} \boldsymbol{K}_{g_1}(\boldsymbol{x}-\boldsymbol{X}_{t-1})\right)^{-1}, \end{aligned}$$

   with $n = T - d$ and $\check{f}_{g_1}$ and $\check{\varphi}_{g_1}$ kernel estimators of f, the density of $\boldsymbol{X}_{t-1}$, and of the function $\varphi = m \cdot f$, respectively. Here $\boldsymbol{K}_{g_1}(\boldsymbol{u}) = \prod_{r=1}^{d} g_1^{-1}K(u_r/g_1)$, with $K(\cdot)$ a univariate kernel function, in general, a symmetric probability density function.

The choice of $\hat{m}_{g_1}(\boldsymbol{x})$ as an initial estimator of $m(\boldsymbol{x})$ ensures the consistency of the autoregression bootstrap when a sufficiently large C_m is taken and K and g_1 satisfy certain regularity conditions [Franke et al.(2002)].

2. Compute the nonparametric residuals, $\widehat{\varepsilon}_t = X_t - \hat{m}_{g_1}(\boldsymbol{X}_{t-1})$, $t = d+1, \ldots, T$.

3. Construct a kernel estimate of the density function, $f_{\tilde{\varepsilon}}$, associated to the centered residuals $\tilde{\varepsilon}_t = \widehat{\varepsilon}_t - \widehat{\varepsilon}_\bullet$ with $\widehat{\varepsilon}_\bullet$ the mean of the $\widehat{\varepsilon}_t$.

   Using the Rosenblatt-Parzen kernel estimator, we obtain

$$\hat{f}_{\tilde{\varepsilon},h}(u) = \int H_h\left(u - v\right) d\,\hat{F}_n(v) = \frac{1}{n} \sum_{t=d+1}^{T} H_h\left(u - \tilde{\varepsilon}_t\right),$$

   where $\hat{F}_n(v) = n^{-1} \sum_{t=d+1}^{T} I(\tilde{\varepsilon}_t \leq v)$ is the empirical distribution function associated to the sample $\{\tilde{\varepsilon}_t : t = d+1, \ldots, T\}$, $H(u)$ is a kernel function, h is the bandwidth and $H_h(u) = h^{-1}H(u/h)$ is the rescaled kernel according to the bandwidth h.

4. Draw a bootstrap-resample ε_k^* of i.i.d. observations from $\hat{f}_{\tilde{\varepsilon},h}$ as follows:

$$\varepsilon_k^* = \hat{F}_n^{-1}(U) + hZ, \quad k = 1, 2, 3, \ldots$$

   where U is a random value from uniform distribution $U(0,1)$ and Z is a random value from a variable with density $H(u)$.

5. Define the bootstrap series X_t^*, $t = 1, \ldots, T$, by the recursion

$$X_t^* = \hat{m}_{g_1}(\boldsymbol{X}_{t-1}^*) + \varepsilon_t^*,$$

   where $\hat{m}_{g_1}$ is defined in Step (2).

6. Estimate the bootstrap autoregressive function, m^*, on the basis of the bootstrap sample $(X_1^*, \ldots, X_T^*)$ obtained in the previous step. Estimation is carried out using again the modified Nadaraya-Watson estimator with bandwidth g_2. The resulting estimator is denoted by $\hat{m}_{g_2}^*$.

7. Compute bootstrap prediction-paths by the recursion

$$X_t^* = \hat{m}_{g_2}^*(\boldsymbol{X}_{t-1}^*) + \varepsilon_t^*,$$

for $t = T+1, T+2, \ldots, T+b$, $b > 0$, where $T+b$ is the horizon pre-selected by the user to carry out the clustering, and $X_t^* = X_t$, for $t \leq T$.

8. Repeat Steps (1)-(7) a large number (B) of times to obtain replications of the b-step-ahead bootstrap future observations.

It is useful to emphasize some remarks on the practical implementation of the resampling method described.

Remark 1. The autoregressive order d must be previously determined to draw out the resampling procedure, or more precisely, a vector $(X_{t-i_1}, \ldots, X_{t-i_d})$ with the lagged variables that contain relevant information on X_t must be previously selected. This is a very important point due to the so-called curse of dimensionality. As it is well-known in nonparametric regression estimation, for a large number of regressors, the estimator becomes inefficient unless the sample size is very large. For this reason, the number of regressor variables should not be too large. There exist several proposals to solve the problem of lag selection: [Vieu(1994)] and [Yao and Tong(1994)] suggested different methods based on cross-validation, [Tjostheim and Auestad(1994)] and [Tscherning and Yang(2000)] proposed a nonparametric version of the final prediction error (FPE) criterion.

Remark 2. If Steps (5) and (6) are omitted and the prediction-paths in Step (7) are computed using $\hat{m}_{g_1}$ instead of $\hat{m}_{g_2}$, then, our resampling plan is a conditional bootstrap procedure (see, v.g., [Cao et al.(1997)]). Both approaches are discussed and compared in the Monte Carlo study of Section 4.

Remark 3. The proposed resampling scheme requires determining up to three bandwidths. An initial bandwidth g_1 is necessary to estimate the regression function $m(\cdot)$ and then to compute the residuals. The cross-validation selector introduced by [Hart(1994)] is specifically designed to deal with dependent data, and hence, this automatic bandwidth selector is a reasonable choice for obtaining g_1. A standard plug-in selector can be used to compute the bandwidth h, which is required to estimate the density $f_{\tilde{\varepsilon}}$. Lastly, a bandwidth g_2 is necessary to estimate the regression m^*. [Franke et al.(2002)] suggest using a bandwidth larger than g_1, such as $g_2 = 1.5g_1$ or $g_2 = 2g_1$. Our numerical study in Section 4 demonstrates that both choices for g_2 work well.

Now, we come back to the clustering procedure. Applying the resampling method to the ith time series under study, $\mathbf{X}^{(i)}$, provides a bootstrap sample $(X_{T+b}^{(i)*1}, X_{T+b}^{(i)*2}, \ldots, X_{T+b}^{(i)*B})$ that allows us to estimate the unknown density of $X_{t+b}^{(i)}$ using kernel estimation techniques. In particular, we consider the Rosenblatt-Parzen kernel smoother to obtain $\hat{f}_{X_{T+b}^{(i)*}}(x)$, the b-step-ahead density estimator at point x for the ith time series, $i = 1, \ldots, s$. Then, distances $D_{u,ij}$, $u = 1, 2$, given in (2.2) and (2.3), can be approximated by the "plug-in" versions

$$\hat{D}_{u,ij}^* = \int \left| \hat{f}_{X_{T+b}^{(i)*}}(x) - \hat{f}_{X_{T+b}^{(j)*}}(x) \right|^u dx, \quad i, j = 1, \ldots, s, \quad u = 1, 2. \tag{2.6}$$

Note that, unlike other dissimilarity measures for time series, distances $\hat{D}_{1,ij}^*$ and $\hat{D}_{2,ij}^*$ can be computed for time series of unequal length. It is also observed that we once again face the problem of choosing a smoothing parameter to compute $\hat{D}_{u,ij}^*$, $u = 1.2$. Here the objective is to minimize the error in the estimation of $D_{u,ij}$. For this reason, in the particular case of $u = 2$, we use the bandwidth proposed by [Sheather et al.(1994)], which is especially designed to estimate the functionals $\int f^2(x)\,dx$. The same approach was considered by [Alonso et al.(2006)]. In Section 3, we state appropriate conditions under which the consistency of the bootstrap distance $\hat{D}_{ij}^*$ as an estimator of D_{ij} holds.

Once the pairwise dissimilarity matrix $\hat{D}_u^* = \left(\hat{D}_{u,ij}^*\right)$ is obtained, a standard clustering algorithm based on D_u^* is carried out. We consider an agglomerative hierarchical clustering method. Application of an agglomerative hierarchical clustering requires us to establish a measurement of proximity between two clusters to determine which groups are to be joined at each step. In particular, our experiments were carried out by using different grouping criteria, including the single linkage, the complete linkage and the average linkage.

3 Asymptotic results

In this section we establish the conditions under which the bootstrap distances $D_{1,ij}^*$ and $D_{2,ij}^*$, given in (2.6), are consistent estimators of $D_{1,ij}$ and $D_{2,ij}$, respectively. First, we consider the precise assumptions a process $\{X_t\}_{t\in\mathbb{Z}}$ should meet to prove the required consistency. Specifically, consistency is obtained in two different situations: first, in the more restrictive case where the processes are assumed to be bounded (Assumption A1(a)), and then, in a more general case (Assumption A1(b)).

Assumption A1. $\{X_t\}_{t\in\mathbb{Z}}$ is a d-order Markovian process, geometrically strong mixing, strictly stationary and satisfies one of the below conditions.

A1(a). $\{X_t\}_{t\in\mathbb{Z}}$ is bounded and the associated process, $Z_t = (\boldsymbol{X}_t, X_{t+1})$, admits functional parameters f and φ continuously differentiable on $C = \overline{\{f > 0\}}$, where $\overline{A}$ denotes the closure of A.

A1(b). The associated process, $Z_t = (\boldsymbol{X}_t, X_{t+1})$, admits functional parameters f and φ satisfying f and $\varphi \in \mathcal{C}_{2,d}(q)$, for some q, and $\mathrm{E}[\exp(a|X_t|^\tau)] < \infty$, for some $a > 0$ and some $\tau > 0$. Here, $\mathcal{C}_{2,d}(q)$ denotes the space of twice continuously differentiable real valued functions g defined on R^d such that $\|g\|_\infty < q$ and $\left\|g^{(2)}\right\|_\infty < q$, $g^{(2)}$ being any partial derivative of order two for g.

Assumption A2. $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ are i.i.d. variables, with $\mathrm{E}\left(\varepsilon_t\right) = 0$, $\mathrm{E}\left(|\varepsilon_t|^s\right) < \infty$ for some $s \geq 1$, and have a twice continuously differentiable density $f_\varepsilon(\cdot)$.
Assumption A3. There is a constant k_1 such that $\int_{-\infty}^{\infty} |f_\varepsilon(x) - f_\varepsilon(x+c)|\, dx \leq k_1\, c$, $\forall c \in \mathbb{R}$.

Additional conditions are also required for the kernel function H and the bandwidth h used to construct the nonparametric estimator of f_ε, and for the kernel function K and the bandwidth g_1 used to estimate $m = \varphi/f$. Specifically, the following conditions are assumed to hold.
Assumption B1. The kernel H is a density function such that $\int_{-\infty}^{\infty} xH(x)\, dx = 0$ and $\int_{-\infty}^{\infty} x^2 H(x)\, dx \neq 0$.
Assumption B2. There is a constant k_2 such that $\int_{-\infty}^{\infty} |H(x) - H(x+c)|\, dx \leq k_2\, c$, $\forall c \in \mathbb{R}$, and $\int_{-\infty}^{\infty} |x|^s H(x)\, dx < \infty$ for the same s as in Assumption A2.
Assumption B3. The bandwidth h satisfies $h = o(1)$ and $h^{-1} = o(n)$, as $n = T - d \to \infty$.
Assumption C1. The kernel K is a density function with $\int_{-\infty}^{\infty} xK(x)\, dx = 0$ and satisfies a Lipshitz condition, i.e. $|K(u) - K(v)| < L\|u - v\|$ for some $L > 0$.
Assumption C2. The bandwidth g_1 satisfies one of the below conditions:

C2(a). $g_1 \simeq \left(\frac{\log(n)^{2-\epsilon}}{n}\right)^{1/(d+2)}$, where $0 < \epsilon < 2$.

C2(b). $g_1 \simeq \left(\frac{\log(n)^{2-1/\tau}}{n}\right)^{1/(d+4)}$, where τ is the same as in Assumption A1(b).

The following condition is necessary to obtain the consistency of $D_{1,ij}^*$ and $D_{2,ij}^*$ in the general case.

Assumption D. There is a sequence, S_n, of regular compact sets and a sequence, $\{\delta_n\}$, of real number satisfying the following conditions:

$$\frac{\delta_n \log(n)^{(1+1/(2\tau))-(1-1/(2\tau))2/(d+4)}}{\beta_n n^{2/(d+4)}} \to 0, \tag{3.7}$$

where β_n is such that $\inf_{x\in S_n} f(x) > \beta_n > 0$, and

$$\delta_n \log(n)^{1/\tau} \Pr(\|X_0\| > \delta(S_n))^{1/4} \to 0, \tag{3.8}$$

where $\delta(S_n)$ denotes the diameter of S_n.

The consistency of $\hat{D}^*_{1,ij}$ as an estimator of $D_{1,ij}$ is established in the following theorem.

Theorem 1 *Let* $\left\{X_t^{(i)}\right\}_{t\in\mathbb{Z}}$, $i = 1, \ldots, s$, *be partial realizations of stochastic processes and suppose that one of the following conditions, (i) or (ii), are fulfilled.*

(i) *Assumptions A1(a), A2, A3, B1, B2, B3, C1 and C2(a) hold and the bandwidth h satisfies* $\left(h\ n^{1/(d+1)}\right)^{-1} \log(n)^{\epsilon+(1-\epsilon)/(d+2)} \leq O(1)$.

(ii) *Assumptions A1(b), A2, A3, B1, B2, B3, C1, C2(b) and D hold and the bandwidth h satisfies* $(h\delta_n)^{-1} \leq O(1)$.

Then,

$$D^*_{1,ij} \to D_{1,ij},\ \textit{in probability, for } i, j = 1, \ldots, s. \tag{3.9}$$

To establish the consistency of $D^*_{2,ij}$ as estimator of $D_{2,ij}$, Assumptions A3 and B2 must be replaced by the following assumptions.

Assumption A3′. There is a constant k_3 such that $\int_{-\infty}^{\infty} (f_\varepsilon(x) - f_\varepsilon(x + c))^2\, dx \leq k_3 c^2 + o(c^2)$, $\forall c \in \mathbb{R}$.

Assumption B2′. There is a constant k_4 such that $\int_{-\infty}^{\infty} (H(x) - H(x + c))^2\, dx \leq k_4 c^2 + o(c^2)$, $\forall c \in \mathbb{R}$, and $\int_{-\infty}^{\infty} |x|^s H(x)\, dx < \infty$ for the same s as in Assumption A2.

Theorem 2 *Let* $\left\{X_t^{(i)}\right\}_{t\in\mathbb{Z}}$, $i = 1, \ldots, s$, *be partial realizations of stochastic processes and suppose that one of the following conditions, (i) or (ii), are fulfilled.*

(i) *Assumptions A1(a), A2, A3′, B1, B2′, B3, C1 and C2(a) hold and the bandwidth h satisfies* $h^{-3} \log(n)^{2\epsilon+2(1-\epsilon)/(d+2)} n^{-2/(d+2)} \leq O(1)$.

(ii) *Assumptions A1(b), A2, A3′, B1, B2′, B3, C1, C2(b) and D hold and the bandwidth h satisfies $h^{-3}\delta_n^{-2} \leq O(1)$.*

Then,

$$D^*_{2,ij} \to D_{2,ij}, \text{ in probability, for } i,j = 1,\ldots,s. \tag{3.10}$$

The proofs of Theorem 1 and Theorem 2 are derived from the validity of the bootstrap procedure and some standard results from nonparametric density estimation theory. In particular, the proof of Theorem 1 is given in the Appendix and similar arguments lead to the proof of Theorem 2.

4 Simulation study

Our clustering procedure is aimed at grouping time series with similar forecast densities at a specific future time $T+b$. Under this similarity principle, the main question about the efficacy of the cluster solution is: how close is the clustering using the bootstrap forecast densities, $\hat{f}_{X^{(i)*}_{T+b}}$, $i = 1,\ldots,s$, to the clustering using the true forecast densities, $f_{X^{(i)}_{T+b}}$, $i = 1,\ldots,s$.

To answer this question it is interesting to analyze the performance with finite samples of the values $d_{u,\mathbf{X}}$, $u = 1,2$, defined by

$$d_{u,\mathbf{X}} = \int \left| \hat{f}_{X^*_{T+b}}(x) - f_{X_{T+b}}(x) \right|^u dx, \text{ u=1,2}, \tag{4.11}$$

where $\mathbf{X} = (X_1,\ldots,X_T)$ is a partial realization of length T from X_t, with X_t an arbitrary process satisfying (2.1), and $\hat{f}_{X^*_{T+b}}(\cdot)$ is the bootstrap forecast density based on $\mathbf{X}$.

The importance of $d_{u,\mathbf{X}}$ relies on that, for a set of series $S = \left\{\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(s)}\right\}$ subjected to cluster analysis, the efficacy of our clustering procedure is basically determined by the closeness to zero of the values $d_{u,\mathbf{X}^{(i)}}$, for $i = 1,\ldots,s$. In fact, if $d_{u,\mathbf{X}^{(i)}}$ is close to zero for all i, then, $D^*_{u,ij}$ is close to $D_{u,ij}$ for all i,j, and hence, the cluster solutions obtained from both dissimilarity matrices should be close as well.

Taking into account the previous arguments, a Monte Carlo study was designed to evaluate the performance with finite samples of the values $d_{u,\mathbf{X}}$ in (4.11), for stationary processes X_t following different autoregressive functions $m(\cdot)$. The details of the Monte Carlo study are given next.

(a) *Autoregressive models*

For the sake of simplicity, our attention was restricted to first order autoregressive models. Here, we present the results attained for the following representative models.

$$\begin{array}{lll}
\text{M1} & \text{AR} & X_t = 0.6X_{t-1} + \varepsilon_t \\
\text{M2} & \text{Bilinear} & X_t = (0.3 - 0.2\varepsilon_{t-1})\, X_{t-1} + 1.0 + \varepsilon_t \\
\text{M3} & \text{EXPAR} & X_t = \big(0.9 \exp\left(-X_{t-1}^2\right) - 0.6\big)\, X_{t-1} + 0.3 + \varepsilon_t \\
\text{M4} & \text{EXPAR} & X_t = \big(0.9 \exp\left(-X_{t-1}^2\right) - 0.6\big)\, X_{t-1} + 1.0\varepsilon_t \\
\text{M5} & \text{SETAR} & X_t = (0.3X_{t-1} - 1.0)\, I\left(X_{t-1} \geq 0.2\right) - \\
 & & \qquad (0.3X_{t-1} + 0.5)\, I\left(X_{t-1} < 0.2\right) + \varepsilon_t \\
\text{M6} & \text{SETAR} & X_t = (0.3X_{t-1} + 1.0)\, I\left(X_{t-1} \geq 0.2\right) - \\
 & & \qquad (0.3X_{t-1} - 1.0)\, I\left(X_{t-1} < 0.2\right) + \varepsilon_t \\
\text{M7} & \text{NLAR} & X_t = 0.7\left|X_{t-1}\right| \left(2 + \left|X_{t-1}\right|\right)^{-1} + \varepsilon_t \\
\text{M8} & \text{STAR} & X_t = 0.8X_{t-1} - 0.8X_{t-1}\left(1 + \exp\left(-10X_{t-1}\right)\right)^{-1} + \varepsilon_t
\end{array}$$

In all cases, the ε_t are i.i.d. zero-mean random variables with variance σ^2.

Except for model M1, an $AR(1)$ process with moderate autocorrelation, the selected models form a wide class of parametric nonlinear models frequently used to characterize the conditional mean $m(\cdot)$. Model M2 is a bilinear process with approximately quadratic conditional mean, and thus, strongly nonlinear. Models M3 and M4 are exponential autoregressive models with a more complex nonlinear structure although in both cases very close to linearity. Models M5 and M6 are self-exciting threshold autoregressive models with a relatively strong nonlinearity, particularly M6. Finally, Models M7, a general nonlinear autoregressive model, and M8, a smooth transition autoregressive model, present a weak nonlinear structure.

These models have already been used in previous Monte Carlo studies in the literature. For instance, models M2-M6 were considered by [Luukkonen et al.(1988)] and the rest by [Giordano et al.(2007)].

(b) *Distributions for innovations*

Three different distributions were considered for the error processes: Gaussian innovations with unit variance, Student-t innovations with 3 degrees of freedom, and centered exponential $Exp(1)-1$ innovations. The reason for using these different distributions is to analyze the performance of our clustering procedure when both kurtosis or skewness are present.

(c) *Forecast horizons*

The forecast densities are estimated at a specific forecast horizon $T+b$, with b denoting the number of steps-ahead. Here, we present the results for $b=1$ (short term), $b=3$ (intermediate term) and $b=10$ (long term).

(d) *Resampling methods to generate bootstrap predictions*

The experiment was carried out with three different resampling methods. First, we used the *autoregression bootstrap* (AB) outlined in Steps (1)-(8) of Section 2. The kernel estimators were always constructed by using a Gaussian kernel. Bandwidths g_1, g_2 and h were determined as suggested in Remark 3, namely, g_1 in Step (1) by the cross-validation method adapted for dependent data [Hart(1994)], h in Step (3) by using the plug-in selector introduced by [Sheather and Jones(1991)] and g_2 was taken to be larger than g_1, $g_2 = 1.5g_1$ or $g_2 = 2g_1$.

Another resampling mechanism was the *conditional bootstrap* (CB) mentioned in Remark 2 of Section 2. The conditional bootstrap works as the autoregression bootstrap, but Steps (5) and (6) are omitted and the bootstrap prediction-paths in Step (7) are obtained by the recursion

$$X_t^* = \hat{m}_{g_1}(\boldsymbol{X}_{t-1}^*) + \varepsilon_t^*, \tag{4.12}$$

with $X_t^* = X_t$, for $t \leq T$. Thus, this procedure does not draw bootstrap replicates of the original sample but only constructs future bootstrap realizations. In contrast to the autoregression bootstrap, this method focuses on replicating the conditional distribution of X_{T+b} given the observed sample $(X_1, X_2, \ldots, X_T)$. In fact, this is the reason why this procedure is called conditional bootstrap [Cao et al.(1997)].

The third approach was the *smoothed sieve bootstrap* (SB) considered by [Alonso et al.(2006)]. In this case, an $AR(\infty)$ representation is assumed for the underlying process, and hence, the strategy is to generate residuals from estimated linear autoregressive models of order $p = p(T)$. These residuals together with the estimated models allow us to obtain the bootstrap replicates. Indeed, it is expected that this procedure will fail with nonlinear underlying models.

(e) *The simulation mechanism*

For each of the considered models, we simulated one thousand time series of length $T = 200$. With every simulated series $\mathbf{X}$, the b-step-ahead forecast density $f_{X_{T+b}}(\cdot)$ was approximated using the autoregression bootstrap $\hat{f}^{AB}_{X^*_{T+b}}(\cdot)$, the conditional bootstrap $\hat{f}^{CB}_{X^*_{T+b}}(\cdot)$, the sieve bootstrap

$\hat{f}^{SB}_{X^*_{T+b}}(\cdot)$, and lastly Monte Carlo forecasts instead of bootstrap predictions, $\hat{f}^{MC}_{X^*_{T+b}}(\cdot)$. For each simulated series, $B = 1000$ replicates were drawn to estimate each bootstrap density. Note that the Monte Carlo replicates are obtained using the true underlying model (autoregression model and innovations distribution), and therefore $\hat{f}^{MC}_{X^*_{T+b}}(\cdot)$ can be considered as a benchmark in our experiment. So, the behavior with finite samples of the values $d_{u,\mathbf{X}}$, $u = 1, 2$ in (4.11) can be studied by analyzing the feasible values of

$$d^{\bullet}_{u,\mathbf{X}} = \int \left| \hat{f}^{\bullet}_{X^*_{T+b}}(x) - \hat{f}^{MC}_{X^*_{T+b}}(x) \right|^u dx, \quad u = 1, 2, \tag{4.13}$$

with $\bullet$ being some of the considered bootstrap procedures, namely AB, CB or SB. All the computations are implemented in the R language [R Development Core Team(2004)] and the code is available upon request from the authors.

Figure 1 shows the results with Gaussian innovations by using the L^1 distance $d^{\bullet}_{1,\mathbf{X}}$. Each panel corresponds to a different horizon, $b = 1$, 3, 10, and contains the boxplots constructed with the 1000 values of d^{SB}_1, d^{AB}_1 and d^{CB}_1 for each model. The results obtained by using the L^2 distance $d^{\bullet}_{2,\mathbf{X}}$ are shown in Figure 2.

Several conclusions are drawn from Figures 1 and 2. First, as expected, similar results were achieved with both distances, $d^{\bullet}_{1,\mathbf{X}}$ and $d^{\bullet}_{2,\mathbf{X}}$. Note that all models in our simulation have the same mean so the point forecasts are always very close. Thus, $d^{\bullet}_{2,\mathbf{X}}$ is not affected by the drawback pointed out in Section 2.

Concerning the bootstrap procedures, it is observed that the sieve bootstrap is clearly the best procedure when the data are generated from the linear model M1. Indeed, this is not surprising because the sieve bootstrap relies on the linear approximation. The other two bootstrap approaches also work reasonably well by presenting small error rates. We also simulated $AR(1)$ processes with higher autocorrelations, 0.75 and 0.9, and similar results were obtained.

On the contrary, when the data generating process presents a moderate or strong nonlinearity, the autoregression bootstrap and the conditional bootstrap outperform the sieve bootstrap. This fact is shown for models M2-M6, and in particular a huge improvement is observed for models with a strongly nonlinear conditional mean, namely M2, M4 and M6. The behavior of the three bootstrap approaches is somewhat similar for models M7 and M8 (in this latter case only in the short time $b = 1$), but this is justified because in both cases the nonlinear structure is weak and the data generating process can be well approximated by a linear process.

As for the comparison between the autoregression bootstrap and the conditional bootstrap, the performance of both procedures is rather similar although, for some models, a slight improvement seems to be observed when the conditional bootstrap is used. Let us recall that the conditional bootstrap does not draw bootstrap replicates of the original series (Steps (5) and (6) of the resampling algorithm are eliminated) and therefore, this procedure is free of the variability generated by the estimation of the conditional mean in Step (6). It is worthwhile stressing that both the autoregression bootstrap and the conditional bootstrap lead to very similar results for all the models considered in our study. This is a very interesting robustness property with respect to the data generating model.

In general, previous considerations are valid for the three prediction horizons. Except for models M2, M6 and M8, the values of $d_u^{\bullet}$, $u = 1, 2$, tend to decrease for larger horizons regardless of the bootstrap method. The nonlinearity of models M2, M6 and M8 is mainly revealed at lags of order higher than one and for this reason the sieve bootstrap behaves poorly for these specific models. Nonparametric AB and CB procedures are not affected by this feature and therefore they provide better results.

The experiments conducted to investigate the sensitivity of our procedure to skewness and kurtosis led to analogous conclusions. For example, Figure 3 shows the boxplots obtained with exponential innovations for prediction horizons $b = 1$ and $b = 3$ using the distance $d_{1,\mathbf{X}}^{\bullet}$. The corresponding boxplots with Student-t innovations are shown in Figure 4. Compared with the Gaussian case (Figure 1), distributions of d_1^{SB}, d_1^{AB} and d_1^{CB} present larger bias and more skewness to the right, especially with exponential innovations and for horizon $b = 1$. In any case, a good performance of d_1^{AB} and d_1^{CB} is again observed. Similar graphs are obtained with $d_{2,\mathbf{X}}^{\bullet}$ and therefore our simulations seem to confirm the robustness of the procedure with respect to the deviations from normality.

5 A case-study with real data

In this section we perform clustering on a real data example involving economic time series. The dataset consists of a collection of time series representing the monthly industrial production indices (seasonally adjusted) for 21 countries from January 1990 to November 2007. All the considered countries are members of Organization for Economic Cooperation and Development (OECD), and in particular, this dataset is available from the Statistics Portal of OECD (http://stats.oecd.org/). Graphs of these single series are depicted in the panel of Figure 5. Our purpose is to classify the 21 countries in accordance with the performance of their industrial production indices on next month, so the short term $b = 1$ was considered as the forecast horizon of interest.

Note that the time series under study are clearly nonstationary. Hence, our clustering procedure cannot be directly applied because the bootstrap predictions are computed under stationary assumption. So, we proceeded as follows. First, each of the time series was transformed using logarithms (if required) and taking an appropriate number of regular differences. In particular, the program TRAMO (*T*ime series *R*egression with *A*RIMA noise, *M*issing observations and *O*utliers) developed by [Gómez and Maravall(1996)] was used to determine the order of regular differences and to test for the log transformation. The bootstrap prediction-paths for the transformed series were constructed following Steps (1)-(8) in Section 2. Then, the resulting bootstrap prediction-paths were backtransformed to obtain the bootstrap predictions for the original series. From this point on, our procedure was carried out as described in Section 2.

It is also worth pointing out that some of the series are nonlinear. In order to test the linearity hypothesis, all the transformed series were subjected to the McLeod-Li test [McLeod and Li(1983)] based on the empirical r-order autocorrelations of the squared residual of the best autoregressive fit, for a wide range of values for r. The test led to rejection of linearity at level 0.05 for the series of the following countries: *Austria*, *Belgium*, *Canada*, *Hungary*, *Luxembourg*, *Mexico*, *Poland*, *Turkey* and *United Kingdom.* This fact justifies the application of our clustering procedure because it performs well with nonlinear series (as shown in the simulation study).

Before running the clustering procedure, we carried out a standard univariate clustering based on the last available observation of each series. This is not an adequate approach for performing time series clustering because the past history of the series is not taken into account, but this is a standard procedure of official statistical institutions. Therefore, this univariate clustering was only performed for comparison purposes. Figure 6 shows the resulting dendrogram using the average linkage method as the criterion of proximity between groups.

At the end of the agglomerative process, the two-cluster solution clearly identifies a group $\mathcal{C}_1$ formed by the 14 countries with the lowest indices and another $\mathcal{C}_2$ formed by the 7 countries with the highest. If a finer partition identifying more compact groups is required, then $\mathcal{C}_1$ and $\mathcal{C}_2$ can be split into the groups described in Table 1.

$\mathcal{C}_1$				$\mathcal{C}_2$	
$\mathcal{C}_{1,1}$	$\mathcal{C}_{1,2}$	$\mathcal{C}_{1,3}$	$\mathcal{C}_{1,4}$	$\mathcal{C}_{2,1}$	$\mathcal{C}_{2,2}$
United Kingdom	*Portugal*	*United States*	*Sweden*	*Finland*	*Poland*
Italy	*Canada*	*Mexico*	*Luxembourg*	*Austria*	*Korea*
Norway	*Greece*	*Spain*	*Germany*	*Ireland*	*Turkey*
	France		*Belgium*		*Hungary*

Table 1: Six-cluster solution for the hierarchical clustering based on the last observation of each series.

Then, we carried out our clustering procedure using the conditional bootstrap as resampling procedure for generating the bootstrap predictions. The dendrograms obtained using the distances $d^{\bullet}_{1,\mathbf{X}}$ and $d^{\bullet}_{2,\mathbf{X}}$ are shown in Figures 7 and 8, respectively.

Classification of the series based on the L^1 distance (Figure 7) is similar to the one based on the last available observation of each series (Figure 6). In both cases, $\mathcal{C}_1$ and $\mathcal{C}_2$ are identified at the end of the process, and in particular, $\mathcal{C}_1$ is formed by the groups $\mathcal{C}_{1,i}$, $i = 1, \ldots, 4$, with the same membership. However, some important differences are observed when the agglomeration processes for the two procedures are compared. Note that our dissimilarity measure modified the order in which some series are joined. In $\mathcal{C}_{1,2}$, for instance, *Mexico* and *Spain* are now grouped together first, joined later by *United States*. Analogously, in $\mathcal{C}_{1,4}$, *Belgium* joins {*Sweden*, *Luxembourg*} first, and *Germany* later. Also some groups are joined in different order. So, $\mathcal{C}_{1,4} =$ {*Sweden*, *Luxembourg*, *Germany*, *Belgium*} is now closer to $\{\mathcal{C}_{1,2}, \mathcal{C}_{1,3}\}$ than the cluster formed by the three countries with the lowest indices, $\mathcal{C}_{1,1} =$ {*Norway*, *Italy*, *United Kingdom*}. More substantial differences are observed in the nested arrangement of series in $\mathcal{C}_2$. Here the series are even regrouped in a different way. Note that, for example, *Turkey* and *Hungary*, two countries grouped together at an early stage of the process in Figure 6, are now in different clusters in Figure 7. Furthermore, with our procedure *Ireland* is the closest neighbor to *Turkey*, in contrast to the dendrogram in Figure 6 where *Ireland* and *Turkey* appear in different clusters.

Indeed, all of these differences rely on the comparison of the estimated forecast densities. For clarity, the estimated forecast densities have been displayed in Figure 9. The same color was used to depict the forecast densities forming one homogeneous cluster when d_1^{CB} was used as dissimilarity measure. The last observation of each series was also indicated at the bottom of Figure 9 to illustrate the changes in the distances. For instance, if the last observations are taken into account, *Mexico* is closer to *United States* than *Spain*, but this order is reversed when the forecast densities are compared. Analogously, the similarity between the forecast densities of *Ireland* and *Turkey* justifies their proximity, while, on the other hand, *Hungary* is farther from these countries due to the high kurtosis of its density. It is also observed that the forecast densities of *Finland* and *Austria* appear to be practically isolated in Figure 9 and for this reason they tend to remain isolated until very late in the agglomerative process in Figure 7. This kind of properties justify the use of the full forecast densities instead of the point forecasts, as already pointed out by [Alonso et al.(2006)].

The classification based on the L^2 distance (Figure 8) led to a disappointing result. For instance, at an early stage of the clustering process, *Norway*, the country with the lowest industrial production index in our dataset, is wrongly joined to the group formed by $\mathcal{C} = \{$ *Turkey*, *Ireland*, *Poland*, *Korea* $\}$, which is obviously inappropriate because these four countries present the highest indices. As discussed in Section 2, the reason for this bad performance is that the estimated forecast density of any of the countries in $\mathcal{C}$ and the estimated density of *Norway* have disjoint supports. In these cases, $d^{\bullet}_{2,\mathbf{X}}$ ignores the distance between the point forecasts and is governed by the similarity between the densities shapes. In fact, Figure 9 shows that the forecast densities of the five mentioned countries are similar and because of this they have been grouped together. Analogously, the forecast densities of *United States*, *Italy*, *United Kingdom* and *Canada* are characterized by presenting the highest kurtosis (see again Figure 9), and hence they are very close to each other but they are the farthest from the remaining series. In summary, the L^2 distance is not valid to cluster the series under study.

References

[Alonso et al.(2006)] Alonso, A.M., Berrendero, J.R., Hernández, A., Justel, A., 2006. Time series clustering based on forecast densities. Comput. Statist. Data Anal. 51, 762-776.

[Bosq(1998)] Bosq, D., 1998. Nonparametric statistics for stochastic processes. Lecture Notes in

Statistics 110. Springer-Verlag, New York.

[Caiado et al.(2006)] Caiado, J., Crato, N., Peña, D., 2006. A periodogram-based metric for time series classification. Comput. Statist. Data Anal. 50, 2668-2684.

[Cao et al.(1997)] Cao, R., Febrero-Bande, M., González-Manteiga, W., Prada-Sánchez, J.M., García-Jurado, I., 1997. Saving computer time in constructing consistent bootstrap prediction intervals for autoregressive processes. Comm. Statist. Simulation Comput. 26, 961–978.

[Corduas and Piccolo(2008)] Corduas, M., Piccolo, D., 2008. Time series clustering and classification by the autoregressive metric. Comput. Statist. Data Anal. 52, 1860–1872.

[Chouakria-Douzal and Nagabhushan(2007)] Chouakria-Douzal, A., Nagabhushan, P.N., 2007. Adaptive dissimilarity index for measuring time series proximity. Adv. Data Anal. Classif. 1, 5–21.

[Devroye and Györfi(1985)] Devroye, L. and Györfi, L., 1985. Nonparametric density estimation: The L_1 view. John Wiley and Sons, New York.

[Franke et al.(2002)] Franke, J., Kreiss, J.P., Mammen, E., 2002. Bootstrap of kernel smoothing in nonlinear time series. Bernoulli 8, 1–37.

[Giordano et al.(2007)] Giordano, F., La Rocca, M., Perna, C., 2007. Forecasting nonlinear time series with neural network sieve bootstrap. Comput. Statist. Data. Anal. 51, 3871–3884.

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

[Grimaldi(2004)] Grimaldi, S., 2004. Linear parametric models applied on daily hydrological series. J. Hydrol. Eng. 9, 383-391.

[Hart(1994)] Hart, J., 1994. Automated kernel smoothing of dependent data by using time series cross-validation. J. Roy. Statist. Soc. Ser. B 81, 1080–1088.

[Kakizawa et al.(1998)] Kakizawa, Y., Shumway, R.H., Taniguchi, M., 1998. Discrimination and clustering for multivariate time series. J. Amer. Statist. Assoc. 93, 328–340.

[Li et al.(2001)] Li, C., Biswas, G., Dale, M., Dale, P., 2001. Building models of ecological dynamics using HMM based temporal data clustering–a preliminary study. In: Hoffmann, F. et al. (Eds.), IDA 2001, Lecture Notes in Comput. Sci. 2189: pp. 53–62.

[Liao(2005)] Liao, T.W., 2005. Clustering of time series data: a survey. Pattern Recognit. 38, 1857-1874.

[Luukkonen et al.(1988)] Luukkonen, R., Saikkonen, P., Terärsvirta, T., 1988. Testing linearity in univariate time series. Scand. J. Statist. 15, 161–175.

[McLeod and Li(1983)] McLeod, A.I., Li, W.K., 1983. Diagnostic checking of ARMA time series models using squared-residual autocorrelations. J. Time Series Anal. 4, 269–273.

[Maharaj(1996)] Maharaj, E.A., 1996. A significance test for classifying ARMA models. J. Statist. Comput. Simulation 54, 305-331.

[Piccolo(1990)] Piccolo, D., 1990. A distance measure for classifying arima models. J. Time Series Anal. 11, 153-164.

[R Development Core Team(2004)] R Development Core Team, 2004. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

[Sheather and Jones(1991)] Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. J. Roy. Statist. Soc. Ser. B, 53, 683-690.

[Sheather et al.(1994)] Sheather, S.J., Hettmansperger, T.P., Donald, M.R., 1994. Data-based bandwidth selection for kernel estimators of the integral of $f^2(x)$. Scand. J. Statist. 21, 265-275.

[Taniguchi and Kakizawa(2000)] Taniguchi, M., Kakizawa, Y., 2000. Asymptotic Theory of Statistical Inference for Time Series. Springer, New York.

[Thombs and Schucany(1990)] Thombs, L.A., Schucany, W.R., (1990). Bootstrap prediction intervals for autoregression. J. Amer. Statist. Assoc., 85, 486-492.

[Tjostheim and Auestad(1994)] Tjostheim, D., Auestad, B.H., 1994. Nonparametric identification of nonlinear time series: selecting significant lags. J. Amer. Statist. Assoc. 89, 1410–1419.

[Tong(1990)] Tong, H., 1990. Non-linear Time Series: A Dynamical System Approach. Oxford Statistical Science Series, 6.

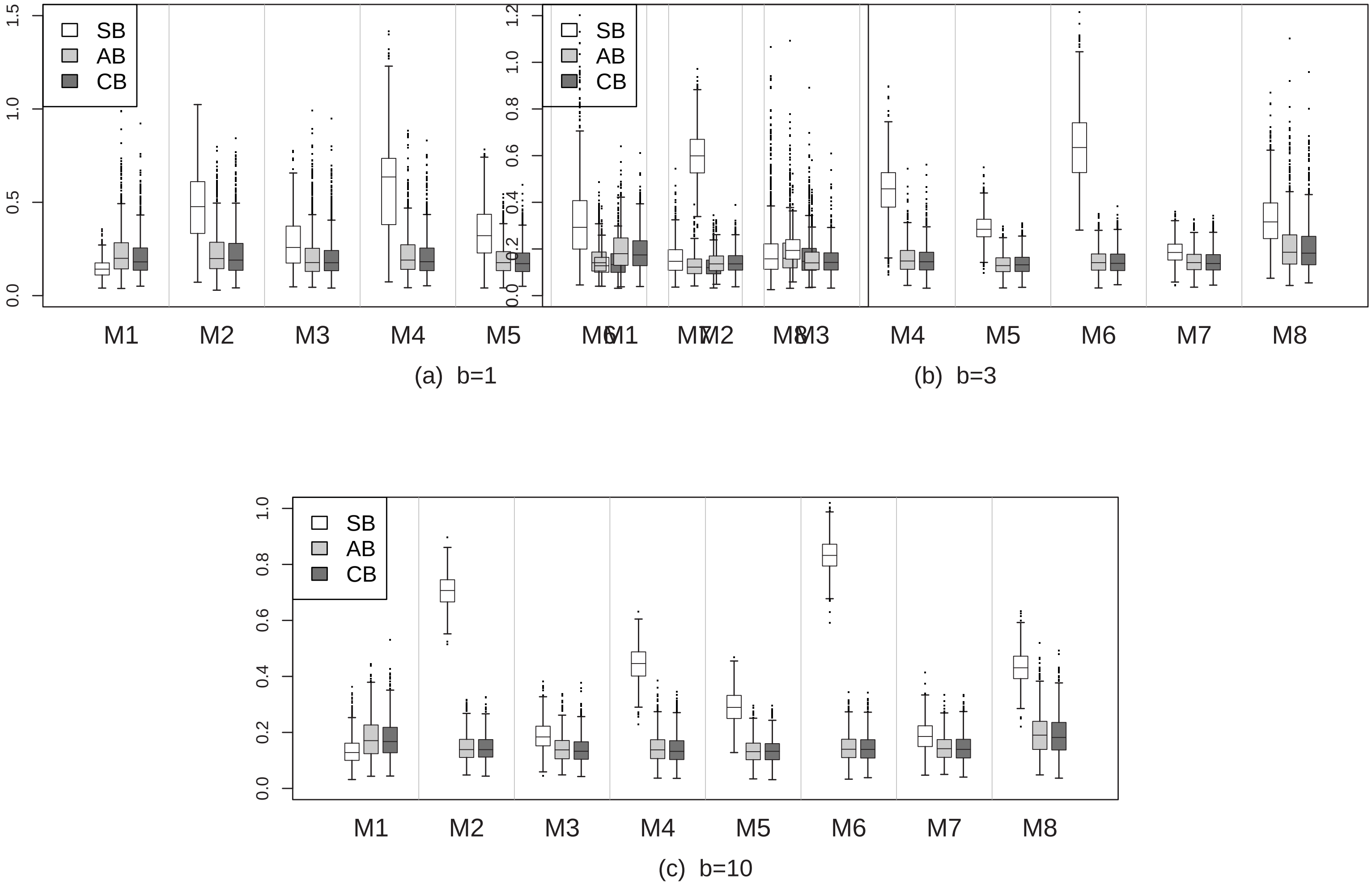
[Tscherning and Yang(2000)] Tscherning, R., Yang, L., 2000. Nonparametric lag selection for time series. J. Time Series Anal. 21, 457–487.

[Vieu(1994)] Vieu, P., 1994. Order choice in nonlinear autoregressive models. Statistics 24, 1–22.

[Vilar et al.(2009)] Vilar, J.A., Alonso, A.M., Vilar, J.M., 2009. Nonlinear time series clustering based on nonparametric forecast densities. Comput. Statist. Data Anal. (*in press*).

[Vilar and Pértega(2004)] Vilar, J.A., Pértega, S., 2004. Discriminant and cluster analysis for Gaussian stationary processes: local linear fitting approach. J. Nonparametr. Stat. 16, 443-462.

[Yao and Tong(1994)] Yao, Q., Tong, H., 1994. On subset selection in non-parametric stochastic regression. Statist. Sinica 4, 51–70.

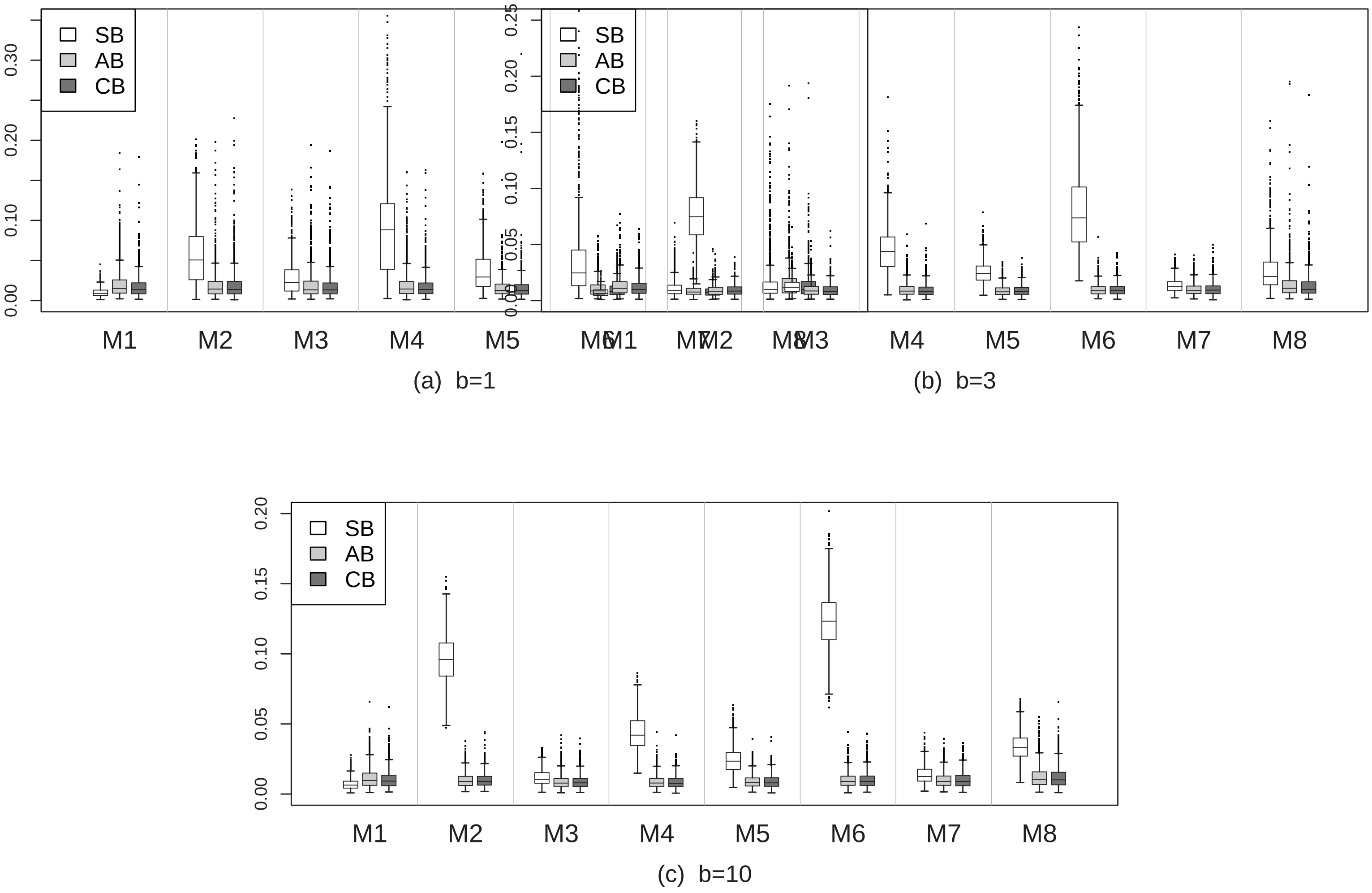


(a) b=1

(b) b=3

(c) b=10

Figure 1: Simulation results for Gaussian innovations and the L^1 distance $d^{\bullet}_{1,\mathbf{X}}$. Boxplots of the values d_1^{SB}, d_1^{AB} and d_1^{CB} for prediction horizons: $b = 1$ (a), $b = 3$ (b) and $b = 10$ (c).

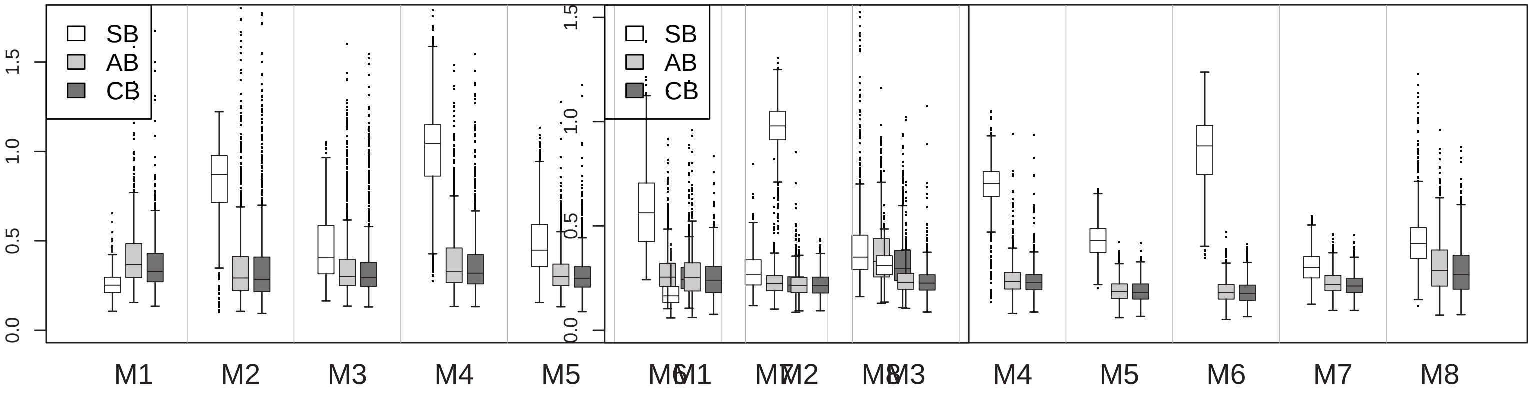


(a) b=1

(b) b=3

(c) b=10

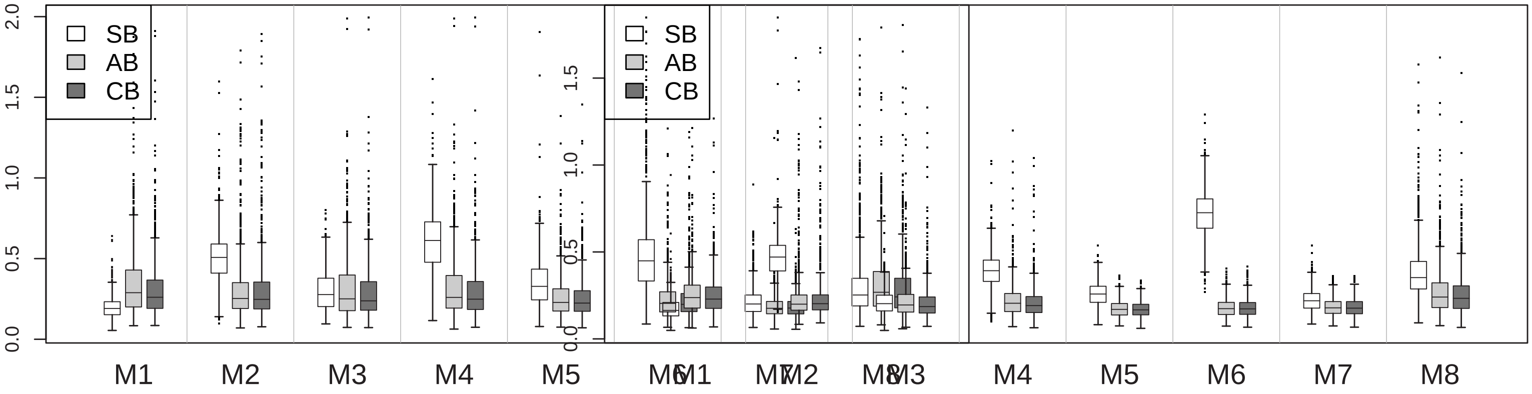
Figure 2: Simulation results for Gaussian innovations and the L^2 distance $d^{\bullet}_{2,\mathbf{X}}$. Boxplots of the values d_2^{SB}, d_2^{AB} and d_2^{CB} for prediction horizons: $b = 1$ (a), $b = 3$ (b) and $b = 10$ (c).



(a) b=1

(b) b=3

Figure 3: Simulation results for centered exponential innovations and the L^1 distance $d_{1,\mathbf{X}}^{\bullet}$. Boxplots of the values d_1^{SB}, d_1^{AB} and d_1^{CB} for prediction horizons: $b = 1$ (a) and $b = 3$ (b).



(a) b=1

(b) b=3

Figure 4: Simulation results for Student-t innovations innovations and the L^1 distance $d_{1,\mathbf{X}}^{\bullet}$. Boxplots of the values d_1^{SB}, d_1^{AB} and d_1^{CB} for prediction horizons: $b = 1$ (a) and $b = 3$ (b).

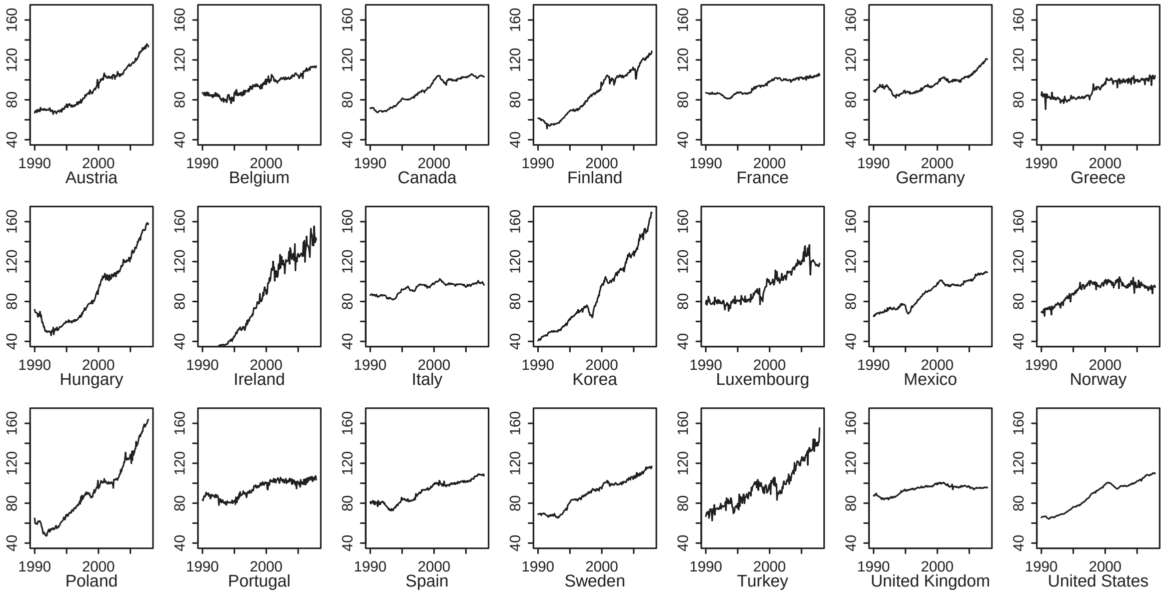


Figure 5: Monthly industrial production indices of 21 countries: January 1990 – November 2007.

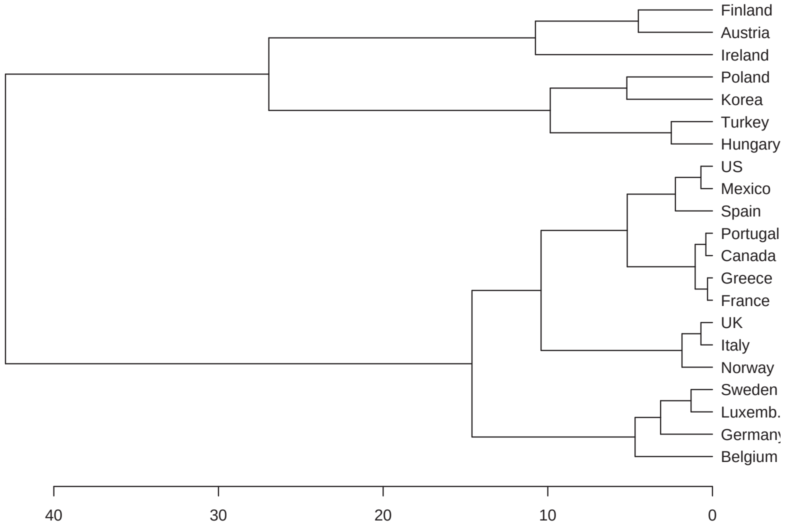


Figure 6: Dendrogram based on absolute differences of the last observation of each series (November 2007) with the average linkage method.

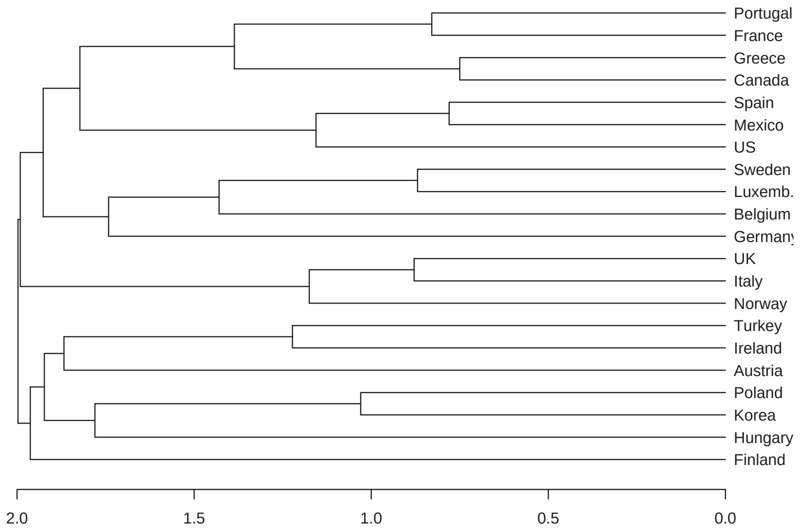


Figure 7: Dendrogram based on our clustering procedure for the prediction horizon $b = 1$ month. The conditional bootstrap, the L^1 distance $d^{\bullet}_{1,\mathbf{x}}$ and the average linkage method were used.

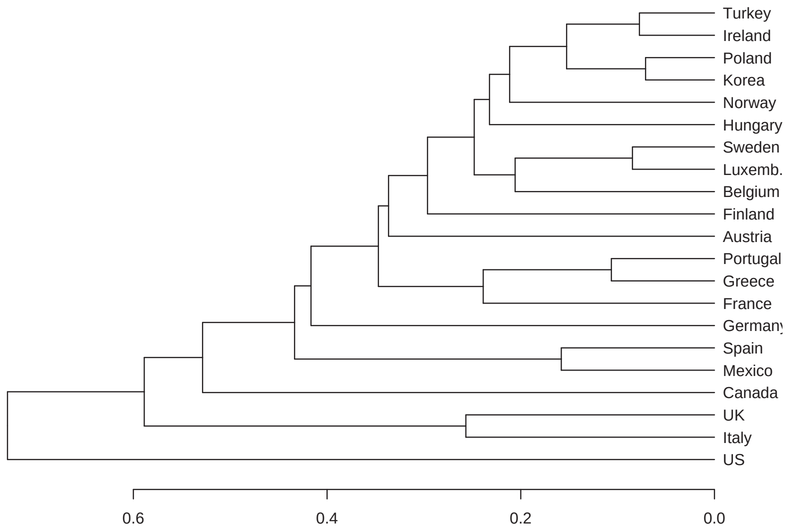


Figure 8: Dendrogram based on our clustering procedure for the prediction horizon $b = 1$ month. The conditional bootstrap, the L^2 distance $d^{\bullet}_{2,\mathbf{X}}$ and the average linkage method were used.

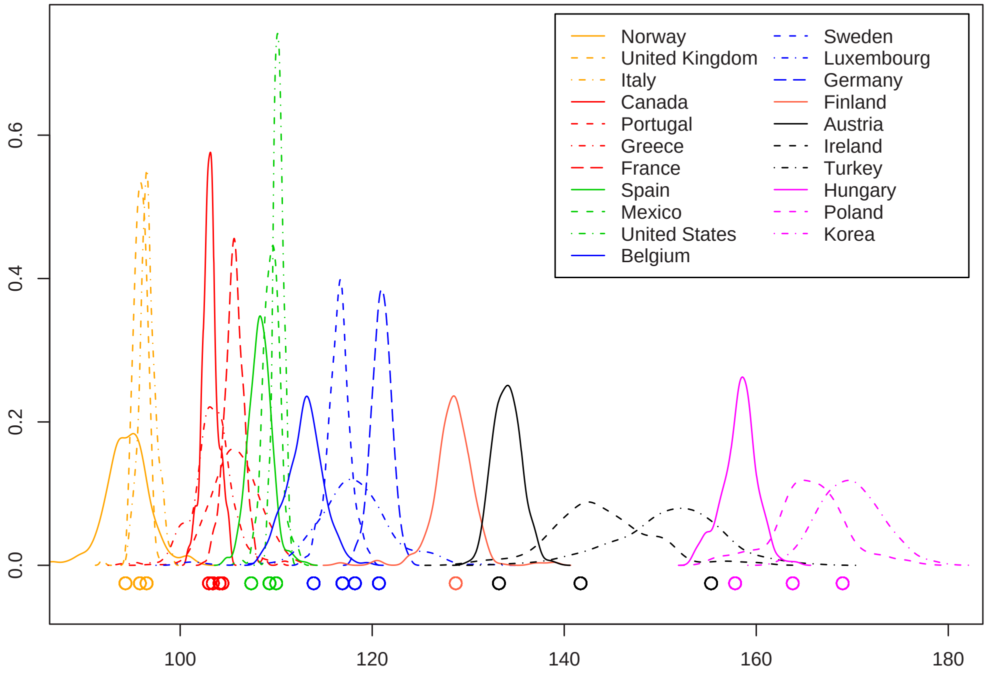


Figure 9: Kernel approximations to the forecast densities based on conditional bootstrap samples of size 1000. Circles at the bottom depict the last available observation of each series.