Communications of the Korean Statistical Society
2010, Vol. 17, No. 3, 377–390

Classification of Time-Series Data Based on Several Lag Windows

Hee-Young Kim[a], Man Sik Park[1,b]

[a]Institute of Economics, Korea University
[b]Department of Statistics, Sungshin Women's University

Abstract

In the case of time-series analysis, it is often more convenient to rely on the frequency domain than the time domain. Spectral density is the core of the frequency-domain analysis that describes autocorrelation structures in a time-series process. Possible ways to estimate spectral density are to compute a periodogram or to average the periodogram over some frequencies with (un)equal weights. This can be an attractive tool to measure the similarity between time-series processes. We employ the metrics based on a smoothed periodogram proposed by Park and Kim (2008) for the classification of different classes of time-series processes. We consider several lag windows with unequal weights instead of a modified Daniel's window used in Park and Kim (2008). We evaluate the performance under various simulation scenarios. Simulation results reveal that the metrics used in this study split the time series into the preassigned clusters better than do the raw-periodogram based ones proposed by Caiado *et al.* (2006). Our metrics are applied to an economic time-series dataset.

Keywords: Clustering, autoregressive model, moving-average model, smoothed periodogram, nonstationary time series.

1. Introduction

Clustering time-series data has been tried in various scientific research fields. For example, in medicine, the functional maps of human brain activities receiving a stimulus are obtained using functional magnetic resonance imaging(fMRI) data (Golay *et al.*, 1998; Wismüller *et al.*, 2002; Goutte *et al.*, 1999). Kakizawa *et al.* (1998) and Shumway (2003) studied seismic phenomena by earthquakes and mining explosions using clustering and discrimination analysis. An issue in environmental studies is to group different locations into some clusters, each of which have similar time-series behavior (Macchiato *et al.*, 1995; Cowpertwait and Cox, 1992).

The first step to cluster time-series data is to clarify how similarity (or dissimilarity) between time series can be measured. Many metrics that measure the similarity between time series have been proposed in the last two decades. There are two different approaches for identifying the similarity, as reviewed by Corduas and Piccolo (2008).

The first approach is to compare time series according to the definition of similarity measures between their underlying processes based on some distributional (parametric) assumptions. Piccolo (1990) considered the Euclidean distance measures between the autoregressive expansions based on a class of autoregressive integrated moving-average(ARIMA) invertible models. Kakizawa *et al.* (1998) proposed dissimilarity measures, J-divergence and Chernoff information divergence, between spectral matrices corresponding to the matrices of autocovariance functions of two zero-mean vector stationary

[1] Corresponding author: Assistant Professor, Department of Statistics, Sungshin Women's University, 249-1, Dongseon-dong 3-Ga, Seongbuk-Gu, Seoul 136-742, Korea. E-mail: mansikpark@sungshin.ac.kr

time series. Maharaj (2000) proposed a hypothesis testing method for the comparison of two stationary time series based on the autoregressive parameters and introduced a clustering method using the p-values obtained from the test. Shumway (2003) employed modified versions of Kullback-Leibler discrimination information in order to measure the dissimilarity between non-stationary time series.

The second approach focuses on nonparametric techniques. These have been of less interest owing to the difficulties of defining reasonable distances between time-series sequences. One possible distance is the Euclidean metric, which is invariant to reordering the observations and, consequently does not take into account the correlation structures inherent in a time-series sequence (see Galeano and Pẽna, 2000). Bohte *et al.* (1980) and Kovačić (1996) considered distance metrics based on autocorrelation and cross-correlation structures of the compared time series. Caiado *et al.* (2006) proposed new metrics based on the raw periodogram Fourier-transformed from autocovariance functions. Using simulation, they compared the raw-periodogram based metrics with the ones proposed by Piccolo (1990) and the ones based on coefficients of autocorrelation, partial autocorrelation, and inverse autocorrelation. Caiado *et al.* (2006) shows that a metric based on the logarithm of the normalized periodogram and a metric based on autocorrelation coefficients distinguish autoregressive moving-average(ARMA) models from ARIMA models with high success rate while neither the Euclidean metric nor the metric by Piccolo (1990) perform very well.

In terms of the second approach, the work most relevant to this study has been done by Caiado *et al.* (2006). They developed new metrics based on the raw periodogram. The raw periodogram is not a good estimator of spectral density function in that the variance of each periodogram ordinate does not decrease as the sample size increases. Chen *et al.* (1994) showed that the lag window was used to estimate the degree of differencing in autoregressive fractionally integrated moving-average(ARFIMA) models. Park and Kim (2008) considered the time-series clustering based on the smoothed-periodogram based metrics. However, they merely focused on the real application and used was only one simple type of lag window called modified Daniell window. Since Priestley (1981) proved that a smoothed periodogram depends on the type of lag windows, we employ the metrics based on smoothed periodograms from several lag windows with unequal weights and different lag widths, and evaluate their performances in the classification of time series by comparing with the metrics proposed by Caiado *et al.* (2006).

This paper is organized as follows. In Section 2, we review the current dissimilarity metrics as well as the metrics based on smoothed periodogram from several lag windows with unequal weights. The performance of the smoothed-periodogram based metrics are evaluated using simulation studies in Section 3. In Section 4, our metrics are applied to the same real dataset analyzed in Caiado *et al.* (2006). Finally, we present some conclusions and remarks in Section 5.

2. Dissimilarity Metrics on the Frequency Domain

In this section, we briefly explain some fundamental concepts of periodogram, summarize some current dissimilarity measures defined on the frequency domain. Most of these were proposed by Caiado *et al.* (2006). We also briefly review smoothed-periodogram based on metrics proposed by Park and Kim (2008).

Suppose that $\{X_t; t \in \mathbb{Z}\}$ is a real-valued stationary process with the autocovariance function denoted by

$$\gamma_{\mathbf{x}}(h) = E\left(X_{t+h}X_t\right) - E\left(X_{t+h}\right)E\left(X_t\right)$$

satisfying $\sum_{h=-\infty}^{\infty} |\gamma_{\mathbf{x}}(h)| < \infty$, where $\mathbb{Z}$ is an integer space. With the assumption, the spectral density of $\{X_t\}$ is obtained as, for all $\omega \in [-\pi, \pi]$,

$$f_{\mathbf{x}}(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma_{\mathbf{x}}(h) \exp(-i\omega h). \tag{2.1}$$

One of the notable characteristics of spectral density, $f_{\mathbf{x}}(\omega)$ is that it completely captures the second order dependence structures of the stationary process $\{X_t\}$. The autocovariance function, $\gamma_{\mathbf{x}}(h)$ is calculated from its spectral density, $f_{\mathbf{x}}(\omega)$ using the following equation

$$\gamma_{\mathbf{x}}(h) = 2 \int_0^{\pi} f_{\mathbf{x}}(\omega) \cos(\omega h) \; d\omega.$$

The basic tool to estimate spectral density is the periodogram. For a stationary time-series process with finite length, $\{X_t; t = 1, \ldots, n\}$, the periodogram at the Fourier frequency $\omega_j = 2\pi j/n$ for $j \in \mathbb{F}_n$, is defined as

$$I_n^{\mathbf{x}}\left(\omega_j\right) = \frac{1}{n} \left| \sum_{t=1}^{n} X_t \exp\left(-it\omega_j\right) \right|^2,$$

where $\mathbb{F}_n = \{j \in \mathbb{Z}; -\pi < \omega_j \leq \pi\} = \{-\lfloor (n-1)/2 \rfloor, \ldots, \lfloor n/2 \rfloor\}$. Here $\lfloor a \rfloor$ is the largest integer less than or equal to a. One of the reasons why the periodogram is regarded as an instrument to estimate spectral density is that the periodogram is completely related to the sample autocovariance, $\widehat{\gamma}_{\mathbf{x}}(k)$ for $|k| < n$ as follows:

$$I_n^{\mathbf{x}}\left(\omega_j\right) = \begin{cases} n \left| \dfrac{1}{n} \displaystyle\sum_{t=1}^{n} X_t \right|^2, & \text{if } \omega_j = 0, \\ \displaystyle\sum_{|k|<n} \widehat{\gamma}_{\mathbf{x}}(k) \exp\left(-ik\omega_j\right), & \text{otherwise.} \end{cases} \tag{2.2}$$

From (2.1) and (2.2), $(2\pi)^{-1} I_n^{\mathbf{x}}(\omega_j)$ is a nonparametric estimator of spectral density $f_{\mathbf{x}}\left(\omega_j\right)$ for $\omega_j \neq 0$.

The raw (or unsmoothed) periodogram has the weakness of being asymptotically unbiased but inconsistent. This means that the variance of a raw periodogram does not decrease as the sample size n increases (Brockwell and Davis, 1991). A solution to this inconsistency is to smooth it. We apply a lag window $W_n(k)$ to the sample autocovariance to obtain a smoothed periodogram, *i.e.*,

$$\widehat{f}_{\mathbf{x}}\left(\omega_j\right) = \frac{1}{2\pi} \sum_{k=-M}^{M} W_n(k) \widehat{\gamma}_{\mathbf{x}}(k) \exp\left(-ik\omega_j\right), \tag{2.3}$$

where M is the truncation point that depends on the sample size n. The reader is referred to the textbooks of Brockwell and Davis (1991), Brillinger (1981) and Priestley (1981) for their detailed examination of lag window spectral density estimators.

Given a set of two stationary time-series sequences, $\mathbf{x} = \{x_1, \ldots, x_n\}$ and $\mathbf{y} = \{y_1, \ldots, y_n\}$, the comparison of the dissimilarity measures of $\mathbf{x}$ and $\mathbf{y}$ defined on the frequency domain is conducted by Caiado *et al.* (2006). Before introducing the metrics in Caiado *et al.* (2006) and explaining the ones based on the smoothed periodogram in (2.3), in order for the measures to be easily understood, we express a metric, $d_A(\mathbf{x}, \mathbf{y})$ as

$$d_A(\mathbf{x}, \mathbf{y}) = \left[\sum_{j=1}^{\lfloor n/2 \rfloor} \left\{ A_n^{\mathbf{x}}(\omega_j) - A_n^{\mathbf{y}}(\omega_j) \right\}^2 \right]^{1/2},$$

where $A_n^{\mathbf{x}}(\omega_j)$, for example, is either a raw (or smoothed) periodogram or a transformed one at the frequency ω_j for the sequence $\mathbf{x}$.

Caiado *et al.* (2006) proposed the following classification metrics: $d_I(\mathbf{x},\mathbf{y})$, $d_{NI}(\mathbf{x},\mathbf{y})$, $d_{LNI}(\mathbf{x},\mathbf{y})$. These were compared with other metrics such as Euclidean distance, Piccolo's distance (1990), ACF distance and PACF distance. The metric, $d_{LNI}(\mathbf{x},\mathbf{y})$ is, for example, defined as

$$d_{LNI}(\mathbf{x},\mathbf{y}) = \left[\sum_{j=1}^{\lfloor n/2 \rfloor} \left\{\ln NI_n^{\mathbf{x}}(\omega_j) - \ln NI_n^{\mathbf{y}}(\omega_j)\right\}^2\right]^{1/2},$$

where $NI_n^{\mathbf{x}}(\omega_j)$ is the normalized periodogram of series $\mathbf{x}$ at the frequency of ω_j. This means that it is the raw periodogram, $I_n^{\mathbf{x}}(\omega_j)$ divided by sample variance, that is, $NI_n^{\mathbf{x}}(\omega_j) = I_n^{\mathbf{x}}(\omega_j)/\widehat{\gamma}_{\mathbf{x}}(0)$. Caiado *et al.* (2006) also proposed the Kullback-Leibler (K-L) information metric (see Kullback and Leibler, 1951; Kullback, 1978, for an introduction) defined as

$$d_{KL}(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{\lfloor n/2 \rfloor}\left[\frac{NI_n^{\mathbf{x}}(\omega_j)}{NI_n^{\mathbf{y}}(\omega_j)} - \ln\frac{NI_n^{\mathbf{x}}(\omega_j)}{NI_n^{\mathbf{y}}(\omega_j)} - 1\right].$$

Since $d_{KL}(\mathbf{x},\mathbf{y}) \neq d_{KL}(\mathbf{y},\mathbf{x})$, we use the following metric adjusting for asymmetry of K-L information metric:

$$d_{KD}(\mathbf{x},\mathbf{y}) = d_{KL}(\mathbf{x},\mathbf{y}) + d_{KL}(\mathbf{y},\mathbf{x}),$$

where $d_{KD}(\mathbf{x},\mathbf{y})$ is called K-L divergence metric in this study.

The smoothed-periodogram based metrics proposed by Park and Kim (2008) are as follows; $d_I^*(\mathbf{x},\mathbf{y})$, $d_{NI}^*(\mathbf{x},\mathbf{y})$, $d_{LNI}^*(\mathbf{x},\mathbf{y})$ and $d_{KD}^*(\mathbf{x},\mathbf{y})$, which are the counterparts of the corresponding metrics based on raw periodogram. The metric, $d_{NI}^*(\mathbf{x},\mathbf{y})$ is defined by

$$\begin{aligned} d_{NI}^*(\mathbf{x},\mathbf{y}) &= \left[\sum_{j=1}^{\lfloor n/2 \rfloor}\left\{N\widetilde{I}_n^{\mathbf{x}}(\omega_j) - N\widetilde{I}_n^{\mathbf{y}}(\omega_j)\right\}^2\right] \\ &= \left[\sum_{j=1}^{\lfloor n/2 \rfloor}\left\{\frac{\widetilde{I}_n^{\mathbf{x}}(\omega_j)}{\widehat{\gamma}_{\mathbf{x}}(0)} - \frac{\widetilde{I}_n^{\mathbf{y}}(\omega_j)}{\widehat{\gamma}_{\mathbf{y}}(0)}\right\}^2\right]^{1/2}, \end{aligned}$$

where $\widetilde{I}_n^{\mathbf{x}}(\omega_j)$ is defined as

$$\widetilde{I}_n^{\mathbf{x}}\left(\omega_j\right) = \sum_{k=-M}^{M} W_n(k)\widehat{\gamma}_{\mathbf{x}}(k)\exp\left(-ik\omega_j\right). \tag{2.4}$$

$\widetilde{I}_n^{\mathbf{x}}(\omega_j)$ is the spectral density estimator for the sequence $\mathbf{x}$, $\widehat{f}_{\mathbf{x}}(\omega_j)$ in (2.3) multiplied by 2π.

The lag window $W_n(k)$ in (2.4) that is the best for this kind of clustering time-series data has not been studied. As mentioned above, Park and Kim (2008) only considered a modified Daniell window which is defined by

$$W_n(k) = \begin{cases} \dfrac{1}{2(M-1)}, & \text{if } |k| = M, \\ \dfrac{1}{M-1}, & \text{if } |k| < M, \\ 0, & \text{if } |k| > M. \end{cases}$$

This lag window has equal weights except the first and the last. In this study, we used various lag windows with unequal weights: Bohman, Bartlett, Parzen, rectangular, and Tukey-Hanning windows. However, using rectangular window and Tukey-Hanning window may lead to negative values of $\widetilde{I}_n^{\mathbf{x}}(\omega_j)$ for some ω_j, so $d^*_{LNI}(\mathbf{x},\mathbf{y})$ and $d^*_{KD}(\mathbf{x},\mathbf{y})$ are not obtained. In Section 3 and 4, we assay the following three lag windows:

1. Bohman window

$$W_n(k) = \begin{cases} \left(1 - \dfrac{|k|}{M}\right)\cos\left(\pi\dfrac{|k|}{M}\right) + \dfrac{1}{\pi}\sin\left(\pi\dfrac{|k|}{M}\right), & \text{if } |k| \le M, \\ 0, & \text{if } |k| > M. \end{cases}$$

2. Bartlett window

$$W_n(k) = \begin{cases} 1 - \dfrac{|k|}{M}, & \text{if } |k| \le M, \\ 0, & \text{if } |k| > M. \end{cases}$$

3. Parzen window

$$W_n(k) = \begin{cases} 2\left(1 - \dfrac{|k|}{M}\right)^3, & \text{if } \dfrac{M}{2} \le |k| \le M, \\ 1 - 6\left(\dfrac{|k|}{M}\right)^2 + 6\left(\dfrac{|k|}{M}\right)^3, & \text{if } |k| \le \dfrac{M}{2}, \\ 0, & \text{if } |k| > M. \end{cases}$$

In this section, we reviewed some current dissimilarity measures proposed by Caiado *et al.* (2006), and explained the metrics based on the smoothed periodogram and several lag windows with unequal weights.

3. Simulation Studies

In this section, we examine the performance of the measures shown above for the classification of time-series data under different simulation scenarios. In order to compare the raw-periodogram based metrics with the smoothed-periodogram based ones, we first generated 12 time-series processes with the preassigned parameter values listed in Table 1 under each scenario. Then we calculated raw periodogram and smoothed periodograms with the three windows shown in Section 2.

Finally, we compute the metrics and conduct classification analysis using the complete linkage hierarchical algorithm. In this section, we evaluate the performance of the dissimilarity measures by means of the empirical probability of the correct classification. The sample sizes of 100, 200 and 500 are used for each simulation scenario and the number of Monte-Carlo simulations set to 1000. The truncation point (or the width of a lag window), M for employing a lag window is set to 5%, 10% and 20% of the sample size. These are chosen based on the practical guidelines of Chatfield (1975, p.141). We also considered the metrics based on autocorrelation functions(ACF) and partial autocorrelation functions(PACF) respectively, defined as

$$d_{ACF}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{l=1}^{L}\left(\hat{\rho}_l^{\mathbf{x}} - \hat{\rho}_l^{\mathbf{y}}\right)^2} \quad \text{and} \quad d_{PACF}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{l=1}^{L}\left(\hat{\phi}_{ll}^{\mathbf{x}} - \hat{\phi}_{ll}^{\mathbf{y}}\right)^2},$$

Table 1: The scenarios conducted in simulation studies.

Scenario	Model	Parameters	
	(p, d, q)	Autoregressive (ϕ)	Moving Average (θ)
1	$(1, 0, 0)$	$0.10, 0.20, 0.40, 0.60, 0.80, 0.90$	
	$(0, 0, 1)$		$0.10, 0.20, 0.40, 0.60, 0.80, 0.90$
2	$(1, 0, 0)$	$0.40, 0.50, 0.60, 0.70, 0.80, 0.90$	
	$(0, 0, 1)$		$0.40, 0.50, 0.60, 0.70, 0.80, 0.90$
3	$(1, 0, 0)$	$0.80, 0.82, 0.84, 0.86, 0.88, 0.90$	
	$(1, 0, 0)$	$0.60, 0.62, 0.64, 0.66, 0.68, 0.70$	
4	$(1, 0, 0)$	$0.65, 0.70, 0.75, 0.80, 0.85, 0.90$	
	$(1, 0, 0)$	$0.10, 0.15, 0.20, 0.25, 0.30, 0.35$	
5	$(1, 0, 0)$	0.90	
	$(2, 0, 0)$	$\phi_1 = 0.95, \phi_2 = -.10$	
	$(1, 0, 1)$	0.95	0.10
	$(1, 0, 1)$	$-.10$	$-.95$
	$(0, 0, 1)$		$-.90$
	$(1, 0, 2)$		$\theta_1 = -.95, \theta_2 = -.10$
	$(1, 1, 0)$	0.10	
	$(0, 1, 0)$		
	$(0, 1, 1)$		0.10
	$(0, 1, 1)$		$-.10$
	$(1, 1, 1)$	0.10	$-.10$
	$(1, 1, 1)$	0.05	$-.05$

Notes: 12 time-series models were generated for each scenario; Scenario 5 was considered by Caiado *et al.* (2006).

where $\{\hat{\rho}_l\}$ are the sample autocorrelation coefficients and $\{\hat{\phi}_{ll}\}$ are the sample partial autocorrelation coefficients. The number of lags, L is set to 25% of the sample size throughout Section 3 and 4. The statistical program R (R Development Core Team, 2006) was used to conduct the simulation studies and the real application.

Before explaining the simulation results, it may be more helpful to address the need for the classification of time-series data. Suppose we have moderate number of observed time-series processes. Then it might not be difficult to analyze each of the processes by means of the time-domain approaches and to make some reasonable clusters by comparing the parameter estimates. However, when many processes are given, it is practically impossible to construct models one by one in order to classify the data. That is why we focus on the time-series clustering based on the frequency domain, where we can capture the underlying characteristics inherent in each of the processes in terms of a spectral density function.

The simulation results are explained in Table 2, Table 3 and Table 4, each of which displays empirical percentages of success in classifying the two true clusters of time-series models under certain scenario, that is, how many times out of 1000 each measure divides the 12 time-series processes into the two true clusters correctly. Before explaining the simulation results for each scenario we discuss some general characteristics of the metrics considered in this study. Any smoothed-periodogram based metrics outperform the corresponding raw-periodogram based ones. Among the raw-periodogram based metrics, the LNI metric is superior to the others and the performance of the LNI and the KD metrics improves as n increases. Among the smoothed-periodogram based metrics, the LNI or the KD metrics can be regarded as the best and their empirical percentages are also proportional to n except for Scenario 1 for the KD metric. When stationary time-series models are considered, the performance of the smoothed-periodogram based metrics under Bartlett window is mostly inversely-proportional to M and the performance of the smoothed NI metric worsens as M increases for any lag windows. The PACF metric is much better than the ACF one. Under Scenario 5 (where the stationary and the

Table 2: Empirical percentages of success on the classification under Scenarios 1 and 2

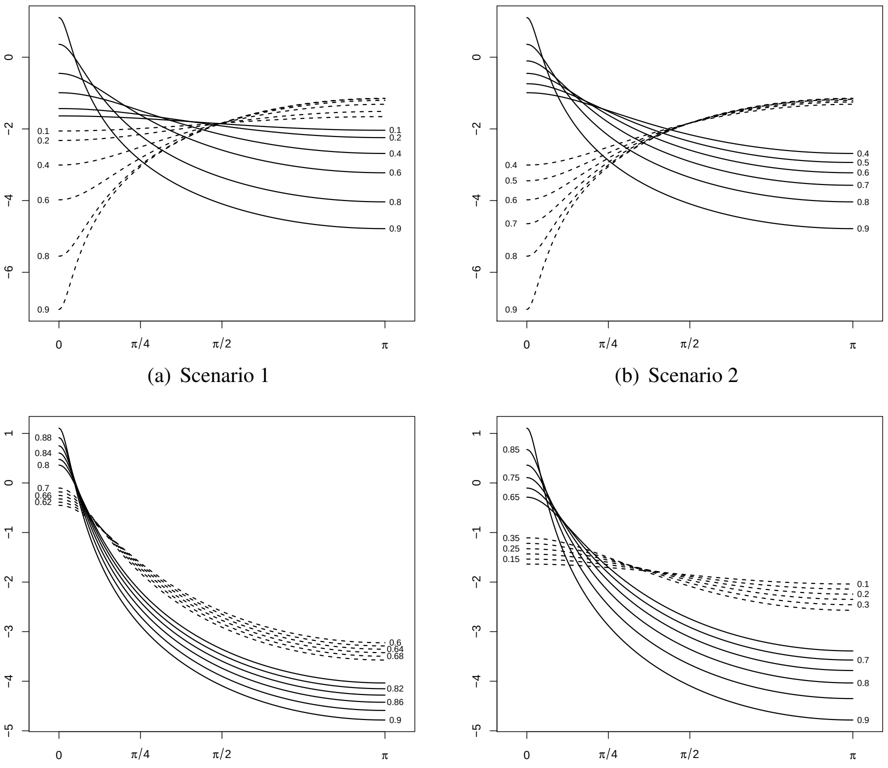
				Smoothed periodogram									Correlation	
	n	metric	Raw†	Bohman window			Bartlett window			Parzen window			ACF	PACF
				5%	10%	20%	5%	10%	20%	5%	10%	20%		
1	100	I	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.1	0.1	0.0	2.3	20.8
		NI	0.8	21.8	10.6	5.1	15.4	9.4	6.0	22.3	11.9	5.8		
		LNI	7.1	8.6	10.6	11.6	17.4	15.2	13.3	8.3	11.4	11.2		
		KD	4.3	8.4	11.7	10.9	17.3	15.8	14.3	7.6	11.9	11.4		
	200	I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	24.7
		NI	0.0	11.3	2.0	1.0	5.6	2.4	1.5	12.8	2.4	1.2		
		LNI	7.6	15.4	14.7	14.3	17.9	18.3	18.0	15.0	15.1	14.8		
		KD	4.2	16.9	13.8	10.1	22.1	22.4	17.7	17.8	14.3	11.1		
	500	I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	30.5
		NI	0.0	0.2	0.0	0.0	0.2	0.1	0.0	0.1	0.0	0.0		
		LNI	9.9	17.8	16.6	15.7	19.1	18.0	17.3	17.1	16.9	16.0		
		KD	4.2	10.4	6.3	6.1	20.5	16.6	11.3	10.8	6.8	6.3		
2	100	I	0.7	5.9	3.4	2.5	3.5	2.8	2.8	5.9	3.6	2.5	26.0	98.2
		NI	14.7	98.5	87.9	65.2	95.8	82.8	67.6	98.5	91.9	68.7		
		LNI	86.5	92.6	93.9	95.0	98.4	97.5	96.7	91.8	93.5	94.7		
		KD	41.1	92.4	96.1	97.0	98.7	98.2	97.7	90.6	95.7	97.3		
	200	I	0.0	0.8	0.3	0.3	0.4	0.3	0.4	0.9	0.3	0.4	16.9	99.9
		NI	8.0	92.5	58.6	35.9	82.9	57.9	39.5	94.5	62.9	37.6		
		LNI	96.5	97.7	98.7	99.2	99.3	99.2	99.2	97.6	98.9	99.2		
		KD	57.6	99.4	99.3	99.3	99.7	99.7	99.8	99.3	99.4	99.4		
	500	I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	9.0	100.0
		NI	1.9	44.4	23.7	17.1	43.1	27.9	19.6	48.4	24.4	17.3		
		LNI	99.6	99.8	100.0	99.9	100.0	100.0	100.0	99.8	100.0	100.0		
		KD	62.5	100.0	99.8	99.6	100.0	100.0	100.0	100.0	99.8	99.5		

Notes: n is the sample size; Raw† denotes the raw periodogram; ACF denotes the autocorrelation; PACF denotes the partial autocorrelation.

nonstationary models are considered) the LNI and the KD metrics with Bohman or Parzen windows perform the best and the ACF metric is superior to the PACF one.

The simulation results under the first two scenarios are displayed in Table 2. First, when the AR processes and the MA ones are considered (Scenario 1 and 2), any smoothed-periodogram based metrics outperform raw-periodogram based ones except for the I metrics, whose percentages are too low to compare to each other. Now we give a full detail of the results from Scenario 1. The NI metric with Parzen (or Bohman) window and $M = 5\%$ is superior to the others when $n = 100$, whereas the LNI or the KD metrics with Bartlett window and $M = 5\%$ or 10% perform the best when $n \geq 200$. The performance of any smoothed-periodogram based metrics largely depends on M. As M decreases, that is, as a periodogram becomes smoothed, any smoothed-periodogram based metric tends to classify the true clusters more correctly especially when $n \geq 200$. For the correlation-based metrics, the PACF metric splits the generated processes into the two true clusters much better than the ACF one, and its performance becomes remarkable as n increases. The smoothed NI metric for $n = 100$ and the PACF metric for $n \geq 200$ are the best among all the metrics considered. Overall performances under Scenario 1 are not quite good in that it is not practically easy to discriminate between the AR process with $\phi = 0.10$ and the MA one with $\theta = 0.10$. This can be easily found from Figure 1(a).

We now discuss the results of Scenario 2, under which, as shown in Figure 1(b), the two clusters are more separable. The smoothed-periodogram metrics also perform better in the classification than the raw-periodogram based ones. when $n = 500$, the raw LNI metric and the smoothed LNI one have similar performance. There is remarkable improvement in the performance of the smoothed NI metric compared to the raw NI one. The NI metric with $M = 5\%$, and the LNI or the KD metrics correctly



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

Figure 1: *Plots of the theoretical spectral densities under Scenarios 1–4. Notes: y-axis is the logarithm of the normalized spectral densities and x-axis is the frequency with range of $(0, \pi)$; solid lines are for the first six models and dashed lines for the other six models listed in Table 1.*

classify the stationary processes at least with 90% even when $n = 100$. The PACF metric performs as well as the smoothed NI metric when $n = 100$ and the smoothed LNI (or KD) metric when $n \geq 200$.

We also compared the empirical percentages of successful classification under Scenarios 3 and 4 (Table 3), where the two true clusters consist of the AR processes with high parameter values and another AR ones with relatively low values. First of all, the smoothed-periodogram based metrics outperform the raw-periodogram based ones regardless of the type of lag window, smoothing rate, and sample size. The smoothed-periodogram based metrics perform even better than the correlation-based metrics under Scenario 3. In terms of Scenario 4, the PACF metric is slightly better than the LNI metric with Bartlett window and $M = 5\%$ only when $n = 500$. The correct classification is more likely to take place under Scenario 4 than under Scenario 5 in that the distinction of the theoretical spectral densities of the two clusters is more remarkable in case of Scenario 4 (Figure 1(d)).

In general, stationarity assumption only allows to the definition of spectral density, but we can apply our approach to integrated processes due to Pẽna and Poncela (2006). Scenario 5 considered by Caiado *et al.* (2006) focuses on the comparison of the classification metrics under the two clusters of the stationary and the nonstationary time-series models. In general, the stationarity is a quite explicit definition while the nonstationarity is not. That is, if a process does not satisfy the stationarity assumption, then it is simply regarded as nonstationary although we have no idea of the type of

Table 3: Empirical percentages of success on the classification under Scenarios 3 and 4

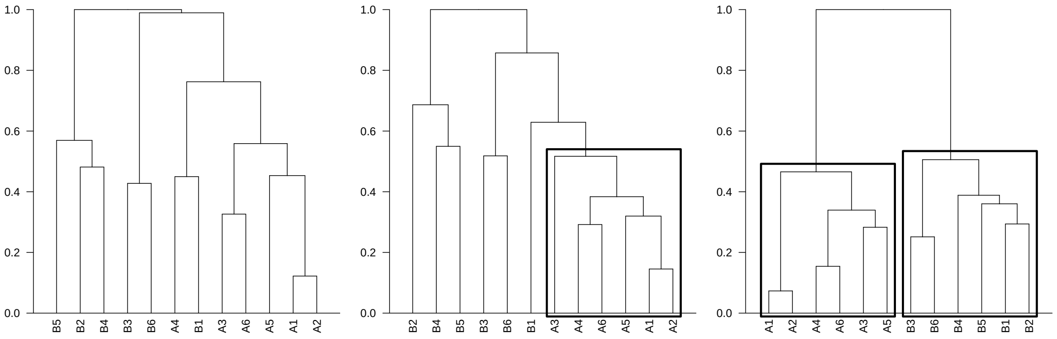
				Smoothed periodogram									Correlation	
	n	metric	Raw†	Bohman window			Bartlett window			Parzen window			ACF	PACF
				5%	10%	20%	5%	10%	20%	5%	10%	20%		
3	100	I	0.1	3.9	3.4	2.7	3.6	3.2	2.3	4.1	3.7	2.9	2.0	0.7
		NI	0.2	14.0	11.4	6.7	12.9	8.3	5.6	13.8	12.1	7.3		
		LNI	2.6	12.5	12.0	11.3	13.5	14.2	12.0	13.0	12.2	11.5		
		KD	0.6	12.5	11.6	11.5	13.7	13.8	12.5	13.0	11.8	11.6		
	200	I	0.0	9.5	8.4	6.4	8.5	8.4	6.2	9.6	8.3	7.0	2.7	2.3
		NI	0.4	29.9	21.2	14.0	25.3	18.3	13.7	30.3	22.8	15.1		
		LNI	13.1	32.9	33.4	31.3	36.8	35.0	34.3	32.9	33.2	31.5		
		KD	2.0	32.4	33.3	31.0	36.9	35.2	33.7	32.8	33.3	31.0		
	500	I	0.0	12.1	10.0	8.1	11.2	9.3	6.7	12.0	10.3	8.2	3.3	7.7
		NI	0.1	46.9	34.5	24.2	42.6	33.5	22.7	47.8	36.2	24.7		
		LNI	38.5	60.4	60.8	60.7	65.3	63.3	61.2	60.6	60.8	61.1		
		KD	2.9	60.6	60.9	60.2	65.3	63.5	61.0	60.6	61.0	60.4		
4	100	I	0.1	6.0	4.6	2.5	5.2	3.5	2.7	6.0	4.7	3.0	10.4	53.7
		NI	2.3	70.9	62.0	44.0	67.5	55.0	42.0	71.6	62.7	46.6		
		LNI	30.6	61.8	58.2	56.8	68.7	64.8	63.8	62.0	59.3	56.9		
		KD	7.3	61.2	58.4	56.6	68.6	65.0	64.8	60.9	58.4	57.2		
	200	I	0.0	2.0	1.1	1.1	1.3	0.9	0.9	2.1	1.0	1.1	7.5	79.7
		NI	0.5	79.6	54.8	34.7	69.8	49.0	33.5	82.2	58.0	35.2		
		LNI	51.0	71.3	70.3	68.1	81.7	78.8	74.3	71.8	70.1	68.3		
		KD	12.2	71.6	71.4	69.7	82.1	79.6	76.0	71.9	71.2	69.9		
	500	I	0.0	0.1	0.1	0.0	0.0	0.1	0.1	0.1	0.1	0.0	2.7	95.0
		NI	0.0	55.4	33.0	24.0	48.5	33.9	23.5	58.4	34.3	24.2		
		LNI	72.1	80.8	79.1	79.4	90.6	86.7	84.8	80.6	79.2	78.7		
		KD	20.2	82.3	81.3	81.8	91.2	87.9	86.9	82.5	81.2	81.0		

Notes: n is the sample size; Raw† denotes the raw periodogram; ACF denotes the autocorrelation;
PACF denotes the partial autocorrelation.

nonstationarity inherent in its features. We examined the performance of the metrics in terms of the following two viewpoints: how many times the two true clusters are classified correctly by each of the metrics (traditional clustering rule) and how many times the stationary or the nonstationary time-series processes are purely combined first by each of the metrics. For the rest of this work, we call the two viewpoints "From Top" direction and "From Bottom" direction, respectively.

For ease of understanding, we illustrate the two viewpoints with Figure 2, where the dendrogram plots of two groups of simulated processes are displayed. Suppose that Group A (or B) consists of six processes, A1 through A6 (or B1 through B6). Then Figure 2(a) is for the case where the two groups are not properly classified since, for instance, B1 prior to A4 was combined with the other members of Group A. As can be seen from Figure 2(b), Group A was purely clustered first before other members of Group B are merged into the group; the classification result is acceptable in terms of the "From Bottom" direction. Figure 2(c) is the ideal situation where the two groups are combined with their own members first according to the "From Bottom" direction, and, consequently, the whole processes are properly split into the two true groups according to the "From Top" direction. In this scenario, we take the "From Bottom" direction along with the "From Top" direction into account.

We now explain the results of Scenario 5 shown in Table 4 in terms of the two viewpoints. In terms of the traditional classification rule ("From Top" direction), the LNI metric performs the best among the raw-periodogram based ones. Among the smoothed-periodogram based metrics, the LNI and the KD metrics with Bohman or Parzen windows are the best. Of interest here is the contrast with the previous scenarios, any smoothed metrics with $M = 5\%$ are not always the best and the smoothed LNI metric performs worse than the raw LNI metric in some situations. Among the correlation-based



(a) (b) (c)

Figure 2: *Dendrogram plots of simulated time-series.*

Table 4: Empirical percentages of success on the classification under Scenario 5

	n	metric	Raw†	Smoothed periodogram									Correlation	
				Bohman window			Bartlett window			Parzen window			ACF	PACF
				5%	10%	20%	5%	10%	20%	5%	10%	20%		
T	100	I	0.1	0.3	0.5	0.3	0.3	0.4	0.2	0.3	0.5	0.3	4.2	0.0
		NI	2.2	0.0	0.0	2.3	0.0	0.6	3.3	0.0	0.0	1.9		
		LNI	1.7	5.1	5.5	6.1	3.4	2.8	3.2	4.8	5.6	6.0		
		KD	1.2	5.1	5.1	5.5	3.7	2.8	2.7	4.7	5.4	5.6		
	200	I	0.5	2.7	2.0	2.2	2.4	2.0	1.9	2.7	2.0	2.3	12.3	0.0
		NI	8.4	0.0	1.7	11.7	0.0	6.4	14.7	0.0	0.9	11.3		
		LNI	15.3	24.2	26.3	26.3	9.8	9.6	13.4	24.0	25.8	26.2		
		KD	6.7	23.5	25.0	25.1	12.0	9.0	11.4	23.5	24.8	24.8		
	500	I	0.4	2.6	2.0	1.5	2.5	1.8	1.3	2.7	2.1	1.7	32.7	0.0
		NI	19.8	2.4	44.7	54.9	19.6	53.0	53.8	1.0	40.9	55.3		
		LNI	66.7	72.8	72.4	72.3	40.5	56.6	70.0	72.8	72.8	72.4		
		KD	36.6	71.8	71.2	71.5	45.0	50.2	62.6	72.1	71.4	71.3		
B	100	I	16.6	13.4	13.0	13.7	13.2	13.2	13.7	13.4	13.2	13.6	10.4	53.7
		NI	3.1	11.1	9.4	9.3	10.0	9.5	8.2	11.3	9.7	9.4		
		LNI	4.1	8.0	7.7	8.2	7.8	8.5	9.1	8.1	7.8	8.3		
		KD	2.8	8.0	7.6	8.0	7.8	8.0	9.1	7.9	7.8	8.2		
	200	I	54.9	46.2	46.3	46.3	46.4	46.5	46.5	46.0	46.2	46.1	5.1	2.4
		NI	15.7	32.5	29.1	23.3	30.2	26.3	20.7	33.1	29.3	23.9		
		LNI	22.3	31.2	30.4	29.3	30.5	33.5	35.3	31.5	30.3	29.3		
		KD	12.4	31.2	30.3	29.6	29.9	33.0	34.8	31.2	30.3	29.5		
	500	I	95.4	89.2	89.4	90.2	89.3	89.2	90.5	89.1	89.4	90.1	57.8	11.9
		NI	57.8	74.8	70.1	61.8	72.6	65.4	57.9	75.2	70.8	62.6		
		LNI	69.2	73.5	72.8	72.6	75.7	78.6	77.5	73.7	73.2	72.7		
		KD	53.6	72.8	72.2	72.7	75.4	78.2	77.5	73.2	72.4	72.5		

Notes: T denotes "From Top" direction; *B* denotes "From Bottom" direction; *n* is the sample size; Raw† denotes the raw periodogram; ACF denotes the autocorrelation; PACF denotes the partial autocorrelation.

metrics, the ACF metric outperforms the PACF one, unlike the results from Scenarios 1 through 4. In terms of the second viewpoint ("From Bottom" direction), whether or not the I metric is smoothed, they yield the best performances irrespective of *n*. Since Figure 2(c) is a special case of Figure 2(b), the phenomenon illustrated in Figure 2(b) occurs quite often to the I metrics. However, the smoothed LNI and the smoothed KD metrics with Bohman or Parzen windows have insignificant differences between the percentages based on the two viewpoints. The differences are nearly 1% when $n = 500$. Overall comparison forms the conclusion that the LNI or KD metrics with Bohman or Parzen windows

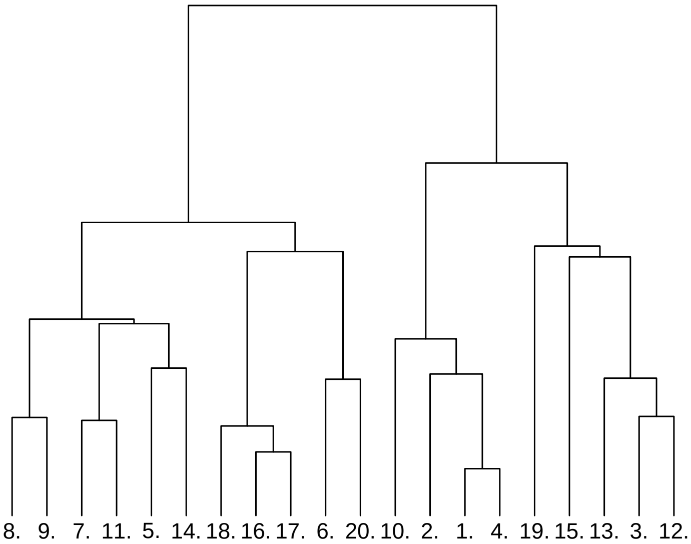


Figure 3: *Dendrogram plots of differenced log series of individual production indices in US by using the smoothed LNI metric under Bartlett window with 5% truncation points.*

tend to split the data into the true clusters much better.

In this section, we conducted various simulation scenarios in order to evaluate the classification metrics. The NI metric with Bohman or Parzen window, and the LNI or KD metrics with Bartlett window are preferred to classify stationary time-series models. However, the correct classification can be done by the LNI or KD metrics with Bohman or Parzen windows when the nonstationary models are also to be considered.

4. Real Application

In this section, we apply the classification metrics shown above to real data, which were analyzed in Caiado *et al.* (2006). We use the Industrial Production (by Market Group) indices in the United States (source: http://www.economagic.com). The data contain seasonally adjusted time-series indices, each of which are monthly measurements from January 1997 to September 2002 ($n = 309$). In order to try direct comparison with the clustering outcomes shown Caiado *et al.* (2006), we also transformed each time-series sequence, $\{x_t\}$ into its differences of the logarithm, $\{\ln x_t - \ln x_{t-1}\}$ to stabilize the variations and remove the linear trend. For the time plots of differenced log series of 12 industrial production indices, see Caiado *et al.* (2006).

Figure 3 displays the dendrogram resulting from the complete linkage classification method of the series using the smoothed LNI metric under Bartlett window with the 5% truncation point, which has the best performance from Section 3. We used Goodman and Kruskal's Gamma coefficient (Baker and Hubert, 1975), which is an index of association between the clustering levels and the dissimilarity metrics, to avoid the subjective choice of the number of clusters based on the real data. The gamma value recommended two clusters in the data as the best hierarchy level. The determined clusters are as follows: $C_1 = \{1, 2, 3, 4, 10, 12, 13, 15, 19\}$ and $C_2 = \{5, 6, 7, 8, 9, 11, 14, 16, 17, 18, 20\}$. When we consider three clusters, Cluster C_1 is divided into two groups: $C_{11} = \{1, 2, 4, 10\}$ and $C_{12} = \{3, 12, 13, 15, 19\}$. When applying the smoothed KD metric under the same window with the same

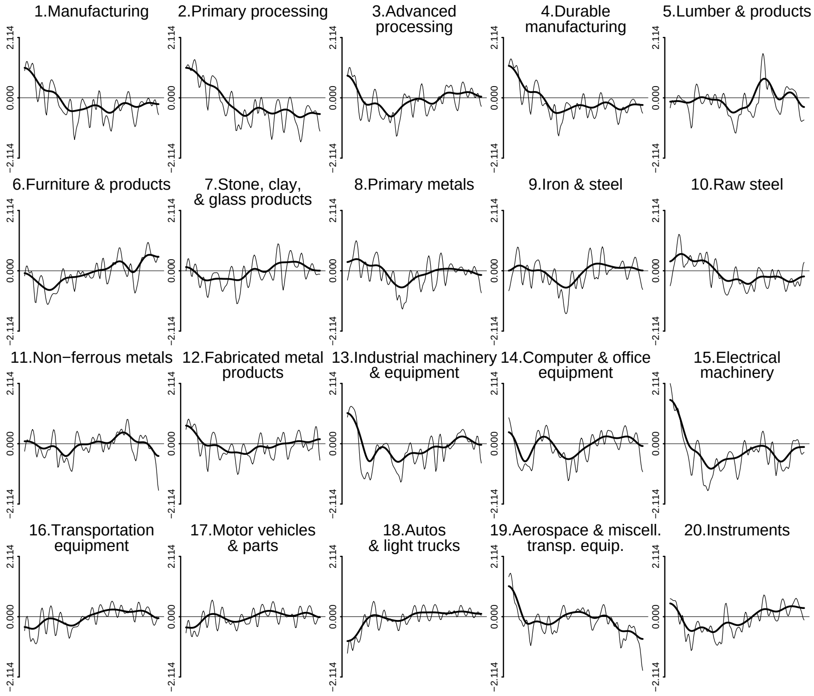


Figure 4: *Plots of logarithm of the normalized smoothed periodograms under Bartlett window. Notes: the thin lines are for* 20% *truncation points and the thick lines for* 5% *truncation points.*

level of truncation points, we obtained the same classification result.

We explain the characteristics of the clusters with the logarithm of the normalized smoothed periodograms illustrated in Figure 4. The indices in Cluster C_1 have large density for very low frequencies compared to high frequencies. Their densities rapidly decrease in the low-frequency region, $(0, \pi/3)$ except Index 10 (Raw steel). Conversely, Cluster C_2 includes the indices in which the densities for low frequencies are not so large as to dominate over all the region or the peaks in the high-frequency region, $(2\pi/3, \pi]$. The two groups, C_{11} and C_{12} consisting of Cluster C_1 are different in that the densities of the indices in C_{12} tend to increase outside the low-frequency region.

It is unreasonable to compare the clustering results based on different dissimilarity metrics, but we briefly describe our classification result against Caiado *et al.* (2006), where the raw LNI metric split the indices into the following three clusters: $G_1 = \{1, 2, 3, 4, 6, 7, 12, 13, 14, 19, 20\}$, $G_2 = \{8, 9, 10\}$ and $G_3 = \{5, 11, 16, 17, 18\}$. Clusters G_1 and C_1 are similar except the indices, $\{6, 7, 10, 14, 15, 20\}$. Index 15 (Electrical machinery) is isolated from the three clusters in Caiado *et al.* (2006). G_3 is included in C_2 and the indices $\{16, 17, 18\}$ are more homogeneous than the others in the same cluster from Figure 3. These indices turned out to be white noise when applying Box-Jenkins models, though the results are not listed in this study.

5. Conclusion

In this study, we reviewed the dissimilarity metrics based on the smoothed periodogram proposed in Park and Kim (2008) and considered several lag windows with unequal weights: Bohman, Bartlett and Parzen windows. We compared the smoothed-periodogram based metrics having the lag windows with the raw-periodogram based ones via simulation. We applied them to the monthly Industrial Production Indices in the United States. The main contribution of the proposed metrics for the classification is that we showed the superiority of the smoothed-periodogram based metrics over the raw-periodogram based ones, in terms of the empirical percentages of success on the classification. One of the reasons is that smoothing the raw periodogram with any lag windows improves the stability of estimation of the spectral density and better captures the underlying characteristics.

Performance of the smoothed-periodogram based metrics is influenced by the type of lag windows. This is an avenue for further research. We can also extend this research to compare autoregressive conditional heteroscedasticity(ARCH) and generalized ARCH(GARCH) models by using transformations of a periodogram.

Acknowledgements

The authors gratefully present their thanks to Professor Jorge Caiado who supplied the dataset used in Section 5.

References

Baker, F. B. and Hubert, L. J. (1975). Measuring the power of hierarchical cluster analysis, *Journal of the American Statistical Association,* **70**, 31–38.

Bohte, Z., Cepar, D. and Kosmelij, K. (1980). Clustering of time series, *In Proceedings of COMPSTAT,* 587–593.

Brillinger, D. (1981). *Time Series: Data Analysis and Theory,* Holden-Day, San Francisco.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods,* Springer-Verlag, New York.

Caiado, J., Crato, N. and Pẽna, D. (2006). A periodogram-based metric for time series classification, *Computational Statistics and Data Analysis,* **50**, 2668–2684.

Chatfield, C. (1975). *The Analysis of Time Series: Theory and practice,* Chapman & Hall, London.

Chen, G., Abraham, B. and Peiris, S. (1994). Lag window estimation of the degree of differencing in fractionally integrated time series models, *Journal of Time Series Analysis,* **15**, 473–487.

Corduas, M. and Piccolo, D. (2008). Time series clustering and classification by the autoregressive metric, *Computational Statistics and Data Analysis,* **52**, 1860–1872.

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.

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

Golay, X., Kollias, S., Stoll, G., Meier, D., Valvanis, A. and Boesiger, P. (1998). A new correlation-based fuzzy logic clustering algorithm for fMRI, *Magnetic Resonance in Medicine,* **40**, 249–260.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F. Å. and Hansen, L. K. (1999). On clustering fMRI time series, *Neuroimage,* **9**, 298–310.

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

Kovačić, Z. J. (1996). Classification of time series with applications to the leading indicator selection, *In Proceedings of the Fifth Conference of IFCS,* **2**, 204–207.

Kullback, S. (1978). *Information Theory and Statistics,* Peter Smith, Gloucester, Massachusetts.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *Annals of Mathematical Statistics,* **22**, 79–86.

Macchiato, M., 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.

Maharaj, E. A. (2000). Cluster of time series, *Journal of Classification,* **17**, 297–314.

Park, M. S. and Kim, H.-Y. (2008). Classification of precipitation data based on smoothed periodogram, *The Korean Journal of Applied Statistics,* **21**, 547–560.

Pẽna, D. and Poncela, P. (2006). Nonstationary dynamic factor models, *Journal of Statistical Planning and Inference,* **136**, 1237–1257.

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

Priestley, M. B. (1981). *Spectral Analysis and Time Series,* Academic Press, New York.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing,* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Shumway, R. H. (2003). Time-frequency clustering and discriminant analysis, *Statistics and Probability Letters,* **63**, 307–314.

Wismüller, A., Lange, O., Dersch, D. R., Leinsinger, G. L., Hahn, K., Pütz, B. and Auer, D. (2002). Cluster analysis of biomedical image time-series, *International Journal of Computer Vision,* **46**, 103–128.

Received February 2010; Accepted April 2010