Tendências em Matemática Aplicada e Computacional, **19**, N. 3 (2018), 391-403

www.scielo.br/tema
doi: 10.5540/tema.2018.019.03.0391

Wavelet Cross-correlation in Bivariate Time-Series Analysis

E.M. SOUZA[1] and V.B. FÉLIX[2]

Received on February 28, 2018 / Accepted on March 26, 2018

ABSTRACT. The estimation of the correlation between independent data sets using classical estimators, such as the Pearson coefficient, is well established in the literature. However, such estimators are inadequate for analyzing the correlation among dependent data. There are several types of dependence, the most common being the serial (temporal) and spatial dependence, which are inherent to the data sets from several fields. Using a bivariate time-series analysis, the relation between two series can be assessed. Further, as one time series may be related to an other with a time offset (either to the past or to the future), it is essential to also consider lagged correlations. The cross-correlation function (CCF), which assumes that each series has a normal distribution and is not autocorrelated, is used frequently. However, even when a time series is normally distributed, the autocorrelation is still inherent to one or both time series, compromising the estimates obtained using the CCF and their interpretations. To address this issue, analysis using the wavelet cross-correlation (WCC) has been proposed. WCC is based on the non-decimated wavelet transform (NDWT), which is translation invariant and decomposes dependent data into multiple scales, each representing the behavior of a different frequency band. To demonstrate the applicability of this method, we analyze simulated and real time series from different stochastic processes. The results demonstrated that analyses based on the CCF can be misleading; however, WCC can be used to correctly identify the correlation on each scale. Furthermore, the confidence interval (CI) for the results of the WCC analysis was estimated to determine the statistical significance.

Keywords: Multiscale Analysis, Time Series, Cross Correlation, Non-Decimated Wavelet Transform.

1 INTRODUCTION

Investigating the correlation between time series is of great interest in several areas. Further, lagged relations are common in many natural systems. For example, a series may have a delayed response relative to another series or one may have a delayed response to a common stimulus that affects both series. A simple zero-lag correlation coefficient is inadequate for such situations.

*Corresponding author: Eniuce Menezes de Souza – E-mail: emsouza@uem.br
[1]Departamento de Estatística, Universidade Estadual de Maringá (UEM), Av. Colombo, 5790, 87020-900, Maringá, PR, Brasil. E-mail: emsouza@uem.br
[2]Programa de Pós-graduação em Biostatística, Universidade Estadual de Maringá (UEM), Av. Colombo, 5790, 87020-900, Maringá, PR, Brasil. E-mail: felix_prot@hotmail.com

A common way in which the relation between two time series can be analyzed is to use the cross-correlation function (CCF), to determine a simple correlation coefficient as a function of the lag or the time offset between the time series. Although the time lag at which the series are correlated can be identified using the CCF, there are a few limitations. First, each series must be normally distributed. Second, if at least one series is autocorrelated, the estimated CCF may be distorted and, therefore, may be misleading as a measure of the lag between the time series. Finally, the relations between different scales cannot be taken into account. As time series are frequently autocorrelated and composed of mixtures various effects of different frequencies, the CCF is usually inadequate. Thus, in this paper, we point out the advantages of using the WCC estimator, which can estimate how strongly two time series are correlated in terms of the lag and the scale. Some successful applications of WCC include fluid engineering [6], including the analysis of the cross-correlation between two velocity signals to investigate the structural similarity of motion on various scales in terms of the time and period delays. In another study [11], WCC was used to analyze the relation between cerebral oxyhemoglobin (O2Hb) and mean arterial blood pressure to identify autonomic failure. The results demonstrated that the frequency of the maximum wavelet cross-correlation is significantly different between patients with autonomic failure and age-matched control subjects.

In these applications, a continuous wavelet transform, which is very redundant, was used. On the contrary, decimated discrete transforms are non redundant, so they are attractive for practical applications; however, they are not the best option for time series because of the translation variance. In this article, WCC estimation will be conducted based on the non-decimated wavelet transform (NDWT), which is also called the maximal overlap discrete wavelet transform. The NDWT is a discrete transform but has the property of shift invariance and offers many advantages. One benefit is that the estimators obtained from the NDWT are asymptotically more efficient than those from the commonly used decimated wavelet transform [8].

To investigate the performance of WCC, simulated time series from various stochastic processes and scenarios as well as real data from bronchiolitis hospitalizations in two health divisions of the Paraná state in Brazil were considered.

This paper is organized as follows: the CCF and WCC are briefly described in sections 2 and 3, respectively. In section 4, several data sets were simulated and in section 5, real data is presented. Final considerations are considered in section 6.

2 CROSS-CORRELATION COEFFICIENT

Considering X_t and Y_t, $t = 1, \ldots n$, stochastic processes, and $B^j X_t = X_{t-j}$ the backward operator, the d_Xth $((1-B)^d X_t)$ and d_Yth $((1-B)^d Y_t)$ order backward differences are stationary Gaussian processes, and the cross-covariance function (CCVF) of n pairs of observations is

$$c_{XY}(k) = \begin{cases} \frac{1}{n}\sum_{t=1}^{n-k}(x_t - \bar{x})(y_{t+k} - \bar{y}), & k = 0, 1, \ldots, n-1, \\ \frac{1}{n}\sum_{t=1-k}^{n}(x_t - \bar{x})(y_{t+k} - \bar{y}), & k = -1, -2, \ldots, -(n-1), \end{cases} \tag{2.1}$$

where n is the series length, $\bar{x}$ and $\bar{y}$ are the sample means, and k is the lag, as seen in [4].
The sample cross-correlation function (CCF) is the CCVF scaled by the variances of the two series:

$$CCF_{XY}(k) = \frac{c_{XY}(k)}{\sqrt{c_{XX}(0)c_{YY}(0)}}, \tag{2.2}$$

where $c_{XX}(0)$ and $c_{XY}(0)$ are the sample variances of $\{X_t\}$ and $\{Y_t\}$. The CCF calculates the linear correlation between the series, ranging from -1 to 1.

3 WAVELET CROSS-CORRELATION

The discrete wavelet transform (DWT) consists on the decomposition of signals according to wavelet functions with discretized translation and dilatation parameters [14]. Unfortunately, the DWT is variant by translation and inadequate for time series analysis [3][7]. Hence, the NDWT [9] was chosen, because its benefits, such as translation invariance and easy interpretation of the coefficients. The NDWT coefficients can be obtained by

$$\bar{W}_{j,t} = \sum_{l=0}^{L_j - 1} h_{j,l} X_{t-l}, \tag{3.1}$$

where $h_{j,l}$, $j = 1, \ldots, J$, $l = 0, \ldots, L_j - 1$, $L_j \equiv (2^j - 1)(L-1) + 1$, are the wavelet filters [15]. So, the WCC of $\{X_t, Y_t\}$ for the scale $\lambda_j = 2^{j-1}, j = 1, \ldots, J$, and an arbitrary positive lag τ can be expressed as seen in [15]

$$\rho_{\tau,XY}(\lambda_j) \equiv \frac{Cov\left\{\bar{W}_{j,t}^{(X)}, \bar{W}_{j,t+\tau}^{(Y)}\right\}}{\left(Var\{\bar{W}_{j,t}^{(X)}\} Var\{\bar{W}_{j,t+\tau}^{(Y)}\}\right)^{1/2}} = \frac{c_{\tau,XY}(\lambda_j)}{c_{XX}(\lambda_j) c_{YY}(\lambda_j)}, \tag{3.2}$$

where

- $\left\{\bar{W}_{j,t}^{(X)}\right\}$ are the NDWT coefficients for $\{X_t\}$ in the scale λ_j;
- $\left\{\bar{W}_{j,t}^{(Y)}\right\}$ are the NDWT coefficients for $\{Y_t\}$ in the scale λ_j;
- and $-1 \leq \rho_\tau \leq 1$, for all τ and j.

For demonstration purposes, if $\{\bar{W}_{j,t}^{(X)}, \bar{W}_{j,t}^{(Y)}\}$ is a bivariate Gaussian process, the NDWT based estimator $\tilde{\rho}_{XY}(\lambda_j)$ of the wavelet correlation for scale λ_j is asymptotically normally distributed with mean $\rho_{XY}(\lambda_j)$. As seen in [14], the large sample variance of $\tilde{\rho}_{XY}$ is given by

$$\begin{aligned} Var(\tilde{\rho}_{XY}) \approx \frac{1}{\tilde{N}_j} \sum_{\tau=-(\tilde{N}_j-1)}^{\tilde{N}_j-1} \Big\{ & \rho_{\tau,X}(\lambda_j)\rho_{\tau,Y}(\lambda_j) + \rho_{\tau,XY}(\lambda_j)\rho_{\tau,YX}(\lambda_j) \\ & -2\rho_{0,XY}(\lambda_j)\left[\rho_{\tau,X}(\lambda_j)\rho_{\tau,YX}(\lambda_j) + \rho_{\tau,Y}(\lambda_j)\rho_{\tau,YX}(\lambda_j)\right] \\ & +\rho_{0,XY}^2(\lambda_j)\left[\frac{1}{2}\rho_{\tau,X}^2(\lambda_j) + \rho_{\tau,XY}^2(\lambda_j) + \frac{1}{2}\rho_{\tau,Y}^2(\lambda_j)\right]\Big\}. \end{aligned} \tag{3.3}$$

where $\tilde{N}_j$ is the number of coefficients associated with scale λ_j, and $\rho_{\tau,X}(\lambda_j)$ is the scale λ_j wavelet autocorrelation for the process $\{X_t\}$.

Using the large sample theory, an approximate CI for the NDWT estimator of the WCC can be constructed. In case of non-normal correlation coefficients in small sample sizes, a nonlinear Fisher's z transformation $h(\rho)$ is sometimes required to produce a sample correlation coefficient with an approximately Gaussian distribution and a shape that is independent of the true correlation coefficient.

The Fisher's transformation is defined by the following expression:

$$\frac{1}{2}\ln\left(\frac{1+\tilde{\rho}_X(\lambda_j)}{1-\tilde{\rho}_X(\lambda_j)}\right) = \operatorname{arctanh}(\tilde{\rho}_X(\lambda_j)). \tag{3.4}$$

Then, an approximate $100\gamma\%$ CI for $\rho_{XY}(\lambda_j)$, based on the NDWT, where γ is the nominal confidence coefficient, is

$$\left[\tanh\left\{h[\tilde{\rho}_{XY}(\lambda_j)] - \frac{\Phi_\gamma^{-1}}{\sqrt{\tilde{N}_j - 3}}\right\}, \tanh\left\{h[\tilde{\rho}_{XY}(\lambda_j)] + \frac{\Phi_\gamma^{-1}}{\sqrt{\tilde{N}_j - 3}}\right\}\right], \tag{3.5}$$

where Φ_γ^{-1} is the $100\gamma\%$ quantile of the standard normal distribution, $\tilde{N}_j$ is the number of coefficients associated with scale λ_j and $\tilde{\rho}$ is the NDWT estimator of the wavelet correlation.

In practice, the NDWT can be easily computed by applying the known pyramidal twice [1]. This algorithm is frequently used for the decimated case, but it requires a time series of dyadic length $n = 2^J$, where J represents the largest scale. Each scale j corresponds to a frequency band from 2^j to $2^{(j+1)}$, whose inversion produces the period of time evaluated in the scale of the WCC.

4 SIMULATED DATA APPLICATION

To investigate the characteristics, advantages, and potential applications of WCC, time series from different first order autoregressive (AR) and moving average (MA) stochastic processes were simulated. The AR process with autoregressive parameter α is given by

$$X_t = \alpha X_{t-1} + \varepsilon_t, \tag{4.1}$$

and the moving average process with parameter θ is

$$X_t = \varepsilon_t + \theta\varepsilon_{t-1}, \tag{4.2}$$

where ε_t is the error component.

The simulation scenarios are described in Table 1. The options from the simulation scenarios (Table 1) were combined to simulate 128 monthly time series.

Series A, B, and C were built with different frequency characteristics to identify patterns in different scales. Series D was derived from series C with a lag of 12 observations to investigate if the correlation can be detected in the presence of a lag. By considering the possible AR and MA parameters, and the various lengths, each simulated series was decomposed into 11 levels by applying the NDWT. Some of the levels were removed before reconstruction to create series with known similarities in some scales and no correlation with other scales. Based on these simulations, the performance of the correlation approach can be evaluated under a range of conditions.

Table 1: Simulation Scenarios.

Parameter	Options
α	0.2; 0.4; 0.6; 0.8
θ	0.2; 0.4; 0.6; 0.8
Length	128; 10,000
Frequency	A: Smoothest behavior; reconstruction without the levels 1, 2, and 3. B: Smooth and noisy behaviors; reconstruction without the levels 5, 6, and 7, i.e., without middle-frequencies. C: Noisy behavior; reconstruction with the levels 1 and 2, i.e, the highest frequencies. D: Noisy behavior and lagged of 12 observations; derived from series C with a lag of 12 observations.

The NDWT decomposition, reconstruction, and WCC analysis were performed in R language [10] using the Daubechies mother wavelet and all graphics were implemented by using a package called ggplot2 [16].

The results from the application of the proposed methodology to the simulated data are illustrated by the time series generated from an autoregressive model with a parameter α of 0.8 and a length of 10,000. The results obtained using other simulation parameters and lengths were similar but are not shown here due to space constraints.

In Figure 1 we have the decomposition of the simulated series, of a process AR(1) with parameter α=0.8 with length 10,000, by the NDWT in 11 levels.

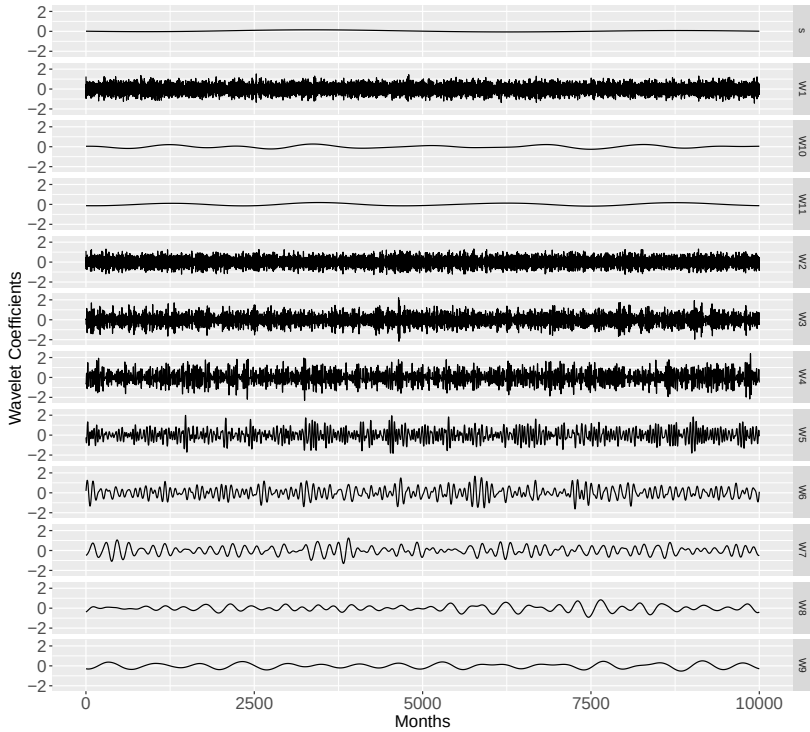

Figure 1: Wavelet coefficients from different levels of decomposition from a simulated AR(1) with parameter α=0.8 and length 10,000.

Figure 1 shows that is hard to comprehend the series behavior, particularly in the first levels where the high-frequency components are found, due to the length of the simulated series. Thus, Figure 2 shows the first 200 observations in the series for better visualization.

The CCF was applied to a set of two simulated series: AR(1) with $\alpha = 0.8$ and one of series A, B, or C, with no lag between them. The correlation coefficients and corresponding 95% CIs of the correlations with filtered series A, B, and C were, 0.871 (0.833; 0.900), 0.797 (0.740; 0.842), and 0.367 (0.240; 0.481), respectively. These results show that the correlations between AR(1) and series B or C are similar despite the different behaviors of these series. This was expected because the CCF is considered to be a general measure of correlation. A weaker correlation between AR(1) and series C was found despite the perfect correlation between the highest frequencies in these series.

Figure 3 shows that there is no correlation in the removed scales of each case. This was expected because the filtered series are reconstructions of the original series without those levels. In addition, as expected, the highest correlations occurred when there was no lag. Figure 4 shows the WCC estimates and their respective 95% CIs for the first 10 scales with no lag.

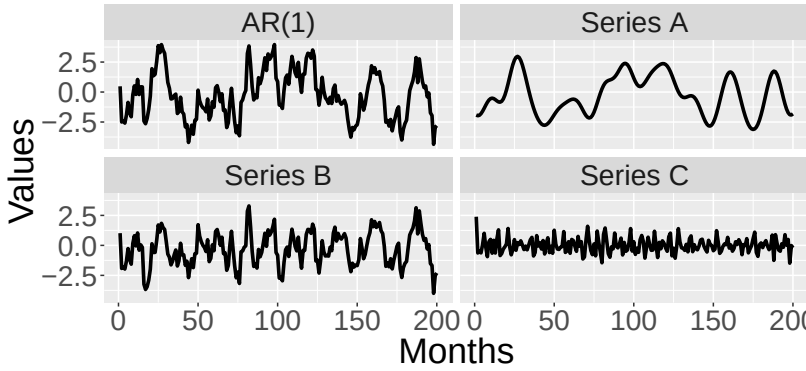

Figure 2: The first 200 observations of the AR(1) with parameter α=0.8, A, B, and C series of length 10,000.

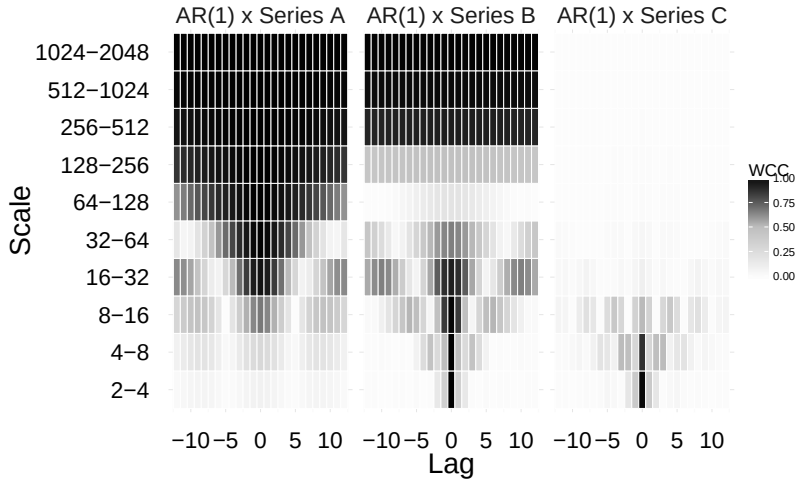

Figure 3: Estimated WCC with the frequency bands related to the first 10 scales, lags ranging from -10 to 10, for AR(1) with $\alpha = 0.8$ and the A, B, and C filtered series.

The results in Figure 4 show that the correlation between the simulated series can be estimated in the appropriate scales, and no correlation is identified in the removed scales of each data set.

In order to evaluate the correlation detection performance in the presence of various lags, Figure 5 illustrates the AR(1) simulated process along with the first 188 observations from series D with a lag of 12 observations.

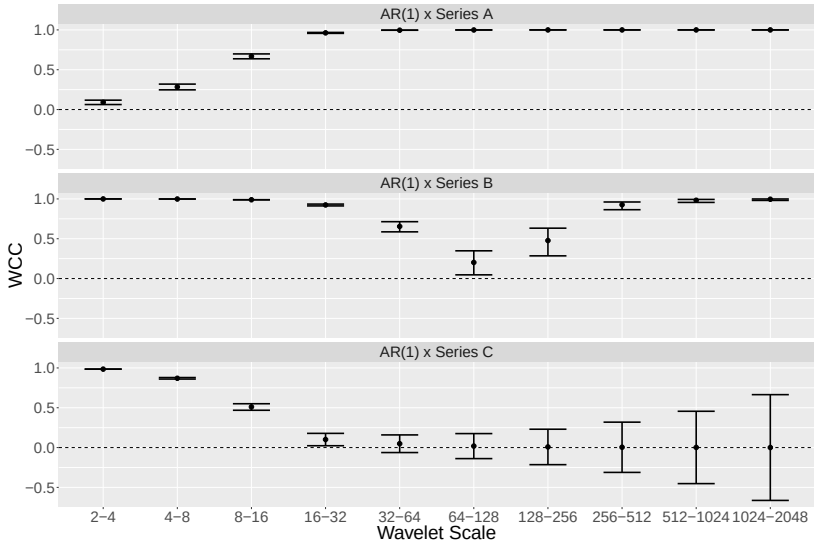

Figure 4: Estimated WCC for each scale on lag 0, considering the AR(1) with $\alpha = 0.8$ simulated series and the A, B, and C filtered series.

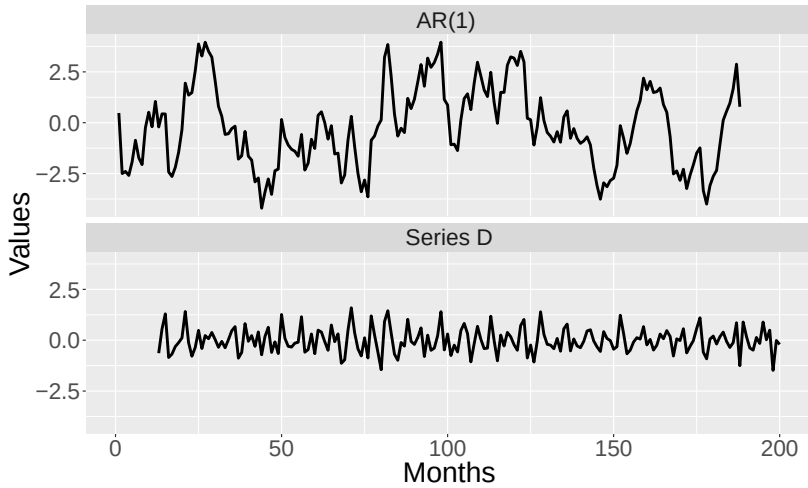

Figure 5: First 188 observations of AR(1) with parameter $\alpha = 0.8$ and D series.

The output of the CCF with AR(1) and series D was -0.010 (-0.030; 0.009) with a zero lag, indicating no correlation, and 0.429 (0.409; 0.449) when a lag of 12 observations was introduced representing a moderate correlation.

To visualize the WCC performance in the presence of different lags and scales, Figure 6 shows a heatmap representing the correlation under various conditions.

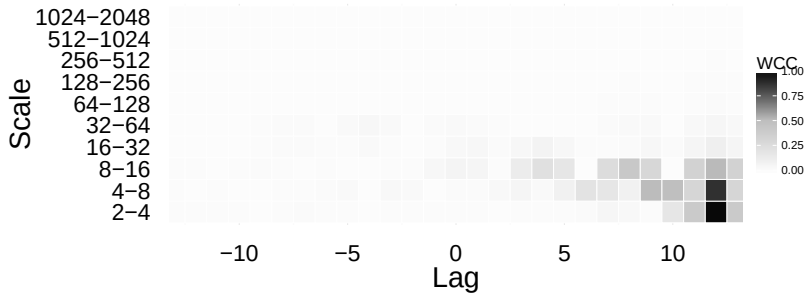

Figure 6: Estimated WCC for each scale and lag, considering the AR(1) with $\alpha = 0.8$ simulated series and the filtered series D.

The data shows that there is no correlation at the largest scales because these scales were removed. However, in the other scales, a strong correlation with a lag of 12 was also detected, which is accurate considering the shift of 12 months in series D. In addition, Figure 7 was created for a lag of 12 to better visualize the behavior and the CI for each scale.

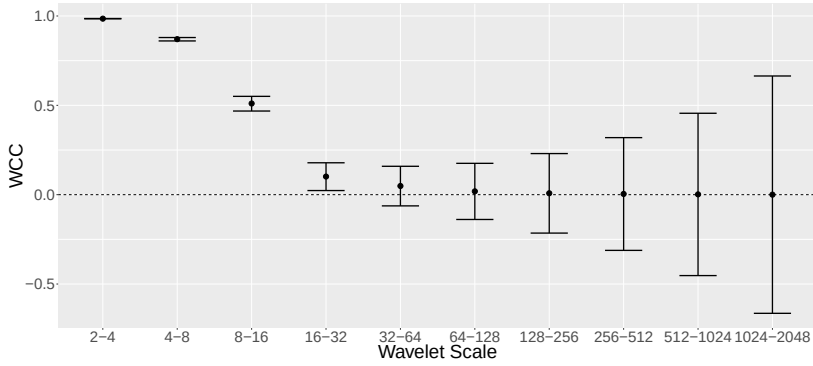

Figure 7: Estimated WCC for each scale and lag 12, considering the AR(1) with α simulated series and the filtered series D.

These findings reinforce the importance of using WCC when investigating the intercorrelation and resemblance between two time series or signals. In addition to providing more accurate point estimates with more precise CIs, the WCC estimator can provide estimates on multiple scales, allowing for the correlation between behaviors or components of two signals to be identified.

A heatmap is used to illustrate the point WCC estimates for different lags and scales for better visualization.

5 REAL DATA APPLICATION

To test the proposed approach on real data, a time series was constructed from the number of bronchiolitis cases in the Metropolitana and Maringá health divisions of Paraná state, Brazil. The number of bronchiolitis cases was taken as the number of patients hospitalized for bronchiolitis each month as reported in the DATASUS (Brazilian Unified Health System database) in the period from January 2002 to December 2012 (Figure 8).

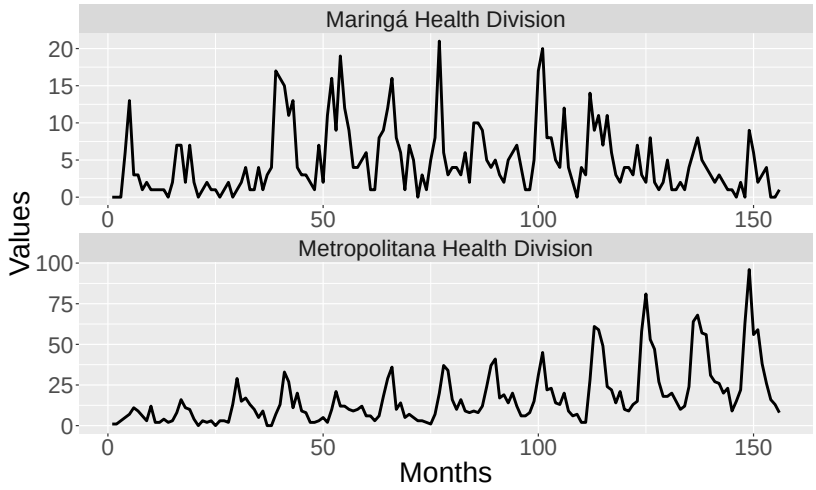

Figure 8: Time series of the number of bronchiolitis cases in the Metropolitana and Maringá health divisions of Paraná state from January 2002 to December 2012.

The results shown in Figure 8 for the Metropolitana Health Division include regular peaks, which indicate seasonality. Conversely, the time series data for the Maringá Health Division do not have consistent patterns and the correlation between the two divisions is not obvious. In Figure 9, the estimated WCC is presented to evaluate the correlation behavior on different scales and lags.

The results in Figure 9 demonstrate that as the lag approaches zero, strong correlations on the scale of 8-16 were observed, which includes the effects (periodicity) of 12 months. This means that both series have annual seasonality and, hence, are correlated. Other correlations were identified on the scale of 32-64. Further, with a lag of approximately 4, which may occur due to some another cyclical effect with periodicity larger than 3 years, the estimated WCC cannot be visually identified in Figure 8.

6 FINAL CONSIDERATIONS

In this study, the capability for investigating the correlation of time series on multiple scales was presented and evaluated. Different stochastic processes were analyzed and the use of the WCC

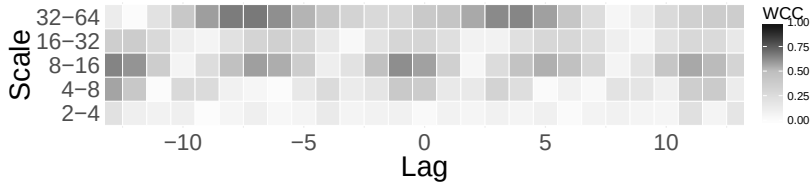

Figure 9: Estimated WCC for each scale and lag, considering the number of bronchiolitis cases in the Metropolitan and Maringá health divisions of Paraná state.

estimator demonstrated far better and more realistic results compared with the results obtained using the classical CCF estimator. Each simulated process was correlated with others having different frequency characteristics and different lags. In this study, we considered a series lagged by 12 observations (series D), but other lags were evaluated over shorter distances. Further, the correlation between signals with different lags was also correctly identified. Lastly, an application with real data to identify seasonal and other cyclical correlations was demonstrated by the WCC.

ACKNOWLEDGMENTS

The authors would like to thank the editor and the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. The authors also thank the financial support from the Brazilian National Council for Scientific and Technological Development (CNPq).

RESUMO. A estimação da correlação entre dados independentes usando estimadores clássicos, como o coeficiente de Pearson, está bem estabelecida na literatura. No entanto, tais estimadores são inadequados para analisar a correlação entre dados dependentes. Existem vários tipos de dependência, sendo as dependências serial (temporal) e espacial as mais comuns, as quais são inerentes aos dados de várias áreas. Usando uma análise de série temporal bivariada pode-se avaliar a relação entre duas séries. Além disso, como uma série temporal pode estar relacionada com outra em alguma defasagem de tempo (seja para o passado ou para o futuro), é essencial também considerar correlações defasadas. A função de correlação cruzada (CCF), que assume que cada série tem uma distribuição normal e não é autocorrelacionada, é usada com freqüência. No entanto, mesmo quando uma série temporal é normalmente distribuída, a autocorrelação ainda é inerente a uma ou ambas séries temporais, comprometendo as estimativas obtidos usando o CCF e suas interpretações. Como uma alternativa a este problema, a análise usando a correlação cruzada wavelet (WCC) foi proposta. O WCC é baseado na transformada wavelet não decimada (NDWT), a qual é uma transformação invariante à translação e decompõe dados dependentes em múltiplas escalas, cada uma representando o comportamento de uma banda de frequências diferente.

Para demonstrar a aplicabilidade deste método, dados simulados e reais de séries temporais de diferentes processos estocásticos foram analisados. Os resultados demonstraram que as análises baseadas em o CCF pode não representar a realidade; no entanto, o WCC pode ser usado para identificar corretamente a correlação em cada escala. Além disso, o intervalo de confiança (IC) para os resultados da análise do WCC foi estimado para determinar a significância estatística.

Palavras-chave: Análise Multiescala, Série Temporal, Correlação Cruzada, Transformada Wavelet Não Decimada.

REFERENCES

[1] P. Abry & P. Flandrin. On the initialization of the discrete wavelet transform algorithm. *IEEE Signal Processing Letters*, **1** (2) (1994), 32–34.

[2] G. O. N. Brassarote. "Análise multiescala de séries temporais do efeito da cintilação ionosférica nos sinais de satélite GPS a partir de wavelets não decimadas". 84 p. Master's thesis. São Paulo State University (UNESP) (2014). Available in: `http://hdl.handle.net/11449/116008`.

[3] G. O. N. Brassarote, E. M. Souza & J. F. G. Monico. Multiscale analysis of GPS time series from non-decimated wavelet to investigate the effects of ionospheric scintillation. *TEMA (São Carlos)*, **16** (2) (2015), 119–130.

[4] C. Chatfield. "The Analysis of Time Series: An Introduction". London, Chapman and Hall (1996).

[5] T. Conlon, H. J. Ruskin & M. Crane. Cross-correlation dynamics in financial time series. *Physica A: Statistical Mechanics and its Applications*, **388** (5) (2009), 705–714.

[6] H. Li, T. Nozaki. Application of wavelet cross-correlation analysis to a plane turbulent jet. *JSME International Journal Series B*, **40** (1) (1997), 58–66.

[7] G. Nason. "Wavelet methods in statistics with R". Springer Science & Business Media, Bristol, United Kingdom (2010).

[8] D. B. Percival & H. O. Mofjeld. Analysis of subtidal coastal sea level fluctuations using wavelets. *Journal of the American Statistical Association*, **92** (439) (1997), 868–880.

[9] D. B. Percival & A. T. Walden. "Wavelet Methods for Time Series Analysis". Cambridge Series in Statistical and Probabilistics Mathematics, New York, USA (2000).

[10] R Core Team. "R: A Language and Environment for Statistical Computing". Vienna, Austria (2016). Available in: `https://www.R-project.org/`.

[11] A. B. Rowley et al. Synchronization between arterial blood pressure and cerebral oxyhaemoglobin concentration investigated by wavelet cross-correlation. *Physiological measurement*, **28** (2) (2007), 161–173.

[12] G. Turbelin, P. Ngae & M. Grignon. Wavelet cross-correlation analysis of wind speed series generated by ANN based models. *Renewable Energy*, **34** (4) (2009), 1024–1032.

[13] P. Vielva, E. Martínez-González & M. Tucci. Cross-correlation of the cosmic microwave background and radio galaxies in real, harmonic and wavelet spaces: detection of the integrated Sachs-Wolfe effect and dark energy constraints. *Monthly Notices of the Royal Astronomical Society*, **365** (3) (2006), 891–901.

[14] B. Whitcher, P. Guttorp & D. B. Percival. Mathematical background for wavelet estimators for cross covariance and cross-correlation, *TR38*, Natl. Res. Cent. Stat. and Environ., Seattle (1999).

[15] B. Whitcher, P. Guttorp & D. B. Percival. Wavelet analysis of covariance with applications to atmospheric time series. *Journal of Geophysical Research: Atmospheres (1984-2012)*, **105** (D11) (2000), 14941–14962.

[16] H. Wickham. "ggplot2: Elegant Graphics for Data Analysis.". Springer-Verlag New York (2009).