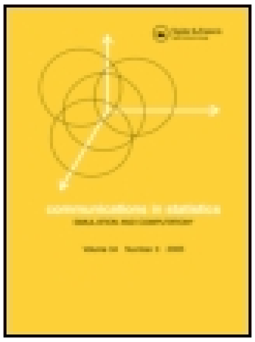

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com/loi/lssp20

Comparison of Times Series with Unequal Length in the Frequency Domain

Jorge Caiado , Nuno Crato & Daniel Peña

To cite this article: Jorge Caiado , Nuno Crato & Daniel Peña (2009) Comparison of Times Series with Unequal Length in the Frequency Domain, Communications in Statistics - Simulation and Computation, 38:3, 527-540, DOI: 10.1080/03610910802562716

To link to this article: http://dx.doi.org/10.1080/03610910802562716

Published online: 06 Feb 2009.

Communications in Statistics—Simulation and Computation®, 38: 527–540, 2009

ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610910802562716

Comparison of Times Series with Unequal Length in the Frequency Domain

JORGE CAIADO[1], NUNO CRATO[1], AND DANIEL PEÑA[2]

[1]CEMAPRE/ISEG, Technical University of Lisbon, Lisboa, Portugal
[2]Department of Statistics, Universidad Carlos III de Madrid, Getafe, Spain

In statistical data analysis it is often important to compare, classify, and cluster different time series. For these purposes various methods have been proposed in the literature, but they usually assume time series with the same sample size. In this article, we propose a spectral domain method for handling time series of unequal length. The method make the spectral estimates comparable by producing statistics at the same frequency. The procedure is compared with other methods proposed in the literature by a Monte Carlo simulation study. As an illustrative example, the proposed spectral method is applied to cluster industrial production series of some developed countries.

Keywords Autocorrelation function; Cluster analysis; Interpolated periodogram; Reduced periodogram; Spectral analysis; Time series; Zero-padding.

Mathematics Subject Classification 37M10; 62H30.

1. Introduction

The classification and clustering of time series has useful applications in several fields. In population studies, one may be interested in identifying similarities among several series of birth and death rates. In finance, one may be interested in classifying and grouping stocks for portfolio design purposes. In international economics, one may be interested in comparing and clustering countries by looking at their main macroeconomic time series indicators.

Methods for comparing time series have been studied by using autocorrelation and spectral analysis and by model fitting methods. Building upon the earlier work of Coates and Diggle (1986), Diggle and Fisher (1991), Dargahi-Noubary (1992), and others, Maharaj (2002), Quinn (2006), and Caiado et al. (2006) proposed

Received September 3, 2008; Accepted October 17, 2008
Address correspondence to Jorge Caiado, CEMAPRE/ISEG, Technical University of Lisbon, Rua do Quelhas 6, 1200-781 Lisboa, Portugal; E-mail: jcaiado@iseg.utl.pt

frequency-domain methods for time series discrimination and clustering. As the last of these articles shows, spectral methods can work very well for these purposes.

A problem that often arises in real applications is dealing with time series of unequal length. For instance, in the business cycle study of some industrialized countries, Caiado et al. (2006) deal with time series of unequal length by truncating the series to the length of the shortest one. They do it in order to use spectral estimates to compute distances across countries. In this article, we propose to deal with this problem by adjusting the number of used periodogram ordinates. We construct an interpolated periodogram for the longer series at the frequencies defined by the shorter series. This method seems to work particularly well for comparison purposes.

The remainder of the article is organized as follows. In Sec. 2, we present a well-known procedure that have been proposed in the literature for handling series of unequal length in the spectral domain; we discuss a natural extension of the usual periodogram, and propose our interpolation method. In Sec. 3, we present the results of a Monte Carlo simulation study where our method is compared to the other procedures. In Sec. 4, we apply the interpolation-periodogram based discrepancy statistic to analyze industrial time series of developed economies. In Sec. 5, we summarize the main results obtained in this article.

2. Periodogram-Based Discrepancy Statistics

Periodograms provide useful statistics for studying and comparing time series. Various authors have used pairwise comparison of periodogram ordinates from different series at the corresponding frequencies. This can be done directly when the series have the same number of data points. A problem arises when the data sets have different lengths and the Fourier frequencies at which the periodogram ordinates are usually computed are not the same.

Let $\{x_t, t = 1, \dots, n_x\}$ and $\{y_t, t = 1, \dots, n_y\}$ be two stationary processes with different sample sizes. Without loss of generality, assume that $n_x > n_y$. The periodogram of series x_t is given by

$$P_x(\omega_j) = \frac{1}{n_x}\left|\sum_{t=1}^{n_x} x_t e^{-it\omega_j}\right|^2, \tag{1}$$

where $\omega_j = 2\pi j/n_x$, for $j = 1, \dots, m_x$, with $m_x = [n_x/2]$, the largest integer less or equal to $n_x/2$. The periodogram of series y_t, $P_y(\omega_p)$, is given by a similar expression with $\omega_p = 2\pi p/n_y$, for $p = 1, \dots, m_y$, with $m_y = [n_y/2]$. When $m_x \neq m_y$, ω_j and ω_p do not coincide. Then, if we want to compare these two time series, a direct distance between the same periodograms ordinates cannot be computed.

A first solution to this problem common in the pattern recognition and signal processing literature (e.g., Wang and Blostein, 2004) consists in extending the shorter series y_t, by adding zeros and getting a new series y'_t with the same length as the longer one. One obtains

$$y'_t = \begin{cases} y_t, & t = 1, \dots, n_y \\ 0, & t = n_y + 1, \dots, n_x, \end{cases}$$

and then computes the periodogram of series y'_t, $P_{y'}(\omega_j)$. This approach, called "zero-padding", matches the frequencies of both series.

A *zero-padding* periodogram discrepancy statistic for handling series of unequal length can then be defined by

$$d_{ZP}(x, y) = \sqrt{\frac{1}{m_x}\sum_{j=1}^{m_x}[P_x(\omega_j) - P_{y'}(\omega_j)]^2}. \tag{2}$$

A second solution to the unequal length problem consists in calculating both periodograms at a common frequency. Although this is a simple and natural way of dealing with unequal sample sizes data, it has not been discussed in the time series classification literature. The procedure can be applied in various ways, but it seems natural to compute the periodogram of the longer series x_t of the shorter series y_t frequencies, that is

$$P_x^{RP}(\omega_p) = \frac{1}{n_x}\left|\sum_{t=1}^{n_x} x_t e^{-it\omega_p}\right|^2, \tag{3}$$

where $\omega_p = 2\pi p/n_y$, for $p = 1, \ldots, m_y < m_x$. We will call it the "reduced periodogram".

A *reduced periodogram* discrepancy statistic can be defined by

$$d_{RP}(x, y) = \sqrt{\frac{1}{m_y}\sum_{p=1}^{m_y}[P_x^{RP}(\omega_p) - P_y(\omega_p)]^2}. \tag{4}$$

The solution we propose is to interpolate the periodogram ordinates of the series with longer length at the frequencies defined by the series with the shorter length. Without loss of generality, let $r = [p\frac{m_x}{m_y}]$ be the largest integer less or equal to $p\frac{m_x}{m_y}$ for $p = 1, \ldots, m_y$, and $m_y < m_x$. The periodogram ordinates of x_t can be estimated as

$$\begin{aligned} P_x^{IP}(\omega_p) &= P_x(\omega_r) + (P_x(\omega_{r+1}) - P_x(\omega_r)) \times \frac{\omega_{p,y} - \omega_{r,x}}{\omega_{r+1,x} - \omega_{r,x}} \\ &= P_x(\omega_r)\left(1 - \frac{\omega_{p,y} - \omega_{r,x}}{\omega_{r+1,x} - \omega_{r,x}}\right) + P_x(\omega_{r+1})\left(\frac{\omega_{p,y} - \omega_{r,x}}{\omega_{r+1,x} - \omega_{r,x}}\right). \end{aligned} \tag{5}$$

This procedure will yield an interpolated periodogram with the same Fourier frequencies of the shorter periodogram $P_y(\omega_p)$.

The *interpolated periodogram* discrepancy statistic we propose is then given by

$$d_{IP}(x, y) = \sqrt{\frac{1}{m_y}\sum_{p=1}^{m_y}[P_x^{IP}(\omega_p) - P_y(\omega_p)]^2}. \tag{6}$$

In practical terms, if we are only interested in the dependence structure and not in the process scale, we can normalize the periodograms dividing the ordinates by the data sample variances: $NP_x^{IP}(\omega_p) = P_x^{IP}(\omega_p)/\hat{\sigma}_x^2$ and $NP_y(\omega_p) =$

$P_y(\omega_p)/\hat{\sigma}_y^2$. Additionally, it is useful for the statistical analysis and testing to attain homoscedasticity in the periodogram. Since the variance of the periodogram ordinates is proportional to the spectrum at the corresponding Fourier frequencies, we may take logarithms of the ordinates. The *interpolated log-normalized periodogram* discrepancy statistic can then be defined as

$$d_{ILNP}(x, y) = \sqrt{\frac{1}{m_y}\sum_{p=1}^{m_y}[\log NP^{IP}(\omega_p) - \log NP_y(\omega_p)]^2}. \tag{7}$$

For reference, we also consider a well-known nonparametric discrepancy statistic based on the estimated autocorrelations (Caiado et al., 2006; Galeano and Peña, 2000). Let $\hat{\rho}_{x,l}$ and $\hat{\rho}_{y,l}$ be the sample autocorrelation functions of the longer series x_t and shorter series y_t, respectively.

The *autocorrelation* discrepancy statistic is given by:

$$d_{ACF}(x, y) = \sqrt{\sum_{l=1}^{L_y}(\hat{\rho}_{x,l} - \hat{\rho}_{y,l})^2}, \tag{8}$$

where the number of autocorrelation lags used, L_y, would depend on the number of data points at hand. Here, L_y will be the largest integer less or equal to $n_y/10$, as recommended by Caiado et al. (2006).

It is straightforward to show that the statistics (2), (4), (6), (7), and (8) fulfill some properties of a *distance*: (i) $d(x, y) = 0$ if $P_x(\omega_j) = P_{y'}(\omega_j)$, $P_x^{RP}(\omega_p) = P_y(\omega_p)$, $P_x^{IP}(\omega_p) = P_y(\omega_p)$, $NP^{IP}(\omega_p) = NP_y(\omega_p)$, or $\hat{\rho}_{x,l} = \hat{\rho}_{y,l}$; (ii) $d(x, y) \geq 0$ as all the quantities are non-negative; and (iii) $d(x, y) = d(y, x)$, as all transformations are independent of the ordering. However, nothing guarantees the triangle inequality, which is the remaining defining property of a distance. For this reason we use the word "discrepancy" instead of "metric" as a convenient qualifier for the statistics under consideration.

3. Monte Carlo Simulations

To illustrate the performance of the autocorrelation and periodogram-based statistics (zero-padding, reduced, and interpolated), we performed a set of Monte Carlo simulations. For each of the considered processes, we simulated pairs of series of different sample sizes, $(n_1, n_2) = \{(50, 100), (200, 100), (500, 250), (1{,}000, 500)\}$. For each case, we performed 1,000 replications. We performed the following comparisons:

(a) AR(1), $\phi = 0.9$ versus AR(1), $\phi = 0.5$;
(b) AR(1), $\phi = 0.9$ versus ARIMA(0,1,0);
(c) AR(2), $\phi_1 = 0.6$, $\phi_2 = -0.3$ versus MA(2), $\theta_1 = -0.6$, $\theta_2 = 0.3$;
(d) ARFIMA(0, 0.45, 0) versus white noise;
(e) ARFIMA(0, 0.45, 0) versus AR(1), $\phi = 0.95$;
(f) ARFIMA(0, 0.45, 0) versus IMA(1, 1), $\theta = 0.4$;
(g) ARMA(1, 1), $\phi = 0.95$, $\theta = 0.74$ versus IMA(1, 1), $\theta = 0.8$;
(h) Deterministic trend, $x_t = 1 + 0.02t + \varepsilon_t$ versus stochastic trend, $x_t = 0.02 + x_{t-1} + (1 - 0.9B)\varepsilon_t$.

In case (a), we compare low-order models of similar type and similar autocorrelation functions. In case (b), we compare a nonstationary process with a near nonstationary AR process. In case (c), we compare selected second-order ARMA processes in order to deal with peak spectra. In case (d), we compare stationary processes with very different characteristics of persistence. In case (e), we compare a near-nonstationary long-memory process with a short-memory one. In case (f), we compare a long-memory process with a nonstationary one. In case (g), we compare a near-nonstationary process with a nonstationary one. The models chosen are those discussed in Wichern (1973). In case (h), we compare a trend-stationary process and difference-stationary one. The models chosen are based on a suggestion of Enders (1995, p. 252), but incorporate a near unit root in the MA component of the stochastic formulation in order to made them more difficult to distinguish. The rational for these choices was to generate processes with similar sample characteristics. Case (d) is an apparent exception to this rule. In this case, we were simply interested in knowing whether our methods could succeed in distinguishing long memory from no memory models.

The fractional noise was simulated using the finite Fourier method of Davies and Harte (1987). The other processes were generated with the well-tested recursive method available in Matlab. In all cases, the series were generated with a zero mean and unit variance white noise. In case (h), the series were first detrended by fitting a simple linear regression before computing the periodograms and the autocorrelation functions. As it is well known, long cyclical periods will not be eliminated by detrending.

For each case, the four generated series were grouped into two clusters by the complete linkage method. This method (also known as the farthest-neighbor method) defines the distance between two clusters by considering all possible pairs of objects (series), one from each cluster. The distance between two clusters is the maximum possible distance calculated for all these pairs of objects. It proceeds recursively. It starts with as many clusters as the number of series. At each step, it groups the existing clusters into fewer clusters by aggregating the two most similar ones. The procedure continues until it groups all objects. In our case, it stops with two clusters. For details, see, for instance, Johnson and Wichern (2007).

Table 1 provides the percentages of comparison successes in cases (a)–(h). Each comparison is defined as a success when the two time series of different length but generated by the same process are classified in the same group. The first rows of each cell show the results for the autocorrelations approach. The second rows of each cell show the results for the zero-padding periodogram approach. The third rows of each cell show the results for the reduced periodogram approach. The fourth rows of each cell show the results for the interpolated log normalized periodogram approach. For instance, the value 67.8 in the upper-left cell means that 67.8% of the times the two AR(1), $\phi = 0.9$, $n_1 = 50$, and $n_2 = 100$ processes were grouped into one cluster and the two AR(1), $\phi = 0.5$, $n_1 = 50$, and $n_2 = 100$ processes were grouped into another cluster using the autocorrelations method.

The interpolated-periodogram discrepancy statistic shows a remarkable good performance on the comparisons among stationary processes with ARMA and ARFIMA formulations, and shows a performance that increases significantly with the sample size on the comparison between ARMA and ARIMA processes and between ARIMA and ARFIMA processes.

The zero padding method works well for classifying longer series of similar length. However, it is not able to separate well near-non stationary processes

Table 1
Percentages of success on the comparison of pairs of simulated time series models: Autocorrelation function (ACF), zero-padding (ZP), reduced periodogram (RP), and interpolated log normalized periodogram (ILNP)

n_1, n_2	Discrep.	50, 100	200, 100	500, 250	1000, 500	50, 100	200, 100	500, 250	1000, 500
		(a) AR(1), 0.9 vs. AR(1), 0.5				(b) AR(1), 0.9 vs. ARIMA(0, 1, 0)			
50, 100	ACF	67.8	69.5	67.8	70.5	19.5	47.4	72.1	74.2
	ZP	63.4	74.0	78.1	80.3	20.8	45.6	88.4	97.6
	RP	50.1	52.0	54.2	51.9	12.5	31.2	74.2	92.7
	ILNP	61.2	73.4	98.4	100.0	16.4	42.4	88.0	99.7
200, 100	ACF	89.2	81.3	82.3	78.1	14.5	39.8	80.1	88.7
	ZP	87.6	91.4	92.4	93.5	25.6	52.0	88.4	97.2
	RP	71.3	76.9	81.4	82.9	11.0	31.0	67.6	89.2
	ILNP	84.8	87.9	95.4	99.9	22.8	36.0	76.6	96.4
500, 250	ACF	98.2	97.4	90.2	90.8	11.8	40.5	78.3	95.6
	ZP	97.6	99.2	99.1	99.3	31.6	61.6	88.4	96.4
	RP	83.6	93.3	97.2	98.7	11.9	25.4	68.7	88.7
	ILNP	99.1	98.6	99.2	99.9	82.4	58.2	74.8	92.0
1000, 500	ACF	98.6	99.7	98.0	94.6	12.7	42.8	78.7	89.3
	ZP	98.8	99.6	100.0	100.0	36.4	60.0	84.0	96.8
	RP	91.0	97.4	99.8	99.9	11.1	24.8	67.6	86.7
	ILNP	100.0	100.0	99.9	100.0	99.8	96.4	79.4	89.0
		(c) AR(2), 0.6, −0.3 vs. MA(2), −0.6, 0.3				(d) ARFIMA(0, 0.45, 0) vs. W. Noise			
50, 100	ACF	32.8	41.7	44.5	49.9	68.2	69.1	65.5	64.3
	ZP	32.5	44.6	62.8	71.0	39.2	41.6	49.6	57.2
	RP	28.7	31.7	42.4	41.7	28.5	31.7	32.8	34.6
	ILNP	34.9	49.5	94.7	100.0	45.5	54.6	95.3	100.0

200, 100	ACF	45.6	42.3	50.7	48.8	83.7	75.4	73.5	70.6
	ZP	40.9	47.8	71.2	83.3	54.8	60.1	70.4	79.3
	RP	30.8	39.9	48.9	54.3	34.8	41.9	51.3	49.6
	ILNP	55.2	58.8	80.7	98.7	63.8	66.7	82.8	99.4
500, 250	ACF	48.6	55.2	56.8	57.7	97.7	90.8	74.9	73.8
	ZP	53.0	67.5	77.6	92.3	82.4	89.2	90.1	94.4
	RP	33.2	47.4	63.6	73.6	43.7	57.1	77.0	83.7
	ILNP	93.4	81.3	88.4	91.3	95.5	87.0	93.7	95.9
1000, 500	ACF	50.8	51.6	65.9	69.7	99.7	97.8	87.0	76.3
	ZP	57.9	76.1	91.2	95.7	90.8	96.4	98.8	99.1
	RP	32.2	46.3	69.3	84.9	49.1	63.5	88.2	94.1
	ILNP	100.0	98.5	93.3	98.8	100.0	99.1	98.2	99.5
		(e) ARFIMA(0, 0.45, 0) vs. AR(1)				(f) ARFIMA(0, 0.45, 0) vs. IMA(1, 1)			
50, 100	ACF	51.0	76.1	88.4	89.8	43.2	75.3	94.1	96.0
	ZP	72.4	91.5	99.2	99.5	53.3	77.5	92.9	98.2
	RP	53.2	80.7	94.6	97.8	30.2	58.7	83.0	94.7
	ILNP	63.5	86.3	99.5	100.0	35.6	66.1	94.6	99.9
200, 100	ACF	45.3	59.6	83.1	89.6	37.9	58.4	92.1	97.4
	ZP	74.1	89.8	99.5	99.8	54.4	73.8	93.9	98.2
	RP	54.8	82.4	97.0	98.9	25.9	52.9	81.9	92.9
	ILNP	74.9	85.2	98.7	100.0	49.5	63.9	85.9	97.8
500, 250	ACF	39.5	49.6	56.8	71.5	31.6	54.8	78.1	94.7
	ZP	75.6	90.4	98.8	99.9	54.9	73.7	91.1	96.9
	RP	53.4	80.7	97.5	99.6	25.5	52.7	83.1	92.4
	ILNP	98.7	93.4	98.6	100.0	94.6	83.7	83.3	93.6

(continued)

Table 1
Continued

n_1, n_2	Discrep.	50, 100	200, 100	500, 250	1000, 500	50, 100	200, 100	500, 250	1000, 500
1000, 500	ACF	36.8	44.6	55.6	53.3	28.8	50.7	73.7	82.4
	ZP	71.3	90.7	99.1	100.0	55.3	68.6	87.3	96.7
	RP	51.6	79.9	97.2	99.8	25.9	52.7	79.9	91.9
	ILNP	100.0	100.0	99.4	97.8	100.0	99.5	92.0	93.7
		(g) ARMA(1, 1) vs. IMA(1, 1)				(h) Determ. trend vs. stoc. trend			
50, 100	ACF	9.2	19.4	56.0	77.3	9.9	21.7	48.6	69.0
	ZP	5.9	13.1	46.5	71.8	9.1	18.6	58.5	81.7
	RP	12.0	14.4	28.8	50.9	10.9	12.8	15.7	16.9
	ILNP	14.5	28.9	82.4	99.8	17.4	29.7	86.9	100.0
200,100	ACF	9.5	13.3	47.4	76.9	10.2	19.1	48.6	63.1
	ZP	10.0	7.0	25.2	56.2	7.4	9.2	38.9	70.6
	RP	11.3	10.2	31.9	48.2	12.3	11.3	18.1	21.3
	ILNP	26.7	22.2	48.1	88.8	26.9	23.3	46.5	93.4
500, 250	ACF	18.1	13.4	45.2	75.1	8.3	15.9	41.8	55.1
	ZP	37.1	12.6	11.6	36.4	16.3	11.6	20.7	52.2
	RP	12.4	13.8	26.5	49.0	13.9	16.2	20.1	28.3
	ILNP	86.9	42.6	42.0	63.2	86.3	48.2	39.7	63.8
1000, 500	ACF	25.3	14.3	43.8	70.7	8.2	14.8	35.7	46.3
	ZP	53.1	21.5	9,5	30.3	32.7	18.6	23.1	41.2
	RP	13.8	13.3	25.5	51.1	11.2	13.6	21.1	36.2
	ILNP	100.0	93.1	54.6	57.4	100.0	94.8	56.3	66.5

with large samples from non stationary processes with short samples, and, more importantly, it does not perform well on the comparison between longer stationary and shorter near-nonstationary ARMA processes. In fact, when sample sizes are very unbalanced, the shorter series periodogram is distorted by the zero-padding method. Zero padding is equivalent to add new ordinate values that are linear combinations of the periodogram ordinates of the original series. Naturally, the resulting statistics and tests suffer from this problem.

The reduced periodogram and the ACF methods are always dominated by the other methods. In particular, the reduced periodogram method displays a very poor performance for distinguishing similar processes with small samples and the ACF method is not able to distinguish near-non stationary processes with large samples from non stationary processes with short samples.

In order to better assess the methods, we have performed additional simulations for other and more dissimilar models. Results were much alike the ones here presented and pointed to the same hierarchy of discrepancy statistics. We have also explored other hierarchical and non hierarchical clustering procedures. Results were again similar and provided the same recommendations for the discrepancy statistics choice.

4. Application

As an illustration of the possibilities of these techniques, we compared the industrial time series of a set of developed countries. We used monthly data of seasonally adjusted industrial production indices for a large set of European and other industrialized economies. Available data are summarized on Table 2 (source data: Camacho et al., 2006). For such large data set, it is unavoidable that sample periods do not coincide. In order to use all available data, it is necessary to apply techniques such as the ones we have described.

In our application, we started by computing the interpolated log normalized periodograms for each of the $k = 30$ production series. The corresponding graphs

Table 2
Industrial production indices series (countries and data avaibility)

Country	Code	Sample	n	Country	Code	Sample	n
Austria	OE	62:01-02:12	492	Canada	CN	62:01-03:01	493
Belgium	BG	62:01-03:01	493	Norway	NW	62:01-03:01	493
Germany	BD	62:01-03:01	493	Japan	JP	62:01-03:01	493
Greece	GR	62:01-03:01	493	USA	US	62:01-03:01	493
Finland	FN	62:01-03:01	493	Cyprus	CY	90:01-03:01	142
France	FR	62:01-03:01	493	Czech Republic	CZ	90:01-03:01	142
Italy	IT	62:01-03:01	493	Estonia	ET	95:01-03:01	97
Ireland	IR	75:07-03:01	331	Hungary	HN	90:01-03:01	142
Luxembourg	LX	62:01-03:01	493	Latvia	LA	90:01-03:01	142
Netherlands	NL	62:01-03:01	493	Lithuania	LI	96:01-03:01	85
Portugal	PT	62:01-03:01	493	Poland	PO	90:01-03:01	142
Spain	ES	65:01-03:01	457	Slovak Republic	SK	93:01-03:01	121
Denmark	DK	74:01-03:01	349	Slovenia	SL	90:01-03:01	142
Sweden	SD	62:01-03:01	493	Romania	RO	90:01-03:01	142
United Kingdom	UK	62:01-03:01	493	Turkey	TK	90:01-03:01	142

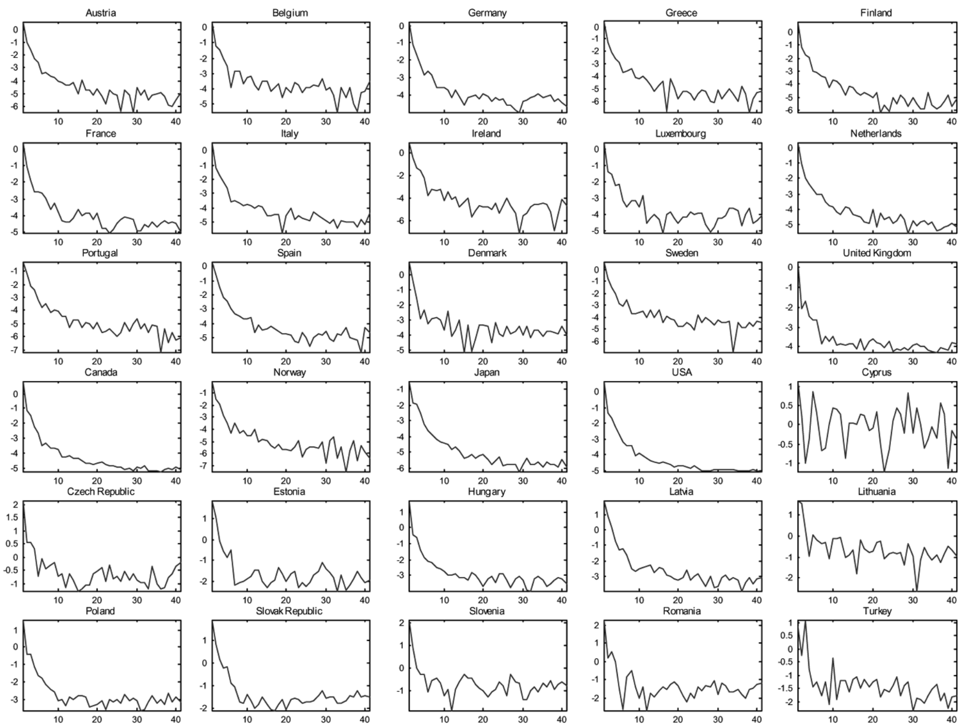

Figure 1. Log normalized interpolated periodograms of 30 European and some developed countries.

are shown in Fig. 1. We then computed all the corresponding $k(k-1)/2$ pairwise distances, by using the ILNP discrepancy statistic given in (7). In order to be able to interpret resulting data, we used two well-known techniques: the multidimensional scaling and the hierarchical clustering tree, or dendrogram (see Secs. 12.6 and 12.3, respectively, of Johnson and Wichern, 2007, for example).

Firstly, we used the multidimensional scaling technique, also often referred to as principal coordinates analysis, which creates a configuration of k points in a lower-dimensional map (usually two or three). Let D be the observed $k \times k$ dissimilarity matrix, applying the multidimensional scaling to the matrix D gives a $k \times s$ configuration matrix T, where the rows of T are the coordinates values of k points in s-dimensional representation of the observed dissimilarities for some $s < k$. The determination of the dimensionality of the spatial configuration is given by the v eigenvectors of $T \times T'$ corresponding to the largest v eigenvalues.

Table 3 shows the eigenvalues resulting from distances between countries and the eigenvectors associated with the first two eigenvalues. The first eigenvalue is equal to 93.76% of the sum of all the eigenvalues. The sum of the first two eigenvalues is equal to 94.66% of the sum of all the eigenvalues. Figure 2 shows a scaling map of the derived first two coordinate values. The first dimension seems to be almost directly related to the countries' development. The second dimension is not easy to interpret. However, looking at the 2-dimensional plot and comparing the relative positions with the periodograms plots, we can make sense of some of the results. Looking at the opposite positions of Cyprus and Ireland, for instance,

Table 3
Eigenvalues and first two eigenvectors for interpolated LNP distances between 30 European and some developed countries

Eigenvalues		Country	Eigenvectors 1	Eigenvectors 2	Country	Eigenvectors 1	Eigenvectors 2
2925.1	4.6	Austria	−8.45	−0.40	Canada	−7.20	0.02
28.3	3.2	Belgium	−2.90	0.78	Norway	−11.39	1.53
21.0	2.3	Germany	−4.09	0.21	Japan	−11.43	0.44
19.5	2.1	Greece	−10.77	−1.86	USA	−7.19	0.20
18.2	1.8	Finland	−9.18	−0.12	Cyprus	21.10	1.70
13.6	1.7	France	−4.87	0.91	Czech Republic	17.22	−0.55
12.7	1.4	Italy	−6.70	0.11	Estonia	10.50	0.31
11.0	1.2	Ireland	−7.80	−3.25	Hungary	2.34	−0.05
9.8	1.0	Luxembourg	−3.48	0.21	Latvia	3.63	−1.40
9.2	0.9	Netherlands	−6.58	0.15	Lithuania	16.91	−0.45
7.3	0.7	Portugal	−10.51	1.41	Poland	3.28	−0.18
7.0	0.3	Spain	−6.62	−0.11	Slovak Republic	11.86	−0.49
6.1	0.2	Denmark	−1.18	0.23	Slovenia	16.77	−0.43
5.0	0.1	Sweden	−5.13	−0.17	Romania	12.36	−0.33
4.6	0.0	United Kingdom	−2.69	1.36	Turkey	12.19	0.19

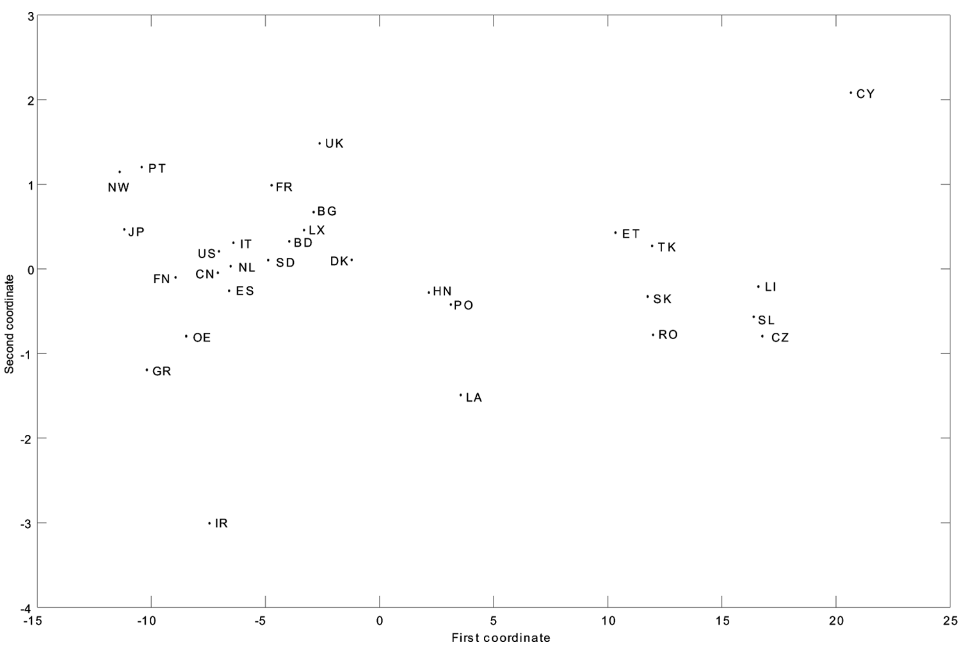

Figure 2. Multidimensional scaling for interpolated LNP distances between 30 European and some developed countries.

we realize that this distance comes from very different spectral peaks at different frequencies—the interpolated LNP of Ireland series reaches the minimum value at frequencies $\omega_{29} = 2\pi(29)/85 = 2.14367$ and $\omega_{38} = 2\pi(38)/85 = 2.80895$, whereas the interpolated LNP of Cyprus series is dominated by large peaks at the same frequencies. It can also be seen that the old European Union (EU) countries (except Ireland) and the U.S., Canada, Japan, and Norway are close to each other and far from the new EU countries and from the then candidate countries (Estonia, Turkey, Slovak Republic, Romania, Lithuania, Slovenia, Czech Republic, and Latvia). More developed Poland and Hungary are in an intermediate position.

Secondly, we consider the method of clustering the series by a hierarchical clustering tree, or dendrogram. This graphical tool shows how the clusters are combined at each stage of the procedure. We begin with each time series being considered as a separate cluster (k clusters). In the second stage, the closest two groups are linked to form $k - 1$ clusters. This process continues until the last stage in which all the time series are in the same cluster.

Figure 3 shows the dendrogram for the industrial production indices series by complete linkage method, from which the clusters of countries can be identified. It can be seen at the tree that the interpolated log normalized periodogram-based method can group the series into three very reasonable clusters: Cluster $1 = \{CN, US, NL, IT, ES, FR, SD, BG, BD, LX, UK, DK, OE, FN, GR, IR, PT, JP, NW\}$; Cluster $2 = \{CY, CZ, SL, LI\}$; and Cluster $3 = \{ET, SK, RO, TK, HN, PO, LA\}$. Cluster 1 includes all the old EU countries and the U.S., Canada, Japan, and Norway. Cluster 2 grouped four new EU countries (Cyprus, Czech Republic, Slovenia, and Lithuania). Cluster 3 includes the other new EU countries (Estonia,

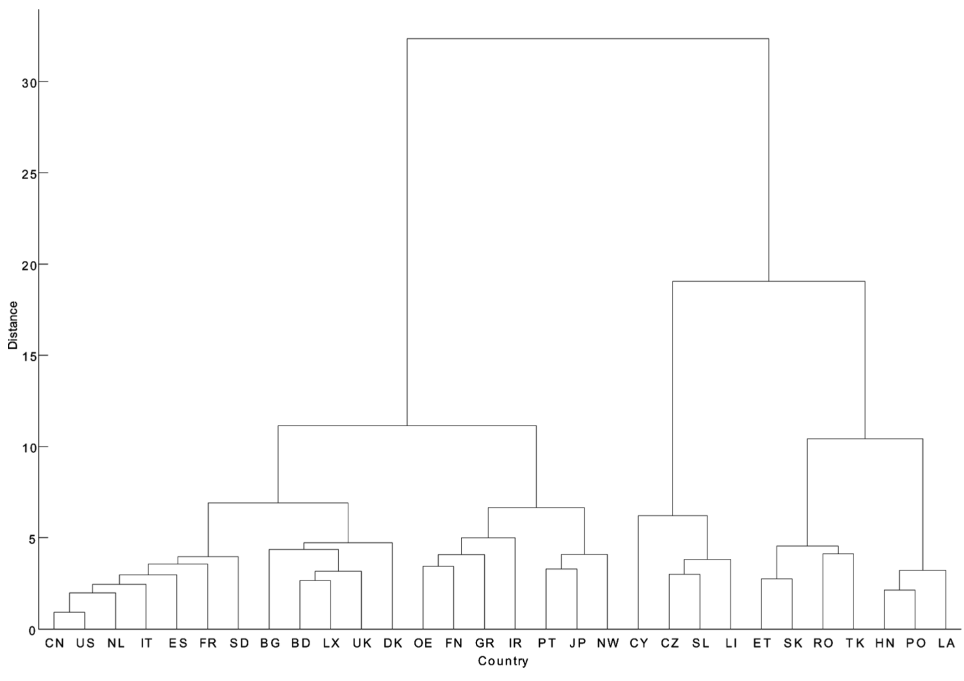

Figure 3. Complete linkage dendrogram for interpolated LNP distances between 30 European and some developed countries.

Slovak Republic, Hungary, Poland, and Latvia) and the then candidate countries (Romania and Turkey).

These results seem to be very reasonable. Not surprisingly, they group together the more developed countries. They essentially confirm the conclusions of Camacho et al. (2006), adding some interesting information.

These authors also used hierarchical clustering and multidimensional scaling techniques to identify cyclical linkages among countries. From cluster analysis, they found four clusters. The first includes most of the old EU countries, the new EU countries Poland, Slovenia, and Hungary, and the industrialized country Japan; the second includes the industrialized countries U.S., Canada, United Kingdom, and Finland; the third includes the other new EU countries (Latvia, Estonia, Czech Republic, Lithuania, and Slovak Republic), the candidate countries (Romania and Turkey), and the industrialized country Norway; and the fourth includes the old European Union countries Portugal, Greece, and Cyprus. From the multidimensional scaling map, they found that most old EU member countries are close to each other and far from the new EU member countries (except Cyprus and Slovenia); and that the very industrialized countries U.S., Canada, and United Kingdom, and new EU member countries Hungary and Finland are close to each other in a distinct location.

By using older information, our analysis is able to distinguish better the old from the new EU countries. It is also able to show U.S., Canada, and Japan close to the EU industrialized countries. Camacho et al. (2006) have only used data from 1992 onwards, while we could use data from 1962 onwards. We thus confirmed some of previous results, but our method allows complementing them with more extensive data.

5. Concluding Remarks

In this article, we presented and discussed two spectral-discrepancy statistics for comparison, classification, and clustering analysis of time series with unequal length. We proposed a novel third statistic based on the interpolated periodogram for the same purposes. We then evaluated this latter statistic against the others. For reference, we also used an autocorrelation-based discrepancy statistic.

A simulation study indicated that the proposed method, the interpolated log normalized periodogram approach, performs very well for a wide type of comparisons: (i) different stationary processes with similar sample properties; (ii) non stationary versus near nonstationary processes; and (iii) short-memory versus long-memory processes; and (iv) deterministic trend versus stochastic trend processes. Moreover, in the comparison of time series of very different sample sizes, the proposed method is preferred to the autocorrelation, the zero-padding periodogram, and the reduced periodogram-based methods. One application to industrial production series also demonstrates the merits of the method.

Acknowledgments

This research was supported by a grant from the Fundação para a Ciência e a Tecnologia (FEDER/POCI 2010) and by MEC project SEJ2007-64500, Spain. The authors gratefully acknowledge the helpful comments and suggestions of an anonymous referee on an earlier version of this article.

References

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

Camacho, M., Pérez-Quiróz, G., Saiz, L. (2006). Are European business cycles close enough to be just one? *Journal of Economic Dynamics and Control* 30:1687–1706.

Coates, D. S., Diggle, P. J. (1986). Tests for comparing two estimated spectral densities. *Journal of Time Series Analysis* 7:7–20.

Dargahi-Noubary, G. R. (1992). Discrimination between Gaussian time series based on their spectral differences. *Communications in Statistics: Theory and Methods* 21:2439–2458.

Davies, R. B., Harte, D. S. (1987). Tests for Hurst effect. *Biometrika* 74:95–102.

Diggle, P. J., Fisher, N. I. (1991). Nonparametric comparison of cumulative periodograms. *Applied Statistics* 40:423–434.

Enders, W. (1995). *Applied Econometric Time Series*. New York: Wiley.

Galeano, P., Peña, D. (2000). Multivariate analysis in vector time series. *Resenhas* 4:383–404.

Johnson, R., Wichern, D. (2007). *Applied Multivariate Statistical Analysis*. 6th ed. Englewood Cliffs, NJ: Prentice-Hall.

Maharaj, E. A. (2002). Comparison of non-stationary time series in the frequency domain. *Computational Statistics & Data Analysis* 40:131–141.

Quinn, B. G. (2006). Statistical methods of spectrum change detection. *Digital Signal Processing* 16:588–596.

Wang, N., Blostein, S. (2004). Adaptive zero-padding OFDM over frequency-selective multipath channels. *Journal on Applied Signal Processing* 10:1478–1488.

Wichern, D. W. (1973). The behaviour of the sample autocorrelation function for an integrated moving average process. *Biometrika* 60:235–239.