Universidade Estadual de Maringá
Departamento de Estatística
Programa de Pós-graduação em Bioestatística

Aline Edlâine de Medeiros

Métodos bootstrap para séries temporais baseados em wavelets

Maringá-PR
Março de 2018

Aline Edlâine de Medeiros

Métodos bootstrap para séries temporais baseados em wavelets

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito para a obtenção do título de mestre em Bioestatística.

Universidade Estadual de Maringá – UEM

Departamento de Estatística

Programa de Pós-Graduação em Bioestatística

Maringá-PR

Março de 2018

Dados Internacionais de Catalogação na Publicação (CIP)
(Biblioteca Central - UEM, Maringá, PR, Brasil)

Medeiros, Aline Edlaine de
M488m Métodos bootstrap para séries temporais baseados em wavelets / Aline Edlaine de Medeiros. -- Maringá, 2018.
69 f. : il. color., figs., tabs.

Orientadora: Profa. Dra. Eniuce Menezes de Souza.
Dissertação (mestrado) - Universidade Estadual de Maringá, Departamento de Estatística, Programa de Pós-Graduação em Biestatística, 2018.

1. Séries temporais. 2. Bootstrap. 3. Walvelets. 4. Expoente de Hurst I. Souza, Eniuce Menezes de, orient. II. Universidade Estadual de Maringá. Departamento de Estatística. Programa de Pós-Graduação em Biestatística. III. Título.

CDD 21.ed.519.55
ECSL-1202/9

Universidade Estadual de Maringá
Centro de Ciências Exatas
Programa de Pós-Graduação em Bioestatística - Mestrado

PARANÁ
GOVERNO DO ESTADO

ALINE EDLAINE DE MEDEIROS

Métodos bootstrap para séries temporais baseados em Wavelets

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de Mestre em Bioestatística.

BANCA EXAMINADORA

Profª. Dra. Eniuce Menezes de Souza
Universidade Estadual de Maringá - UEM

Thelma Sáfadi

Profª. Dra. Thelma Sáfadi
Departamento de Estatística
Universidade Federal de Lavras

Profº Dr. Diogo Francisco Rossoni
Universidade Estadual de Maringá - UEM

Maringá, 28 de março de 2018.

Avenida Colombo, 5790, Bloco E90, sala 211 – Campus Universitário – CEP 87020-900 – Maringá – Paraná
Fone: (44) 3011-5450- http:/www.pbe.uem.b – email: sec-pbe@uem.br

Acknowledgements

I would like to thank all the professors of the graduate Program in Biostatistics, who teach me much of I know about Statistics, especially my advisor Eniuce, for all the scientific contribution, and unlimited patience. I am very grateful to Academic Master's Degree in Biostatistics[1] by the opportunity of to be here today, and to CAPES [2] for the financial support. I also would like to acknowledge all my masters colleagues for the many days of company and shared study, I can say that I learned a lot with all of you. And last, but not least, I would like to thank my parents Pedro and Lidia, my brother Edmilson and my boyfriend Marcelo for the comprehension and contribution in all the spheres of support, I really could not have done nothing without you all.

[1] <http://www.pbe.uem.br/?page_id=351&lang=en>
[2] <http://www.capes.gov.br/>

*"When I was 5 years old, my mother always told me,
that happiness was the key to life. When I went to
school, they asked me what I wanted to be when I grew up. I
wrote down 'happy'. They told me I didn't understand the assignment,
and I told them they didn't understand life.´´
(John Lennon)*

Resumo

A análise de Séries Temporais constitui uma importante área da Estatística, com a qual é possível explicar e predizer uma série de eventos, em particular fenômenos observados ao longo do tempo. Este estudo aborda conceitos e teorias pertinentes a área de Séries Temporais, em particular, aqueles que contemplam a Análise de ondaletas (wavelets). A junção destas duas teorias possibilita analisar mesmo os menores detalhes presentes em uma dada série temporal. Além disso, métodos bootstrap tem sido elaborados a partir desta junção, os quais são capazes de contemplar a autocorrelação presente nas observações de uma série temporal. Nesta dissertação são investigados os métodos existentes de bootstrap baseados em wavelets, e propostas algumas abordagens baseadas nas transformada wavelet decimada (DWT) e na transformada wavelet não decimada (NDWT), a qual é promissora devido a sua invariância por translação e inerente redundância. Além disso, propõe-se estender o método proposto para estimar a incerteza para a média evolucionária em curvas ou séries temporais, um problema ainda em aberto. Os métodos propostos foram avaliados quanto aos erros, erro quadrático médio, e preservação do Expoente de Hurst (H). Os resultados preliminares da análise de alguns mecanismos para estimar o expoente de Hurst indicaram que tais métodos são bastante sensíveis quanto a escolha de parâmetros de cortes, e precisam ser utilizados com cautela. Durante a análise da preservação do expoente de Hurst utilizou-se o método de Higuchi, o qual se mostrou uma métrica mais consistente. Os três métodos propostos e suas versões com fator de correção da correlação permitiram estimar intervalos de confiança para a média evolucionária de um processo estocástico, cujos resultados foram constatados em séries temporais simuladas. Neste contexto, o expoente de Hurst foi preservado, e tanto os erros quanto o erro quadrático médio tenderam a zero. Por fim, os métodos propostos foram empregados para estimar a incerteza associada a série temporal da taxa de hospitalizações por bronquiolite no estado do Paraná-BR (2000-2014), os quais apresentaram resultados satisfatórios.

Palavras-chave: Transformada wavelet não decimada, Bootstrap, Reamostragem, Intervalo de Confiança, Expoente de Hurst.

Abstract

The time series analysis constitutes an important area of statistics, which is possible to explain and predict a series of events, in particular, phenomena observed over time. This study approaches concepts and theories pertinent to the area on Time Series, especially, those that contemplate Wavelet Analysis. The combination of these two theories makes it possible to analyze even the smallest details present in a given time series. In addition, bootstrap methods have been elaborated from this junction, which is able to contemplate the autocorrelation present in the observations of a time series. This dissertation investigates the existing wavelet-based bootstrap methods and proposes some approaches based on Wavelet Transform (DWT) and the Non-Decimated Wavelet Transform (NDWT), which is promising because of its shift invariance and inherent redundancy. Furthermore, it is proposed to extend the proposed method to estimate the uncertainty for the evolutionary mean in curves or time series, a problem still open. The proposed methods were evaluated for errors, mean square error, and preservation of the Hurst Exponent (H). Preliminary results from the analysis of some mechanisms to estimate the Hurst exponent indicated that such methods are quite sensitive as to the choice of cut parameters and need to be used with caution. During the analysis of the preservation of the Hurst exponent, the Higuchi method was used, which proved to be a more consistent metric, between those who were analyzed. The three proposed methods and their versions with correlation correction factor allowed to estimate intervals of confidence for the evolutionary average of a stochastic process, such results were observed in simulated time series. In this context, the Hurst exponent was preserved, and both the error and the mean squared error tended to zero. Finally, the proposed methods were used to estimate the uncertainty associated with the time series of the hospitalization rate for bronchiolitis in the state of Paraná-BR (2000-2014), which presented satisfactory results.

Keywords: Non-decimated Wavelets, Multiscale Analysis, Boostraping, Resampling, Confidence Interval.

List of Figures

List of abbreviations and acronyms

ACF	Autocorrelation function
CWT	Continuous wavelet transform
DWPT	Discrete Wavelet Packet Transform.
DWT	Discrete Wavelet transform
H	Hurst exponent
IWT	Inverse wavelet transform
MODWT	Maximal overlap wavelet transform
MRA	Multi-resolution Analysis
NDWT	Non-decimated wavelet transform
PACF	Partial Autocorrelation function
SB	stationary Bootstrap
TSW	Two step wavestrapping
TSWDWT	Two step wavestrapping based in DWT
TSWNDWT	Two step wavestrapping based in NDWT
WT	Wavelet transform

Contents

Introduction

A time series is a sequence of observations usually ordered in equally spaced time intervals. The main purposes of studying time series are describing, modeling, and forecasting data from a stochastic process (CHATFIELD, 2013; WEI, 1994). Because of time series autocorrelation structure, maintaining the data order is of great importance, and some specific techniques need to be used for modeling or forecasting a time series. Depending on the time series features, such as non-stationarity and sparsity, several classical models are not an adequate approach. In this sense, the wavelet analysis have been a successful methodology.

Among the several advantages of wavelet methods are their scale/time adaptively to erratic fluctuations and non-linearity in time series as well as the excellent mean squared error properties when they are used for estimating functions containing non-stationarities and irregularities of different forms, such as cups or chirps (PICARD; TRIBOULEY, 2000). Because of that, wavelet regression, e. g., have being successfully used in several areas (GIMENES, 2015; DONOHO; JOHNSTONE, 1994; DONOHO; JOHNSTONE, 1995; NASON, 2010). The wavelet coefficients have less autocorrelation than the observed time series, and this allows applying methodologies that are only suitable for non-dependent data, such as bootstrap (GOLIA, 2002; TANG; WOODWARD; SCHUCANY, 2008).

So, in this dissertation, we are going to analyze the existent wavelet-based bootstrapping methods aiming to propose a method based on the non-decimated wavelet transform (NDWT) (NASON, 2010; VIDAKOVIC, 1999; NASON; SAPATINAS; SAWCZENKO, 1997; NASON; SILVERMAN, 1995; NASON; SACHS; KROISANDT, 2000), since the NDWT translation /shift-invariance and redundancy seem to be interesting properties to be combined with bootstrap.

Furthermore, wavelet-based bootstrapping methods have been used to estimate the CI to some statistic estimated from the time series. Although we are also going to do that, we aim to built the CI for the mean on each observed time. In this approach we are considering the time series as the mean of a stochastic process. Thus, the uncertainty would be available for the mean of such process, which can also be represented by a wavelet regression model where the error term can be estimated. This approach for this open problem would be very useful for several applications where wavelets are used for both modeling or forecasting with uncertainty estimated for each observed time.

However, some conditions need to be guaranteed so that the surrogate time series from boostrap method can be used to determine the IC, either for a statistic or for the mean of a stochastic process. Among them, is the autocorrelation or self-similarity parameter preservation (RESTA, 2012; KANG, 2016), which indicates that time series obtained by bootstrap has similar characteristics to the series originally analyzed.

Although the confidence interval (CI) estimation is of considerable interest of a range of applications, first we evaluated the development of an appropriate method to estimate CI using simulated time series, containing diverse types of behaviors: trend, long memory, among other characteristics. But, in a second moment, the introduced methods is applied in order to estimate the uncertainty of data on bronchiolitis in Paraná State - Brazil.

Thus, the aims of this dissertation can be expressed as:

- Estimation of the uncertainty for the evolutionary mean, μ_t, of a stochastic process using bootstrap resampling of wavelet coefficients;
- Studying methods for estimating Hurst exponent.
- Evaluating methods for resampling time series;
- Application of these methods to Biostatistics data.

This work is organized in articles as follows. Chapter 1 presents a literature review of the wavelet methods for dealing with time series, illustrating advantages and limitations of these methods in resampling time series. Chapter 2 discuss about the application of empirical methods for estimating Hurst exponent. Chapter 3 presents three wavelet-based bootstrapping and evaluates if the Hurst exponent is preserved after appling those methods. Chapter 4 contains an application of wavelet-based bootstrapping methods for estimating the uncertainty of the mean rate bronchiolitis time series [3]. In Chapter 5, general conclusions and discussions of this dissertation are presented.

[3] Chapter accepted for publication in: Frontiers of Biostatistics in Bioinformatics, for more information see Medeiros and Souza (2018).

Chapter 1

Literature Review

Suppose we have a parametric space T and a probability space (Ω, A, P), a stochastic process is a family $Z = \{Z(t), t \in T\}$, such that, for each $t \in T$, $Z(t)$ is a random variable (MORETTIN; TOLOI, 2006). Stochastic process is a central statistical approach to define the modern concept of time series. Indeed, a time series is considered as the finite realization of a stochastic process, in which each observed time series is a trajectory of a stochastic process.

Time series requires some special treatment because of its dependence structure. Some mechanisms are of great importance to deal with time series particularities such as autocorrelation function (ACF), the partial autocorrelation function (PACF) and Hurst exponent (H), which in general are used to analyze the presence of periodicity, long memory and to find the order of classical models as ARIMA model, among others (WEI, 1994; MORETTIN; TOLOI, 2006; CHATFIELD, 2013).

When the time series is stationary more options of approaches are available, but those with non-stationarity and/or long memory does not dispose of so many alternatives. In general, dealing with wavelet methods is an interesting alternative, since they contemplate all these characteristics.

Wavelets are functions located in time and frequency, simultaneously. These structure is an advantage when compared to other approaches to deal with time series, as Fourier analysis, which comprises only the frequency domain, or ARIMA models, which contemplate just the time domain. Furthermore, as well as the Fourier coefficients characterize the global behavior of a time series, which is important to analyze those time series containing periodicity, the wavelet coefficients comprise the local behavior of a time series, allowing the analysis of its transients and singularities (MORETTIN, 1999). In the next section, we present an overview of the main concepts of the wavelet theory.

1.1 Wavelet Theory

As mentioned by Percival and Walden (2006) and considering wavelets has been used since the eighties in Geophysics (DAUBECHIES, 1992), in many aspects wavelets are a synthesis of older ideas with new elegant mathematical results and efficient computational algorithms. In few words, a wavelet is a small localized wave (ψ) with some attractive mathematical properties (VIDAKOVIC, 1999). Specially, we can derive a family of daughter wavelets translating and dilating ψ which forms a wavelet base to represent a time series in the time and frequency domain (wavelet domain) simultaneously (VIDAKOVIC, 1999; NASON, 2010; KENDERDINE, 2012).

Given a function f, we say that f belongs to the space of all square-integrable functions, denoted by $\mathbb{L}_2(\mathbb{R})$, if

a. $\int |f|^2 < \infty$;

b. $||f|| = \sqrt{\int f^2}$;

c. $\langle f, g \rangle = \int fg$.

A wavelet is a function ψ belonging to $\mathbb{L}_2(\mathbb{R})$, satisfying the following conditions:

1. $\int_{-\infty}^{+\infty} \psi(x)dx = 0$;
2. $\int_{-\infty}^{+\infty} \psi^2(x)dx = 1$;
3. If $\hat{\psi}(f)$ is the Fourier transform of $\psi(x)$ then

$$C_\psi = \int_0^\infty \frac{|\hat{\psi}(f)|}{f} df.$$

The property 1 motivates the name wavelet (small wave) and the property 3 is called of admissibility condition (KENDERDINE, 2012; VIDAKOVIC, 1999).

For each wavelet ψ we have an associated function ϕ, called of scaling function. A classical example is the Haar wavelets, whose the scaling and wavelet functions are defined, respectively, by

$$\phi(t) = \begin{cases} 1 & \text{if } 0 \leq t < \frac{1}{2}, \\ -1 & \text{if } \frac{1}{2} \leq t < 1, \\ 0 & \text{otherwise} \end{cases} \qquad \psi(t) = \begin{cases} 1 & \text{if } 0 \leq t \leq 1, \\ 0 & \text{otherwise} \end{cases}$$

A family of the functions $\psi_{a,b}$ can be obtained scaling and translating the function ψ, that is

$$\psi_{a,b} = \frac{1}{\sqrt{a}} \psi \left(\frac{t-b}{a} \right) \tag{1.1}$$

with $a \in \mathbb{R}^+$ and $b \in \mathbb{R}$ (VIDAKOVIC, 1999). The family of functions $\psi_{a,b}$ represents the "daughter wavelets", and because of that, the functions ψ and ϕ are also called of "the mother" and "the father" functions, respectively. Although the daughter wavelets appear to be derived only from the mother wavelet, is possible to define them in terms of the father wavelet, as described by Aboufadel and Schlicker (2011).

Let $f(x)$ be any function of $\mathbb{L}_2(\mathbb{R})$, then

$$C(a, b; f(t), \psi(t)) = \langle f, \psi_{a,b} \rangle = \int f(t)\overline{\psi_{a,b}(t)}dt. \tag{1.2}$$

is the continuous wavelet transform (CWT) of f with respect to the wavelet ψ. For continuous case a family $\psi_{a,b}$ can be obtained translating and dilating the Haar function (VIDAKOVIC, 1999).

The CWT of a function has information redundancy, but this not always is a desirable characteristic. A classical approach to deal with this redundancy is performing a discretization, replacing a and b in Equation 1.1 by 2^{-j}, and $k2^{-j}$, respectively, where $j, k \in \mathbb{Z}$, what results in the decimated DWT with respect to the mother wavelet ψ. However, there are many others discretization forms, some of them are described by Daubechies (1992), Vidakovic (1999), Heinlein (2003), and Donoho and Candes (2005).

The choice of a particular wavelet transform (discrete or continue) depends on certain characteristics desirable in each context. In time series analyses DWT has been widely applied. However, NDWT (another discretization) also has been used because of its special features such as shift-invariance.

When a wavelet transform is applied, a natural decomposition of the function $f \in \mathbb{L}_2(\mathbb{R})$ occurs, what are called multiresolution analysis (MRA) (MALLAT, 1989). MRA can be represented by a sequence of closed subspaces $\{V_j, j \in \mathbb{Z}\}$ of $\mathbb{L}_2(\mathbb{R})$ such that the following properties are satisfied:

MRA1: (Nested interval) $\ldots \subset V_0 \subset V_1 \subset V_2 \ldots$

MRA2: (Density) $\bigcup_{j \in \mathbb{Z}} = \mathbb{L}_2(\mathbb{R})$;

MRA3: (Separation) $\bigcap_{j \in \mathbb{Z}} = \{0\}$;

MRA4: (Scale invariance) $f(t) \in V_j \Leftrightarrow f(2^j t) \in V_0, \ \ \forall j \in \mathbb{Z}$;

MRA5: (Translation invariance) $f(t) \in V_0 \Leftrightarrow f(t-k) \in V_0, \ \ \forall k \in \mathbb{Z}$;

MRA6: $\exists\, \phi \in V_j$ for which the set $\{\phi_{j,k}, k \in \mathbb{Z}\}$ is an orthonormal basis of V_j.

The MRA1 property indicates that a $f \in \mathbb{L}_2(\mathbb{R})$ function can be approximated in several resolution levels determined by the vector spaces Vj. The best approximation is obtained by considering the orthogonal projection of f (MORETTIN, 1999). The MRA2 property indicates that as the resolution 2^j increases, then the approximate function for f converges to its true value. Similarly, if the resolution decreases the approximation of f converges to the null function (MRA3). The MRA4 indicates that the details present at a resolution level 2^j must be present at the level 2^{j+1}. The MRA5 property indicates that moving the f function in k units causes no change in the resolution level. The last property (MRA6) is of great importance to calculate scaling and wavelets coefficients (see equation 1.4 and 1.5) (MORETTIN, 1999; VIDAKOVIC, 1999).

Considering we have a MRA, given a time series $Y = (y_0, y_1, \cdots, y_{n-1})$, with n observations, it can be represented as a function f in terms of the scaling function ϕ and wavelet functions ψ as

$$f(t) = \sum_{k=0}^{n-1} c_{J_0,k}\phi_{J_0,k}(t) + \sum_{j=J_0}^{J-1}\sum_{k=0}^{n-1} d_{j,k}\psi_{j,k} \tag{1.3}$$

where $J-1 < \log_2 n \leq J$, $j = J_0, \cdots, J-1$ representing a multiresolution level, and $k = 0, \cdots, n-1$. Taking $\phi_{J_0,k}(t) = 2^{J_0/2}\phi(2^{J_0}(t-k))$, and $\psi_{j,k}(t) = 2^{j/2}\psi(2^j(t-k))$, the coefficients $c_{J_0,k}$ and $d_{j,k}$ called the smooth (scaling) and detail (wavelet) coefficients, respectively, comprise the NDWT of the time series Y. Analogously, Taking $\phi_{J_0,k}(t) = 2^{J_0/2}\phi(2^{J_0}t-k)$, $\psi_{j,k}(t) = 2^{j/2}\psi(2^jt-k)$, the coefficients $c_{J_0,k}$ and $d_{j,k}$ comprise the DWT of the time series Y.

The wavelet and scaling coefficients $c_{J_0,k}$ and $d_{j,k}$ are given by:

$$c_{J_0,k} = < f(x), \phi_{J_0,k} >= \int_{-\infty}^{+\infty} f(t)\phi_{J_0,k}dt \tag{1.4}$$

$$d_{j,k} = < f(x), \psi_{j,k} >= \int_{-\infty}^{+\infty} f(t)\psi_{j,k}dt \tag{1.5}$$

where $<,>$ is the usual internal product operator. To pass from the time series domain to the wavelet domain (obtaining the wavelet coefficients) is necessary to use some decomposition algorithm. A well-knowing algorithm for calculating these coefficients is the Pyramidal or Cascade Algorithm (MALLAT, 1989). Given a wavelet function $\psi \in \mathbb{L}_2(\mathbb{R})$ and $m \in \mathbb{R}$ so that $m \geq 1$, we say that ψ has m vanishing moments if

$$\int_{-\infty}^{+\infty} t^l\psi(t)dt = 0,\ l = 0, 1, \ldots, m-1, \tag{1.6}$$

under certain technical conditions (NASON, 2010). The number of vanishing moments of a wavelet function is an important argument to consider in decision of the wavelet family, because for m big enough the correlations between coefficients within and among levels decay rapidly, on the other hand, increasing the number of vanishing moments also may have undesired effects, specially, when DWT is being used (BREAKSPEAR; BRAMMER; ROBINSON, 2003).

1.2 Wavelet based bootstrap methods for resampling time series

Wavelet Analysis have been successfully applied to deal with time series and its characteristics, such as seasonal and/or cyclic effects, trend, irregular fluctuations, non-stationarity, with long memory, among others. The non-parametric regression using wavelet multiscale decomposition is an alternative in situations where classical models can not be applied. Furthermore, wavelets can also be used for the estimation of spectral densities of stationary time series and of time-varying spectra using localized periodogram based estimators (PERCIVAL; WALDEN, 2006).

Although DWT coefficients have desirable features, as weakly autocorrelation between nearby coefficients, among its limitations are the applicability only for power-of-two time series size, variant with respect to translations and posses only a few coefficients to be resampled in each decomposition level (KENDERDINE, 2012). In conditions where translation or shift-invariance is important, the non-decimated wavelet transform (NDWT) (NASON; SILVERMAN, 1995) is a good alternative. NDWT has the same number of wavelet coefficients in each resolution level, overcoming the DWT limitation of few coefficients to be resampled. Furthermore, NDWT is more flexible with respect to the time series length, being appropriate for all those which are a multiple of two, and has an easy implementation with more than one algorithm, including the pyramidal algorithm (MALLAT, 1989).

Operationally, the pyramidal algorithm performs of NDWT consists of applying DWT twice at each stage, wherein the second round is applied to the coefficients at previous level translated by a single index. The coefficients from each DWT can be considered almost independent and normally distributed. In addition to DWT, NDWT and CWT already discussed, some other important WT have been reported in the literature such as discrete wavelet package (DWPT) (NASON; SILVERMAN, 1995; VIDAKOVIC, 1999; KENDERDINE, 2012). At first, such transforms could be used for bootstrapping time series. However, throughout this text, some limitations for the use of most of these transforms will be presented.

Politis and Romano (1994) developed a resample method called Stationary Bootstrap (SB), which is propitious to deal with stationary weakly dependent time series. Whose the pseudo-time series generated is a stationary time series, which is an advantage when compared with the methods developed earlier in the time domain.

The existent time and frequency domain bootstrapping approaches to access the uncertainty of estimates in time series, such as parametric, residual, block bootstrapping, and periodogram based bootstrapping, they also are mainly designed for stationary series with short-range dependence and are not adequate for those time series exhibiting long-range dependence (KENDERDINE, 2012). Furthermore, parametric and frequency-domain bootstraps work best for series that follow a Gaussian distribution, but can be problematic for time series exhibiting non-Gaussianity (PERCIVAL; SARDY; DAVISON, 2000).

Non-Gaussian series are better handled by block bootstrapping, but the quality of this approach depends critically on chosen size of the blocks. By taking these difficulties into account, Percival, Sardy and Davison (2000) proposed a DWPT wavestrapping method which is an adaptive wavelet-based scheme for bootstrapping time series that can be modeled by either stationary short or long-memory processes. Since then, some other developments have been proposed in wavelet-based bootstrapping methods.

Golia (2002) applied the Stationary Bootstrap (SB) to the wavelet coefficients of time series exhibiting long memory. This application was possible because the wavelet coefficients are wide-sense stationary and weakly correlated in each scale (WORNELL; OPPENHEIM, 1996). In her work, she used the Daubechies wavelet with 4 vanishing moments and coarsest level of details equals to 4 in DWT, the results were good, however the author comments on the need of evaluating this approach for other long memory processes.

In Breakspear, Brammer and Robinson (2003) the resampling of time series in the wavelet domain revealed to have all the desired properties of a nonlinear surrogate technique, including the preservation of linear properties, multiple possible realizations, removal of nonlinear structure, acceptable computational demands and extension to multivariate cases. They compared three classes of resampling wavelet coefficients within each scale: free permutation; cyclic rotation and block resampling. The results showed that block resampling of wavelet coefficients optimizes the mentioned properties in comparison to the other wavelet resampling schemes.

In Angelini et al. (2005) a DWT wavelet based resampling scheme was presented and compared to the traditional Fourier based phase randomization bootstrapping within the context of turbulence energy cascades. The comparison between two the resampling methods and observed ensemble statistics constructed by clustering similar meteorological conditions demonstrated that the wavelet method reproduced several features related to intermittency of the ensemble series, what did not happen with the Fourier based method. Feng, Willemain and Shang (2005) compared the wavelet-based bootstrap with the time domain moving block bootstrap for estimating the standard errors of the unit lag sample autocorrelation and the sample standard deviation. The results showed the achieved performance of wavelet-based bootstrap as better than the moving block bootstrap for both short-range and long-range dependent data.

In Keylock (2006), Keylock (2007), Keylock (2008), a method is presented for generating surrogates that preserve the local mean and variance of the original time series. The author discusses the difficulty of a method consisting of simple shuffling of wavelet coefficients that does not preserve the inherent periodicity of the wavelet coefficients. To overcome this difficulty, he used the iterated amplitude adjusted Fourier transform (IAAFT) to randomize the wavelet coefficients at each level of the non-decimated wavelet transform (NDWT) or Maximal Overlap Discrete Wavelet Transform (MODWT) as in Percival and Walden (2006).

In Yi et al. (2007) a two-step wavestrapping method was introduced and applied to complex time series acceleration data from mobile computing users. This method consists of applying the Stationary Bootstrap (POLITIS; ROMANO, 1994) in one-step of DWT, which is called of Stationary Parallel Bootstrapping (first step), followed by an adjustment of trend and energy (second step). Two obstacles of dealing with acceleration data, that had not adequately be accounted for in the prior wavelet-based approaches were solved by Yi et al. (2007). First, the vertical correlation of wavelet coefficients among scale levels, which yields less disruptive surrogate data. Second, since the block length of the wavestrap method cannot cover the general trend of the original data adequately, the general trend of the acceleration data can be broken. With the two-step wavestrapping, the vertical relationship among levels could be preserved since the scaling and wavelet coefficients on the same time frame are resampled together. The second problem was overcome by a proposed energy adjustment techniques.

Kenderdine (2012) has discussed a method of combining DWPT with Wavelet Lifting that was used to decompose time series into independent components. However, there are some challenges and difficulties related to these methods. One question is related to the correlations between the wavelet coefficients, we know that it becomes small even if the time series itself is highly autocorrelated in the time domain correlation (GOLIA, 2002; YI et al., 2007). Actually, this correlation decays rapidly with increasing the number of vanishing moments of the wavelet filter. On the other hand, high-order wavelets with larger supports may produce more undesirable boundary artifacts (YI et al., 2007). Hence, the choice of vanishing moments depends on the properties of the data, with narrow support for weakly correlated or short data sets and larger support for strongly correlated or long data sets (BREAKSPEAR; BRAMMER; ROBINSON, 2003).

Considering that wavelet-based bootstrapping methods have been used to estimate CI for parameters from time series, Tang, Woodward and Schucany (2008), discussed about the lack of coverage of parameters for simple linear regression for time series data using simulated data and DWT. The authors also presented a parametric wavelet bootstrap that showed better results only for the simulated white noise stochastic process, however, this method is not applicable in general. Furthermore, the authors indicated the possibility of better results by resampling DWPT. As aforementioned one of the difficulties of resampling wavelet coefficients in DWT is the fact that there are only few coefficients to be resampled in the so-called smooth resolution levels. With the DWPT, where both smooth and detail coefficients are decomposed in each level promises to be more advantageous than the DWT since all decomposition levels have the same number of coefficients as the original time series. On the other hand, the shift-invariance is an important characteristics to deal with wavestrapping. In this sense, the NDWT is shift-invariant and it preserves the original length of the data in all decomposition levels, seeming to be a promising choice.

Although of the redundancy of the wavelet coefficients generated by NDWT implies higher intra-level correlation, one can identify in which levels of decomposition the bootstrap can be applied without any intervention to reduce the correlation. The first level of decomposition is an example where we can do it since the coefficients of this level are close to pure noise (DAUBECHIES, 1992) even for the NDWT.

Kang and Vidakovic (2017) proposed a method called MEDLA, which reduces the autocorrelation in NDWT decomposition levels. The MEDLA consists on resampling m random pairs of wavelets coefficients keeping the distance between them at least q_j, wherein $q_j = 2^{J-j}$, and take the logarithm of an average of the two squared wavelet coefficients in each one of the pairs. Furthermore, these methods presented some estimates for the Hurst statistics (H). When compared to standards approaches Kang and Vidakovic (2017) methods presenting smaller MSE. Feng and Vidakovic (2017) presented a robust method for estimating Hurst exponent, where trimean estimator is applied on NDWT coefficients, this method reduces the variance of the estimators in most cases. These studies show that is possible to use NDWT to obtain good estimates.

Chapter 2

Hurst Exponent Estimation

Introduction

Long memory (LM), long-range dependence (LRD), long-range persistence (LRP) or fractal process are widely observed in many fields such as Economy, Physics, and Statistics. These process comprises an important subject of study of the series analysis. While in the classical time series approaches as autoregressive moving average model (ARMA) the autocorrelation function (ACF) decreases rapidly to zero, in LM process this decay occurs slowly (hyperbolically) (MORETTIN; TOLOI, 2006).

Although long memory process are reported in many periods of the human history, including the biblical passage: "seven years of great abundance" and "seven years of famine", whose could be understood as a strong serial correlation, both the studies and the development of methods to work with such processes had ascension in the 20th century (BERAN et al., 2013).

Among the many contributors to the development of techniques for analysis of long memory time series is the hydrologist Hurst (HURST, 1951), who developed a measure called of Hurst exponent (H). This statistics indicates the intensity of the present autocorrelation in a time series. Followed by Mandelbrot (1982), which related the Hurst coefficient to the fractal dimension, Beran (1994), which gathered the memory long theory, as well as provided codes for estimating Hurst exponent in the Splus software, and Taqqu, Teverovsky and Willinger (1995) who compared the main methods for estimating the Hurst statistics.

Currently, the Hurst coefficient is frequently used in the time analysis. However, most of the methods for estimating H are empirical. Beran et al. (2013) suggests that such methods should be used with caution because these "methods involve tuning (or cut-off) parameters that are usually based on a subjective visual impression", i. e., for different parameter choices "one may arrive at completely different conclusions for the same data set", and even if these difficulties are supered "the statistics used in the heuristic methods have poor convergence properties".

In this paper, we intend to analyze the performance of the classical empirical methods in estimating H statistics. Although there are some easy implementation available for estimating the Hurst exponent (BERAN, 1994; VEDOVATTO, 2015) we are going to use those shared in the statistical software R Core Team (2016). The initial proposal is to show how such methods can lead to different results starting from the same cut-off parameters. At a second moment, we are going to use the graphical method to estimate the Hurst exponent for each one of the evaluated estimation methods.

To analyse the quality of this estimation we simulated time series of different lengths and Hurst measure, representing a Fractional Gaussian Noise (FGN). For each time series, the statistic H was estimated using nine methods.

This paper is structured as follows. Section 2.1 contains a brief description of the main Hurst estimation methods. Section 2.2 presents the computational aspects of this paper. Results are discussed at Section 2.3. Finally, in Section 2.4 are the final considerations of this study.

2.1 Methods for estimating Hurst exponet

As aforementioned, who first observed the H statistics was Hurst (1951), which resulted in the creation of the rescale range statistic, usually called of R/S statistics. For estimating R/S statistics of a given time series $Y = (y_1, \ldots, y_p)$ containing p observations, one needs to follow procedure below (RESTA, 2012; KHAREL, 2010):

1. Calculate the time series of returns $R = \{r_t,\ t = 1, \ldots, p-1\}$, where r_t is given by
$$r_t = \frac{y_{t+1} - y_t}{y_t}. \tag{2.1}$$
2. Divide R into k sub series with length n.
3. For each sub-series estimate the mean value (E_m) and the standard deviation (S_m), where $m = 1, \ldots, k$.
4. Normalize the data by subtracting the sample mean:
$$Z_{a,m} = R_{a,m} - E_m,\ a = 1, \ldots, n. \tag{2.2}$$
5. Determine the cumulative time-series $Y_{a,m}$
$$Y_{a,m} = \sum_{d=1}^{a} Z_{d,m},\ a = 1, \ldots, n. \tag{2.3}$$

6. Find the range:

$$R_m = \max\{Y_{1,m}, \ldots, Y_{a,m}\} - \min\{Y_{1,m}, \ldots, Y_{a,m}\} \tag{2.4}$$

7. Obtain the mean value of the rescale range for the sub-series, given by

$$(R/S)_n = \frac{1}{k} \sum_{m=1}^{k} \frac{R_m}{S_m} \tag{2.5}$$

8. Repeat the steps 1 through 7 increasing the n value, for all possible integers divisors of $p-1$.

9. Plot $(R/S)_n$ statistics against $\log(n)$ and use a simple regression to estimate the slope, which is an estimative for H;

If the slope is between 0 and 0.5 we have an anti-persistent (short memory), and for H between 0.5 and 1 the process is persistent (long memory). However, Taqqu, Teverovsky and Willinger (1995) suggest that to obtain reliable estimates one should not use the low and very high end of the graph to estimate H. Instead, the authors recommend setting a cut-off point. Then, these values between the lower and the upper cut-off pints should be used to estimate H.

In addition to the R/S statistics, other methods of frequent use in the literature for estimating the Hurst exponent are: aggregated variance method, differenced aggregated variance method, aggregated absolute value (moment) method, Higuchi's or fractal dimension method, Peng's or variance of residuals method, periodogram method, and boxed (modified) periodogram method. A description of these methods are presented in Taqqu, Teverovsky and Willinger (1995), Beran et al. (2013), Wuertz, Setz and Chalabi (2013).

For simplifying the referencing of the aforementioned methods we are going to follow the notation described in the Table 1.

Table 1 – Notation for Hurst exponent estimation methods

Methods	Notations
Aggregated Variance	E1
Differenced Aggregated Variance	E2
Aggregated Absolute Value	E3
Higuchi	E4
Variance of Residuals	E5
R/S	E6
Periodogram	E7
Modified Periodogram	E8
Wavelet	E9

2.2 Simulation

Currently, there is a diversity of methods used for calculating the Hurst exponent. The statistical software R (R Core Team, 2016) contain many packages that possess different ways of estimating it. In particular, the package fArma (WUERTZ; SETZ; CHALABI, 2013), which we are going to use in this paper, posses 9 methods.

For illustrating the performance of these methods, some time series with sizes: 128, 512, and 2048 were simulated, representing Fractional Gaussian Noise (FGN). For each time size three time series contemplating the H values: 0.35, 0.55, 0.75, and 0.95 were simulated by Beran (1994) method. Time series of size 128 representing FGN for H equal $0.35, 0.55, 0.75,$ and 0.95 can be seen, respectively, in Figure 1.

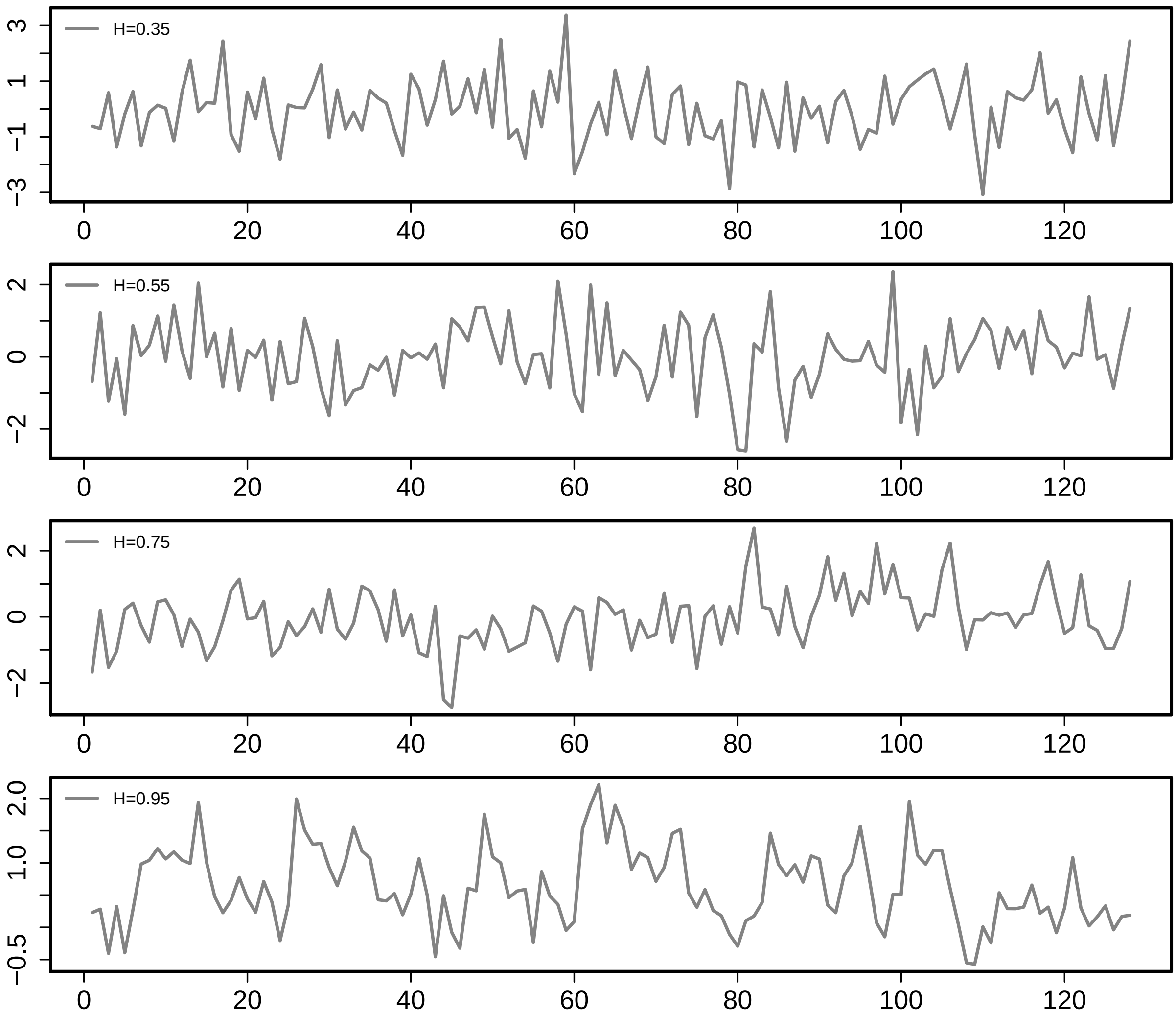

Figure 1 – Time series of length 128, generated from H equal to 0.35, 0.55, 0.75, and 0.95, respectively.

2.3 Results

Table 2 represents the estimates obtained from the functions of the R package fArma. Except for the methods E7, E8, and E9, which have distinct mechanisms for estimating H, all the other needs of cut-off points. For this first overview we use the default cut-off points of the R package aforementioned.

Table 2 – Hurst exponent estimated from FGN

TS length / H method	E1	E2	E3	E4	E5	E6	E7	E8	E9
L=128, H=0.35	-0.104	0.697	-0.000	0.411	0.392	0.863	0.274	0.300	0.258
L=128, H=0.55	0.252	0.878	0.342	0.535	0.593	0.987	0.547	0.566	0.409
L=128, H=0.75	0.900	1.425	1.083	0.832	0.594	0.572	1.127	0.602	0.301
L=128, H=0.95	0.129	0.720	0.247	0.968	0.780	0.781	1.090	1.003	1.009
L=512, H=0.35	0.484	0.726	0.558	0.402	0.425	0.531	0.474	0.266	0.309
L=512, H=0.55	0.287	0.767	0.372	0.287	0.487	0.662	0.260	0.365	0.560
L=512, H=0.75	0.458	0.766	0.578	0.852	0.741	0.886	0.689	0.693	0.889
L=512, H=0.95	0.645	1.143	0.735	0.942	0.923	0.863	1.141	0.855	0.745
L=2048, H=0.35	0.335	0.400	0.367	0.343	0.325	0.369	0.356	0.272	0.372
L=2048, H=0.55	0.474	0.621	0.500	0.488	0.562	0.586	0.627	0.502	0.553
L=2048, H=0.75	0.662	0.879	0.688	0.635	0.721	0.756	0.729	0.653	0.739
L=2048, H=0.95	0.768	1.053	0.785	0.915	0.891	0.929	0.952	0.891	0.908

From Table 2 we can notice that H estimates can be very different from the simulated values. In general, the methods underestimate or overestimate a lot the H value, and in some cases values greater than one are observed. For $H = 0.35$ and time series size 2048, the estimated values are more consistent with the real value, but, this does not occur with the other simulated values of H. This variability could be explained by the assumption of cut-off points for all the methods. So, on Table 3 we present the results of doing a graphical analysis for each one of the Hurst exponent estimation method. We graphically analyze the choice of cut-off points, block-size and all the available arguments of each method.

Table 3 – Hurst exponent estimated from FGN by using graphical analysis

TS length / H method	E1	E2	E3	E4	E5	E6	E7	E8	E9
L=128, H=0.35	0.363	0.343	0.367	0.363	0.361	0.346	0.370	0.315	0.355
L=128, H=0.55	0.578	0.579	0.552	0.553	0.552	0.592	0.547	0.559	0.582
L=128, H=0.75	0.751	0.825	0.762	0.832	0.751	0.742	0.764	0.581	0.692
L=128, H=0.95	0.840	0.819	0.940	0.957	0.970	0.945	0.816	0.973	0.999
L=512, H=0.35	0.352	0.359	0.452	0.399	0.387	0.484	0.352	0.365	0.310
L=512, H=0.55	0.503	0.551	0.507	0.405	0.560	0.616	0.548	0.559	0.562
L=512, H=0.75	0.688	0.799	0.656	0.763	0.742	0.738	0.756	0.746	0.786
L=512, H=0.95	0.853	0.933	0.866	0.922	0.937	0.929	0.909	0.953	0.949
L=2048, H=0.35	0.341	0.350	0.362	0.348	0.348	0.365	0.356	0.349	0.349
L=2048, H=0.55	0.533	0.541	0.540	0.542	0.566	0.566	0.550	0.553	0.553
L=2048, H=0.75	0.713	0.762	0.706	0.646	0.734	0.761	0.748	0.745	0.751
L=2048, H=0.95	0.821	0.948	0.814	0.901	0.946	0.939	0.952	0.941	0.966

Comparing Table 2 and 3 we could notice the difference generated by the choice of block sizes and cut-off points. Since the choice followed a graphical analysis in determining the regression line, if other values, for example, of the block-size is established other estimates will be generated, that may be better or worse than those presented in Table 3.

Clearly, the results presented are more close of the simulated Hurst exponent values. It can be observed that some methods seem to be more affected by the cut-off points and the block-size than others, e. g., the Differenced Aggregated Variance method, which presented inclusive values larger than one but presented similar values to those simulated after the study of these measures.

Figure 2 represents adjusted regression line by using the R/S statistics for simulated time series with size 2048 and generated from $H = 0.55$. Analogous graphics were generated for all the estimation methods for the Hurst exponent. In Figure 2 also is possible to observe the cut-off points.

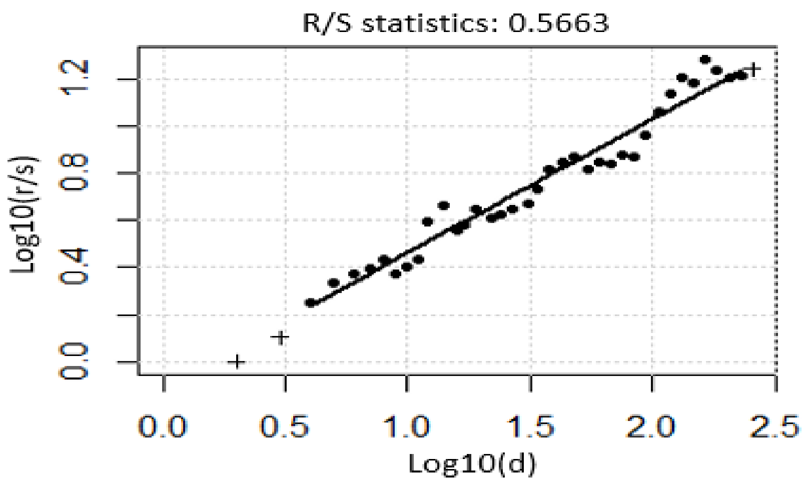

Figure 2 – Estimated regression line for simulated time series with size 2048 and generated from $H = 0.55$, using the R/S statistics.

2.4 Final Considerations

Using simulated time series we estimated Hurst exponent from 9 distinct methods implemented in the R package fArma. The estimation obtained from default procedures was quite divergent from each other. From a graphical analysis, where cut-off points and block-size were selected, the true values used in the simulation were approximated by estimated values. An important point to emphasizes is that in simulation procedures, for example, estimating the Hurst exponent by the graphical method is almost impracticable. In synthesis, methods for estimating Hurst exponent are important to understand the intensity of auto-correlation present in a time series, but, mainly the empirical methods need to be used with caution.

In order to verify if a high number of simulations would yield more accurate estimates in the default procedure, 1000 time series were simulated for each time series size and H value reported in this text. However, the estimates were not satisfactory and similar to those observed when a single time series was simulated for each time series size and H value.

The methods of Higuchi, Variance of Residuals and R/S even in the default procedure presented values more acceptable to the real ones, without values outside of $[0, 1]$. Some discussions about solutions for determination of cut-off points are presented in Beran et al. (2013). Another possibility is to opt for more robust methods to estimate H, e. g., Feng and Vidakovic (2017) method.

Chapter 3

Does bootstrap based on wavelets preserve the Hurst Exponent?

Abstract

In recent years, resampling techniques have been used in several areas of Statistics. However, most of these methodologies were developed for uncorrelated data, and areas such as Time Series, wherein the observations present autocorrelation, have required the development of techniques that take autocorrelation into account. Important methods of resampling for time series consists of combining the decimated wavelet transform (DWT) and a resampling technique (bootstrap) applied to its decomposition levels. Bootstrapping can be applied because the coefficients obtained through the DWT show less correlation than the observations in the time domain. In this paper, we implement three wavelet-based bootstraps (wavestrap) the first based on DWT, and the others based on non-decimated wavelet transform (NDWT). NDWT presents some advantages compared to DWT as the free length of time series, shift-invariance and the same number of wavelet coefficients in each multi-resolution level, what facilitates the bootstrap. Although successful applications of these methods have being presented in the literature, some questions still remain open. Among them, the preservation of the autocorrelation inherent to the replicated time series, in which the intensity is measured by estimating Hurst exponent (H). So, in this paper, we aim to analyze the performance of these methods on preserving the H statistics. To evaluate this performance we simulated 24 time series composed by different behaviors, and lengths $64, 128, 256, 512, 1024, 2018$. Their Hurst exponent were estimated using Higuchi's method before the resampling process. So, 5000 resamples were generated for each time series and method, as well as the Hurst exponent of each time series replica. The results obtained from bootstrap were compared to the estimates computed from the original time series using the errors and Mean Square Error (MSE) measures. The results point that the available methods properly presents similar H value as the original time series, indicating that the methods successful surrogate the time series behavior.

Keywords: Hurst exponent; Bootstrap; Wavelets; Time series.

Introduction

An univariate time series is a sequence of observations usually ordered in equally spaced time intervals, which represents a trajectory of a stochastic process. The analyses of time series comprises an useful tool for explaining and solving problems in several areas such as Economy, Geology, Statistics, Engineering, among others (WEI, 1994; BOX et al., 2015).

Hardly accessing more than one trajectory of a same stochastic process is possible. Even in the case of being possible to observe more time series for describing a specific experiment, as acceleration data (YI et al., 2007), researchers have to deal with other aspects such as costs, time, and human demand. Thus, techniques for bootstrapping time series, replicating more trajectories of a stochastic process, and considering the autocorrelation not contemplated by classical bootstrapping methods has been formulated (BÜHLMANN et al., 1997; LAHIRI, 2013).

In the latest decades, several approaches for resampling time series were developed based on decorrelation property of the wavelet transform (GOLIA, 2002; PERCIVAL; SARDY; DAVISON, 2000; YI et al., 2007). Essentially, the multiresolution decomposition of a time series produces detail and wavelet coefficients arranged in levels. The wavelet-based techniques for bootstrapping time series propose to resample these coefficients with classical or sophisticated methods. So, a replica of this time series is obtained applying the inverse transform.

In general, the Discrete Wavelet Transform (DWT) is chosen to perform wavelet-based bootstrapping (BREAKSPEAR; BRAMMER; ROBINSON, 2003; GOLIA, 2002; YI et al., 2007; TANG; WOODWARD; SCHUCANY, 2008). However, these approaches have limitations, especially, the few number of coefficients available at higher levels of the decomposition difficult the bootstrap procedure.

Recently Medeiros and Souza (2018) investigated three resampling methods: Naive bootstrapping based on Non-decimated Wavelet Transform (NDWT), DWT two step wavestrapping (TSWDWT), and NDWT two step wavestrapping (TSWNDWT). TSWDWT is a particular case of (YI et al., 2007) method, and the others two methods are wavelet-based proposals using NDWT, which instead of DWT preserve the same number of coefficients in each decomposition level.

In this paper, we aim to verify if the methods presented in Medeiros and Souza (2018) preserve the Hurst exponent (H). This measure is used for explaining self-similarity features and providing the intensity of autocorrelation inside a given data set. Preserving the H value after resampling means to maintain the original autocorrelation characteristics of an evaluated time series.

This paper works with 24 simulated time series, which were resampled 5000 times for each method. Hurst exponent was compared before and after resampling by checking the errors and MSE of its estimator. The results point that the evaluated methods properly preserve the autocorrelation features of the analyzed time series.

The paper is organized as follows. Session 3.1 presents a brief review of the three employed wavelet-based bootstrapping. Session 3.2 describes the method for estimating Hurst exponent, followed by evaluation criteria of this statistics in session 3.3. Simulation procedures are described in Session 3.4. The main results and discussions are presented in Session 3.5. Final considerations are mentioned in Session 3.6.

3.1 Wavelet-based technique for resampling time series

The literature provides a range of books that developed in details the wavelet theory and application, such as Vidakovic (1999), Percival and Walden (2006), and Nason (2010). In summary, the wavelet analysis approximates a time series by a linear combination of wavelet functions. Such combinations are obtained by dilations and translations of scaling and wavelet functions, denoted by ϕ and ψ, respectively. In other words, a time series $Y = (y_0, y_1, \cdots, y_{n-1})$ can be represented as a function f belonging to the space of all the square-integrable functions, as follows

$$f(t) = \sum_{k=0}^{n-1} c_{J_0,k}\phi_{J_0,k}(t) + \sum_{j=J_0}^{J-1}\sum_{k=0}^{n-1} d_{j,k}\psi_{j,k} \tag{3.1}$$

where $J-1 < \log_2 n \leq J$, $j = J_0, \cdots, J-1$ representing a multiresolution level, and $k = 0, \cdots, n-1$ (KANG; VIDAKOVIC, 2017).

Considering the wavelet base generated by ψ, $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$, the obtained coefficients $c_{J_0,k}$ and $d_{j,k}$ represent the DWT of the time series Y. Analogously, choosing the wavelet base generated by ψ, $\psi_{j,k}(t) = 2^{j/2}\psi(2^j(t-k))$, the NDWT of time series Y is comprised by $c_{J_0,k}$ and $d_{j,k}$ coefficients. The coefficients $c_{J_0,k}$ and $d_{j,k}$ are called the smooth and detail coefficients, which can be efficiently assessed using the Pyramidal Algorithm, developed by Mallat (1989).

Based on the works of Golia (2002) and Yi et al. (2007), which proposed the Stationary Bootstrap and the Parallel Wavestrap, respectively, Medeiros and Souza (2018) analyzed two methods TSWDWT and TSWNDWT, which are an extension of both the aforementioned methods.

For implementing TSWDWT one must reproduce the Parallel Wavestrap for a single time series Y, which the length is a power of two, and following the next procedure:

1. Decomposing Y using DWT to obtain the coarsest level of detail and wavelet coeffi-

cients;

2. Resampling detail and wavelet coefficients using Stationary Bootstrap;

3. Applying the Inverse Discrete Wavelet Transform (IDWT) to obtain a replica Y_b of Y;

4. Implementing an energy and trend adjustment:

    - Trend adjustment consists of decomposing both the time series Y and its surrogate Y_b using the DWT. Then, the scaling coefficients c_b obtained from Y_b are surrogated by the scaling coefficients c generated from Y;

    - The energy adjustment consists of generating the average energy for each decomposition level of Y and Y_b, denoted by $\bar{e}_j$. So, adjusting the average energy for each scale level of Y_b to the average energy of the Y levels:

    $$d_{baj,k} = d_{bj,k}\sqrt{\frac{\bar{e}_j}{\bar{e}_{bj}}}, \tag{3.2}$$

    where d_{ba} represents the adjustment made in each decomposition level of Y_b;

5. The steps 1, 2, and 3 are repeated a sufficient great number of times.

Analogously, TSWNDWT procedure follows the same steps of TSWDWT replacing DWT by NDWT. Due to the use of NDWT, this method does not require a power of two length for Y.

The third approach contemplated by Medeiros and Souza (2018), the naive bootstrap based on NDWT, for a given time series Y consists of:

1. Decomposing Y using NDWT for obtain detail and wavelet coefficients;

2. Resampling detail coefficients using the classical (naive) bootstrap;

3. Applying INDWT to obtain a replica Y_b of Y;

4. The steps 1, 2, and 3 are repeated a sufficient great number of times.

For the wavelet decomposition procedure the authors used the Daubechies (1992) compactly supported wavelet $d4$. Daubechies (1992) wavelet family has been frequently used in studies involving wavelet-based bootstrap methods, especially $d4$ has been chosen for reducing boundary effect of wavelets (PERCIVAL; SARDY; DAVISON, 2000; GOLIA, 2002; TANG; WOODWARD; SCHUCANY, 2008).

3.2 Estimating Hurst Exponent

The Hurst exponent statistic (HURST, 1951) was considered to evaluate if the analyzed methods provide reliable replicas, i. e., if the main characteristics of the simulated time series were preserved. As mentioned before, Hurst's exponent is considered a measure of the long-range dependence intensity, or the self-similarity parameter H. The H values belong to the interval (0,1), if H>0.5 the trend behavior of a given time series is called persistent, otherwise, this behavior is so-called anti-persistent.

The literature provides many methods for estimating Hurst's exponent as the Aggregate Variance method, Higuchi's method, the Periodogram method, the R/S method, among others (HURST, 1951; HIGUCHI, 1988; TAQQU; TEVEROVSKY; WILLINGER, 1995; BERAN et al., 2013). Each of these methods has advantages and disadvantages. In the latest years, several papers have been written comparing these methods and pointing that finding the optimal method is not easy (TAQQU; TEVEROVSKY; WILLINGER, 1995; CLEGG, 2006; VEDOVATTO, 2015).

In this paper, we are estimating Hurst's exponent using Higuchi's method (HIGUCHI, 1988). This method has been used as an efficient method to analyse non-stationary and irregular time series. This algorithm is similar to the method of aggregated variance and is available in free softwares, such as R. Furthermore, this method presents similar or better results in comparative studies (TAQQU; TEVEROVSKY; WILLINGER, 1995; KRAKOVSKÁ; KRAKOVSKÁ, 2016).

Given a time series Y, with a finite number of observations, $Y_1, Y_2, \ldots, Y_n$, Higuchi's method (HIGUCHI, 1988) is described as follows:

1. A new time series Y_k^m must be constructed following the expression:

$$Y_k^m = \{Y_m, Y_{m+k}, Y_{m+2k}, \ldots, Y_{m+[\frac{N-m}{k}]k}\} \tag{3.3}$$

   wherein $m = 1, 2, \ldots, k$, indicates the first observation time to be putted in the new time series, k is the period followed by the choices of the rest observations of Y_k^m, and $[\ ]$ denotes the greatest integer function.

2. The length of the curve associated to each time series Y_k^m must be found, as follows:

$$L_m(k) = \frac{1}{k}\left(\sum_{i=1}^{[\frac{N-m}{k}]} |Y_{m+ik} - Y_{m+(i-1)k}|\right) \frac{N-1}{[\frac{N-m}{k}]\,k} \tag{3.4}$$

   where the term $\frac{1}{k}\frac{N-1}{[\frac{N-m}{k}]k}$ represents a normalization factor.

3. Define $\langle L(k) \rangle$ as the average over the k sets of $L_m(K)$ associated to Equation 3.3. If the power law $\langle L(k) \rangle \propto k^{-D}$ is satisfied, then the curve is fractal with dimension D.

The Hurst exponent H is obtained from fractal dimension D by the following relation:

$$H = 2 - D \tag{3.5}$$

if Y is a self-similar process (MANDELBROT, 1982). The estimates of H were made using the R package *fArma*, which posses a Higuchi's method implementation.

3.3 Evaluation Criteria

To evaluate if Hurst's exponent is preserved after applying the three resample methods, we analyzed the Error (E) and the Mean Squared Error (MSE). Considering B bootstrap replicates of $\hat{\theta}$ ($\hat{\theta}_b$), the error estimate is given by:

$$\widehat{E}(\hat{\theta}) = \overline{\hat{\theta}^*} - \hat{\theta} \tag{3.6}$$

where $\overline{\hat{\theta}^*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b$, and $\hat{\theta}$ is the estimate computed from the original observed sample.

MSE is an important measure because it incorporates both the variance of the estimator as well as its bias. The MSE considering B bootstrap replicates of $\hat{\theta}$ can be estimated from the following procedure:

$$\widehat{MSE}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b - \hat{\theta})^2 \tag{3.7}$$

where $\hat{\theta}$ is, again, the estimate computed from the original observed sample.

3.4 Simulation

To illustrate the performance of the methods in question, we simulated time series with four characteristics, represented by $X_t{}^*$, $Y_t{}^*$, $Z_t{}^*$ and $W_t{}^*$, as follows

$$Y_t{}^* = 4Y_t + \frac{t}{100} + \cos\left(\frac{\pi(t-1)}{360}\right) \tag{3.8}$$

$$Z_t{}^* = 0.5Z_t{}^2 + 0.25Z_t + \frac{t^2}{25000} + \cos\left(\frac{\pi(t-1)}{180}\right) \tag{3.9}$$

$$W_t{}^* = 1.5W_t + \cos\left(\frac{\pi(t-1)}{180}\right) \tag{3.10}$$

wherein,

- Y_t was generated from a MA(1) with coefficient $\theta_1 = 0.8$, that is

$$Y_t = (1 + 0.8B)\epsilon_t;$$

- Z_t was generated from a SARIMA$(1,1,1)(1,1,1)_{12}$ with coefficient $\phi_1 = 0.5$, $\Phi_1 = 0.8$, $\theta_1 = 0.6$, and $\Theta_1 = 0.3$, that is

$$(1 - 0.5B)(1 - 0.8B^{12})(1 - B)(1 - B^{12})Z_t = (1 + 0.6B)(1 + 0.3B^{12})\epsilon_t;$$

- W_t was generated from a ARFIMA$(1, d, 0)$, wherein $\phi_1 = 0.95$ and $d = 2/5$, that is

$$(1 - B)^{2/5}(1 - 0.95B)W_t = \epsilon_t.$$

The time series ${Y_t}^*$ has a linear trend and periodicity, ${Z_t}^*$ has a nonlinear component, with polynomial trend and periodicity, and ${W_t}^*$ has a seasonal component as a systematic effect.

The time series ${X_t}^*$ is a Fractional Gaussian Noise generated from Beran (1994) algorithm with with $H = 0.96$. Time series generated from ${X_t}^*$ developed a special role because with this was possible to evaluate, indeed, the quality of Higuchi's method to estimate H.

For each of the four technical features aforementioned, time series of sizes 64, 128, 256, 512, 1024, and 2048 were generated, totaling 24 simulated time series. Each one of these time series was resampled 5000 times for each of the three analyzed methods. All the simulation procedure also was developed in the statistical language R.

Figure 3 illustrates time series of length 256 generated from ${Y_t}^*$, ${Z_t}^*$, ${W_t}^*$, ${X_t}^*$ at the graphics (a), (b), (c), and (d), respectively.

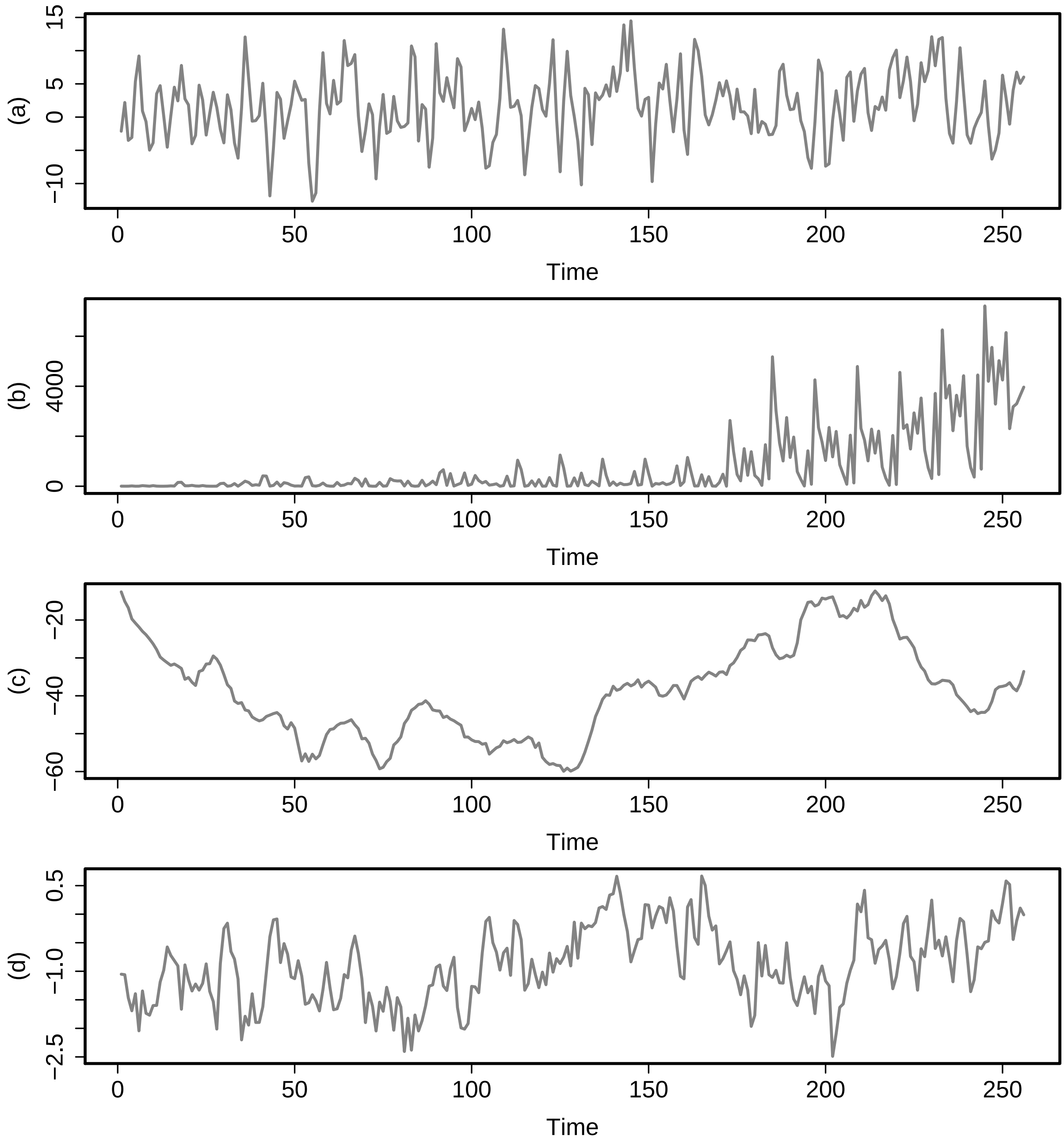

Figure 3 – Time series: (a) generated from ${Y_t}^*$, (b) generated from ${Z_t}^*$, (c) generated from ${W_t}^*$, (d) generated from ${X_t}^*$

From Figure 3 we can observe the behavior of the time series ${Y_t}^*$, although this time series was simulated for containing certain periodicity and trend, these behaviors are not so apparent as showed in this graph in (a). On the other hand, a persistent behavior existent in the time series constructed from ${W_t}^*$, ${X_t}^*$, graphics (c) and (d), respectively. Also it is remarkable the non-linear trend and periodicity in the time series on the graphic (b). Thus, several behaviors found in the analysis of time series are contemplated in these simulated time series.

In resampling process was developed in the first level of decomposition for both detail and wavelet coefficients, using the Daubechies wavelet (DAUBECHIES, 1992), with two vanishing moments (d4). Daubechies wavelet $(d4)$ and $(d8)$ has usually been used in similar works (GOLIA, 2002; TANG; WOODWARD; SCHUCANY, 2008).

3.5 Results

Table 4 describes the estimated value of H for each time series generated from ${X_t}^*$.

Table 4 – Hurst exponent estimated from Higuchi's method

Time Series Length	64	128	256	512	1024	2048
Hurst exponent	0.970	0.965	0.965	0.963	0.962	0.960

Clearly, the estimates of H are very close to the real value, and as the size of the time series increases, such values become even closer. Thus, we can notice that the Higuchi's method provides reliable estimates for the true measure H used to generate such time series.

The bias of TSWDWT is described in Figure 4, where we observed small negatives and positives error values. In general, for TSWDWT, the magnitude of the bias decays to zero as the size of the time series increases. Especially, for the time series obtained from ${Z_t}^*$, which contains a nonlinear component, the error presented major magnitude.

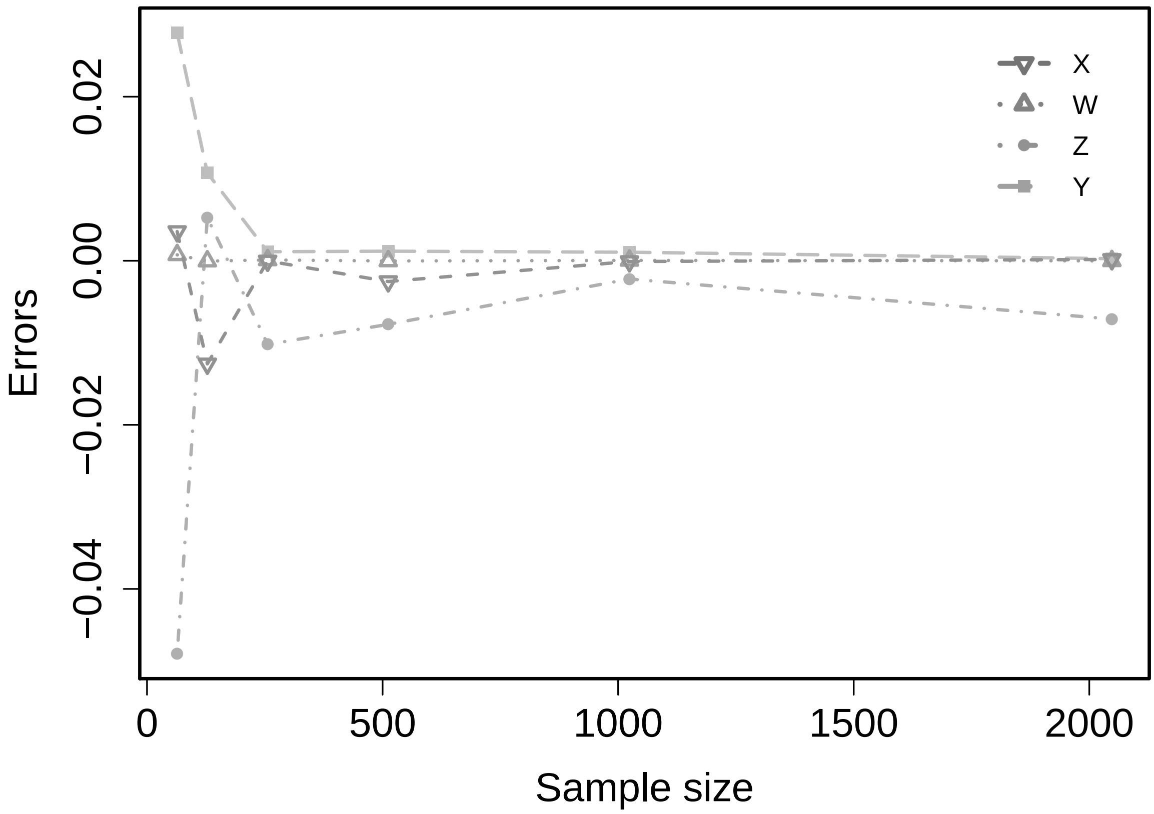

Figure 4 – Replicated time series errors by the TSWDWT method.

Figure 5 presents the error obtained from TSWNDWT method. As in TSWDWT, positives and negatives error values were observed. However, the magnitudes of these bias values also decay rapidly to zero for time series with a larger size.

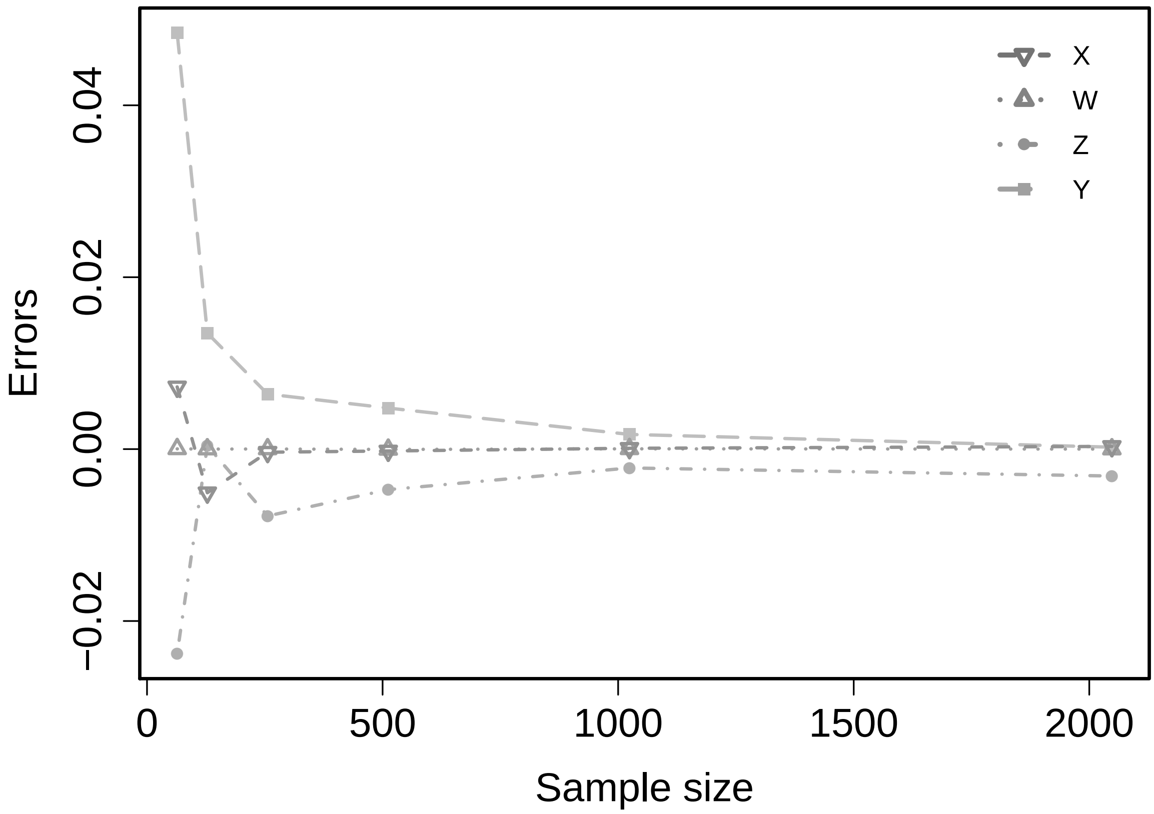

Figure 5 – Replicated time series errors by the TSWNDWT method.

The naive bootstrap based on NDWT was also evaluated with respect to error as represented in Figure 6. As in the two others approaches, small values of error were observed decreasing for zero.

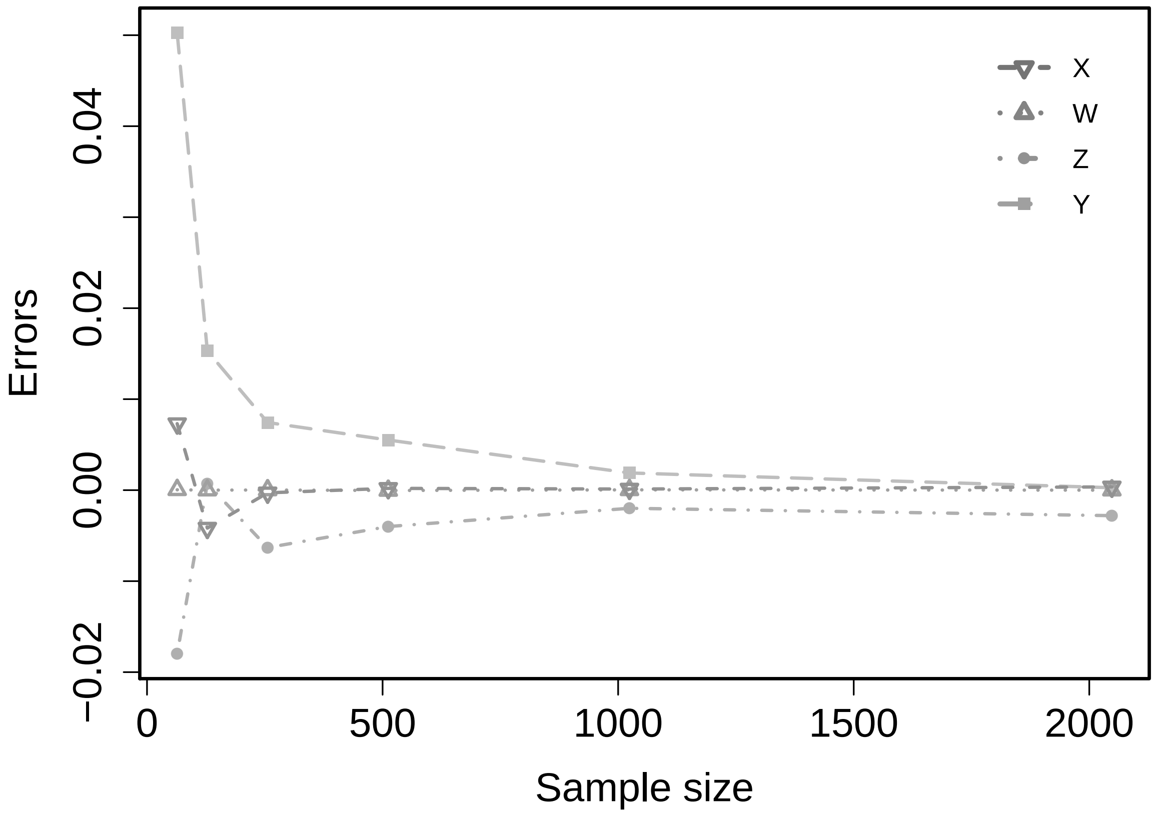

Figure 6 – Replicated time series errors by the naive bootstrap based on NDWT method.

From the error analyzes, the H estimates seem to be preserved after the resampling methods. The MSE of the methods TSWDWT, TSWNDWT, and naive bootstrap based on NDWT are represented in the Figure 7, 8, 9, respectively.

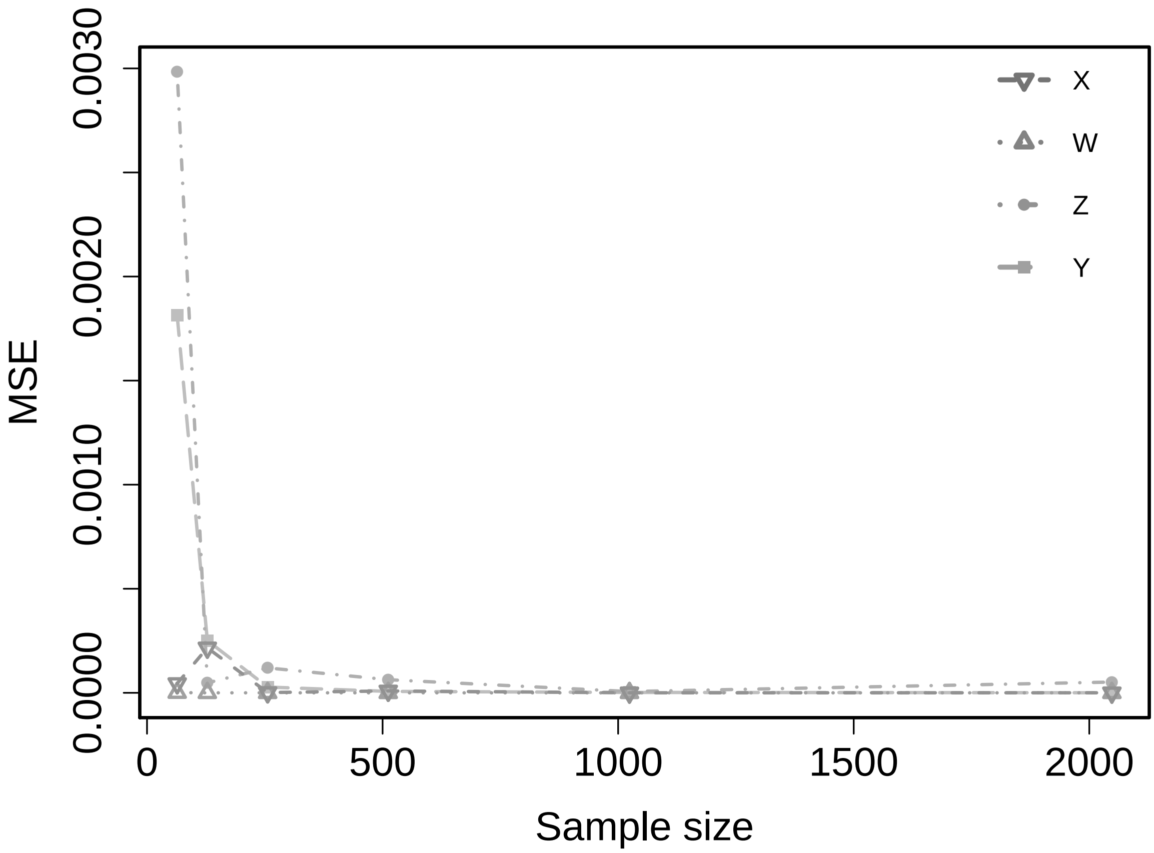

Figure 7 – MSE obtained from TSWDWT method.

From Figure 7 we notice that TSWDWT presents less precision for small time series, which rapidly decreases while increasing the time series length. The time series generated from ${Y_t}^*$ and ${Z_t}^*$ are a little more disturbed by the resampling technique for small lengths as 64.

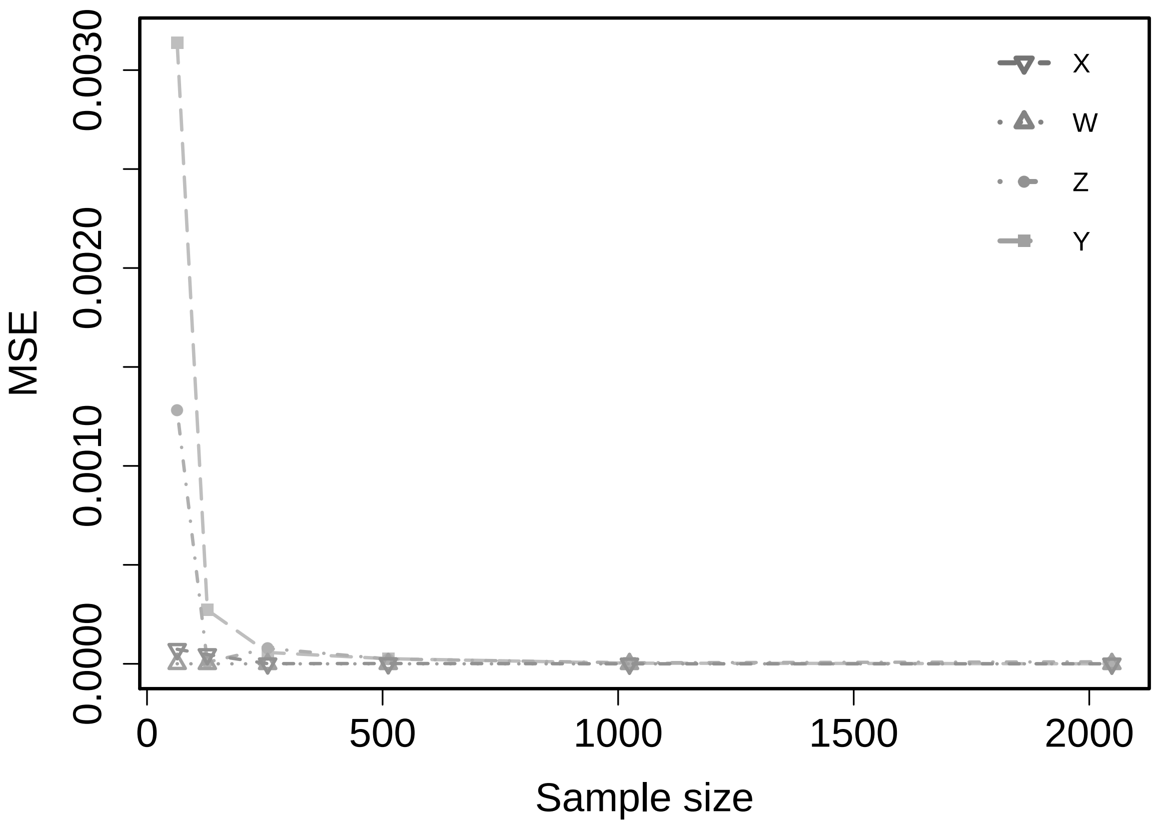

Figure 8 – MSE obtained from TSWNDWT method.

TSWNDWT has similar behavior to TSWDWT, as the size of the time series increases the estimates presents more precision. The time series generated from Y_t^* presents a little imprecision for small time series, but, its MSE also tends to zero when this time series length is increased.

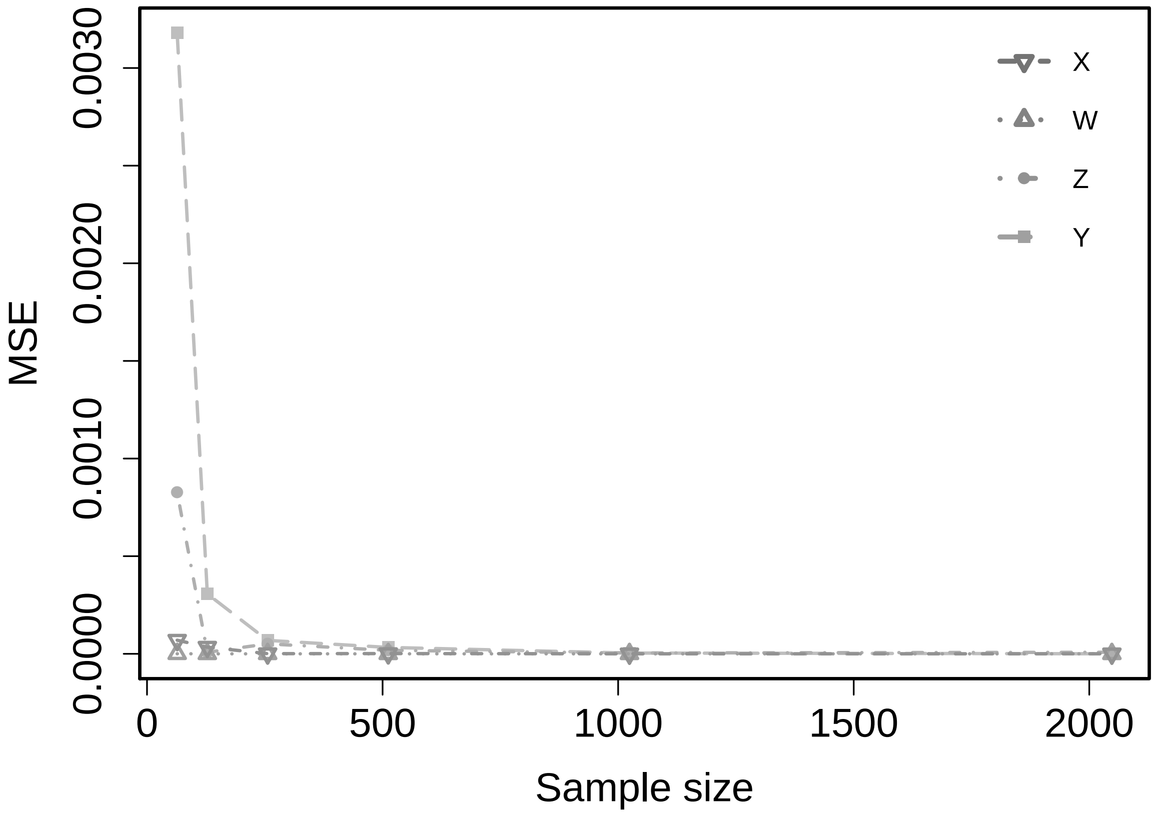

Figure 9 – MSE from naive bootstrap based on NDWT method.

From Figure 9 we notice that for the generated time series, the precision and accuracy has desirable behavior. In other words, for small time series length, MSE value is close to zero, and it tends to zero as the size of the time series increases. Less precision is observed for small time series generated by Y_t^*.

From errors and MSE analyzes we observe that the H value before and after resampling are very close. So that the resampling based on wavelets does not seems to disturb the main behavior of the time series.

3.6 Final considerations

For evaluating if wavelet-based resampling methods preserve the Hurst exponent, one method based on discrete wavelet transform and two other methods based on non-decimated wavelet transform were investigated in this paper. The analysis of the H statistics after resampling was evaluated for bootstrapping simulations applied in 24 time series containing different lengths and characteristics.

Each evaluated method had a similar behavior with respect to errors and MSE. The results contained in this paper show that the influence of series length is important in all the evaluated methods; smaller errors and MSE were observed for larger time series.

We observed that TSWNDWT, TSWDWT, and the naive bootstrap based on NDWT preserved the main correlation features of the simulated times series. Wavelet-based bootstrapping become possible to resample correlated data, thus, trajectories of a stochastic process can be replicated from an observed time series, overcoming the various limitations on truly observing them.

Higuchi's method showed up a consistent methodology to estimate the H value for the time series generated from X_t^*. At the moment, we are comparing the main technique for estimating Hurst Exponent in Fractional Gaussian Noise, as well as in Fractional Brownian Motion. The existence of methods to generate such time series from a chosen H value makes it possible to evaluate the quality of each estimation method.

Acknowledgment: The authors acknowledge the financial support of the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES).

Chapter 4

Estimating the Confidence Interval of Evolutionary Stochastic Process Mean from Wavelet based Bootstrapping

Abstract

A time series is a realization of a stochastic process, where each observation is considered in general as the mean of a Gaussian distribution for each time point t. The classical theory is built based on this supposition. However, this assumption may be frequently broken, mainly for non-stationary or evolutionary stochastic process. Thus, in this work we proposed to estimate the uncertainty for the evolutionary mean, μ_t, of a stochastic process based on bootstrapping of wavelet coefficients. The wavelet multiscale decomposition provides wavelet coefficients that have less autocorrelation than the observations in time domain, allowing to apply bootstrap methodologies. Several bootstrap methodologies based on discrete wavelet transform (DWT), also called wavestrapping, have been proposed in the literature to estimate the confidence interval of some statistics for a time series, such as, e.g., the autocorrelation. In this paper we implemented these methods with few modifications and compared them to newly proposed methods based on non-decimated wavelet transform (NDWT), which is a translation invariant transform and more adequate for dealing with time series. Each realization of the bootstrap provides a surrogate time series, that imitates the trajectories of the original stochastic process, allowing to build a confidence interval for its mean for both stationary and non-stationary processes. As an application, the confidence interval of the mean rate of bronchiolitis hospitalizations for Paraná-BR state were estimated as well as its bias and standard errors.

4.1 Introduction

Time series data are naturally found in a range of fields such as Agriculture, Geophysics, Meteorology, Health, Economy and Social Sciences, among several others (CHATFIELD, 2013; WEI, 1994). Given a parametric space T and a probability space (Ω, A, P), a stochastic process is a family $Z = \{Z(t), t \in T\}$, such that, for each $t \in T$, $Z(t)$ is a random variable (MORETTIN; TOLOI, 2006). A time series is considered as the finite realization of a stochastic process. In others words, an observed time series is a trajectory of a stochastic process.

Indeed, $Z(t)$ is a two variable function $Z(t, w)$ wherein $t \in T$, and $w \in \Omega$. Considering $f_Z(z)$ is the probability density function of $Z(t, w)$, the Figure 10 represents a stochastic process as aforementioned.

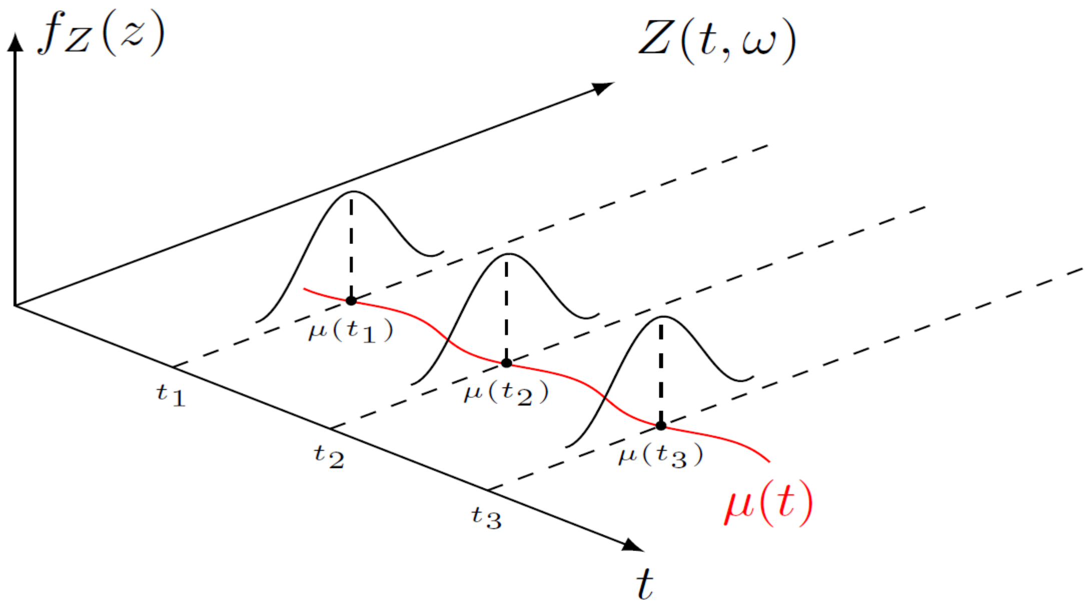

Figure 10 – A stochastic process represented as a family of random variables

In many situations, accessing more than one observation of a phenomenon for each instant of time is impossible. In general, the function $Z(t, w)$ is assumed to follow a Gaussian distribution for each instant, and the observed time series represents the mean μ_t of the stochastic process for each $t \in T$. But, this assumption not always is true and $f_Z(z)$ can be different at each instant of time. It would be very important to estimate the uncertainty associated to μ_t, from its confidence interval.

The technique called bootstrap (EFRON; GONG, 1983), which is an appropriate methodology for solving a variety of inferential problems, could be a good alternative to estimate the uncertainty for μ_t. However, bootstrap is more designed for uncorrelated data, and not for those exhibiting short or long-range dependence as time series. Fortunately, in the last years, several methods were developed to deal with resampling time series, some of them based on wavelets, especially using DWT (PERCIVAL; SARDY; DAVISON, 2000; GOLIA, 2002; ANGELINI et al., 2005; YI et al., 2007).

To estimate the confidence interval for μ_t, we aim to evaluate and implement some methods from the literature but with some modifications, and propose others involving NDWT:

M1: Naive bootstrapping based on NDWT;

M2: DWT two step wavestrapping (TSWDWT);

M3: NDWT two step wavestrapping (TSWNDWT);

M4: Reinflation of the bootstrap resamples of TSWDWT.

M5: Reinflation of the bootstrap resamples of TSWNDWT.

The methods M1, M2, M3, M4, and M5, are going to be applied and compared for estimation of the uncertainty for bronchiolitis hospitalization rate in Paraná State from 2000 to 2014. Bias, standard errors, and coefficients of variation can evaluate the ensembles of resampled time series.

This work is organized as follows. Section 2 presents a brief review of the techniques usually used to resample time series. In Section 3, we describe the methods we are using to estimate the uncertainty associated with the bronchiolitis hospitalization rate. In sections 4 and 5, the main results and conclusions of our study are presented, respectively.

4.2 Resampling time series

One of the most important characteristics of a time series is the dependence on nearby observations. Because of this correlation structure, maintaining the data order is of great importance. So, resampling time series requires appropriate techniques that consider the dependence and the order of the observations. One of the usual approach is the Stationary Bootstrap (SB) (POLITIS; ROMANO, 1994).

Considering Y_t is a strictly stationary and weakly dependent time series, the SB is a special case of blocks resampling, which consists in defining two sequences of random variables $L_1, L_2, \ldots$ and $I_1, I_2, \ldots$, both independent of each other and independent of Y_t, and such that $L_1, L_2, \ldots$ follow a geometric distribution with parameter p and $I_1, I_2, \ldots$ follow an uniform distribution on $\{1, 2, \ldots, n\}$. Then, the random blocks B_{I_i,L_i}, with random blocks length L_{i-1}, are given by

$$B_{I_i,L_i} = (Y_{I_i}, Y_{I_i+1}, \ldots, Y_{I_i+L_{i-1}}). \tag{4.1}$$

However, SB is not applicable to those time series exhibiting non-stationary and long-range dependence.

In the last years, the wavelet analysis has been standing out as a tool for resampling time series (PERCIVAL; SARDY; DAVISON, 2000), (ANGELINI et al., 2005), (BREAKSPEAR; BRAMMER; ROBINSON, 2003), (GOLIA, 2002). Basically, considering we have a multiresolution analyses (MRA) (MALLAT, 1989), a time series $Y = (y_0, y_1, \cdots, y_{n-1})$ can be represented as a function f in terms of the scaling function ϕ and wavelet function ψ as

$$f(t) = \sum_{k=0}^{n-1} c_{J_0,k}\phi_{J_0,k}(t) + \sum_{j=J_0}^{J-1}\sum_{k=0}^{n-1} d_{j,k}\psi_{j,k} \tag{4.2}$$

where $J-1 < \log_2 n \leq J$, $j = J_0, \cdots, J-1$ representing a multiresolution level, and $k = 0, \cdots, n-1$. The coefficients $c_{J_0,k}$ and $d_{j,k}$ are called the smooth (scaling) and detail (wavelet) coefficients, respectively (KANG; VIDAKOVIC, 2017).

When we take $\phi_{J_0,k}(t) = 2^{J_0/2}\phi(2^{J_0}t - k)$ and $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$, the coefficients $c_{J_0,k}$ and $d_{j,k}$ comprise the DWT of the time series Y. On the other hand, taking $\phi_{J_0,k}(t) = 2^{J_0/2}\phi(2^{J_0}(t-k))$ and $\psi_{j,k}(t) = 2^{j/2}\psi(2^j(t-k))$, the detail and smooth coefficients represent the NDWT of the time series Y.

The DWT wavelet coefficients have less autocorrelation than the observed time series, and this allows applying bootstrap (wavestrap), even for non-stationary time series (GOLIA, 2002; YI et al., 2007; TANG; WOODWARD; SCHUCANY, 2008). However, in conditions where translation or shift-invariance (NASON; SILVERMAN, 1995) is important, as for time series, the NDWT is a good alternative.

One difficulty in applying bootstrap on DWT is the number of wavelet coefficients which becomes smaller at each resolution level. NDWT has the same number of wavelet coefficients in each resolution level, overcoming this DWT limitation. NDWT is also more flexible with respect to the time series length, being appropriate for all those which are a multiple of two. Furthermore, NDWT has an easy implementation with more than one algorithm, including the pyramidal algorithm (MALLAT, 1989).

4.2.1 Bootstrap based on wavelets

Golia (2002) applied the stationary bootstrap to the wavelet coefficients of time series exhibiting long memory (GOLIA, 2002). This application was possible because the wavelet coefficients are wide-sense stationary and weakly correlated in each scale (WORNELL; OPPENHEIM, 1996). In her work, she used the Daubechies wavelet with 4 vanishing moments and coarsest level of details equals to 4 in DWT. The results were good, however, the author comments the need of evaluating this approach for other long memory processes. The procedure of this wavelet based stationary bootstrap is described in Figure 11 .

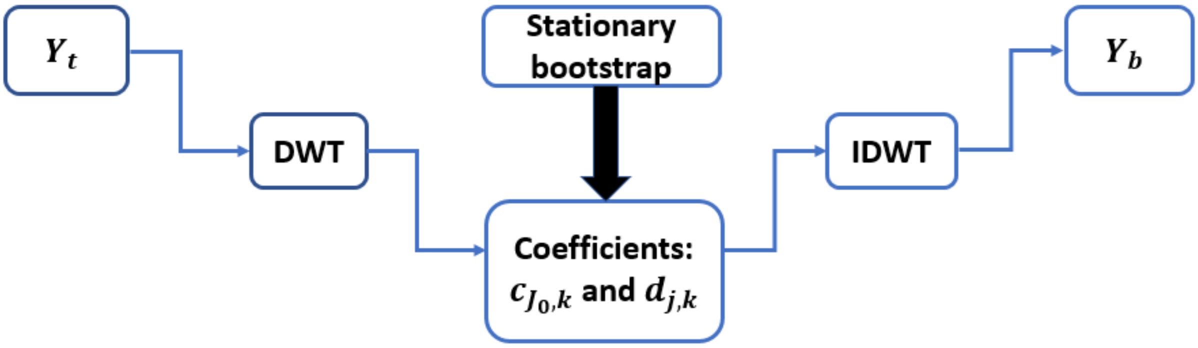

Figure 11 – Wavelet based stationary bootstrap

Considering the SB and its use combined with wavelets, Yi et al. (2007) developed a DWT-based method called of Two-Step Wavestrapping (TSW), to simulate non-stationary acceleration data in the mobile computing context (YI et al., 2007). In this context, they intended to simulate the acceleration data collected of a group composed by one hundred twenty six undergraduate students. In this sort of data, each student provides one time series for each one of the three evaluated axis forming a group of time series. Each group of time series was divided into subgroups statistically characterized by Hurst exponents, and then TSW procedure is applied by subgroups.

Describing TSW for only one time series, the first part of the TSW consists on performing the SB in one-step of DWT, which is called of Stationary Parallel Bootstrapping. In other words,

1. Given a time series Y of power-of-two length, apply the DWT to generate the coarsest level of detail (J_0) and scale coefficients;
2. Resample these scaling and wavelet coefficients using the Stationary Bootstrap (POLITIS; ROMANO, 1994);
3. Apply the inverse discrete wavelet transform (IDWT) to the resampled wavelet coefficients to generate a surrogate time series Y_b, wherein b indicates the performed bootstrap.

The second step consists in adjusting the trend and energy. For trend adjustment, both the time series Y and its surrogate Y_b are decomposed using the DWT. Then, the scaling coefficients c_b obtained from Y_b are surrogated by the scaling coefficients c generated from Y. For the energy adjustment, the following steps can be performed:

i. Generate the average energy for each decomposition level of Y and Y_b, given by

$$\bar{e}_j = \sum_{k=0}^{2^j-1} \frac{{d_{j,k}}^2}{2^j}, \quad j = J_0, \cdots, J-1, \tag{4.3}$$

where $d_{j,k}$ is the kth wavelet coefficient in the jth decomposition level.

ii. Adjust the average energy for each decomposition level of Y_b to the average energy of the levels of Y, doing

$$d_{baj,k} = d_{bj,k}\sqrt{\frac{\bar{e}_j}{\bar{e}_{bj}}}, \tag{4.4}$$

where d_{ba} represents the adjustment realized in each decomposition level of Y_b;

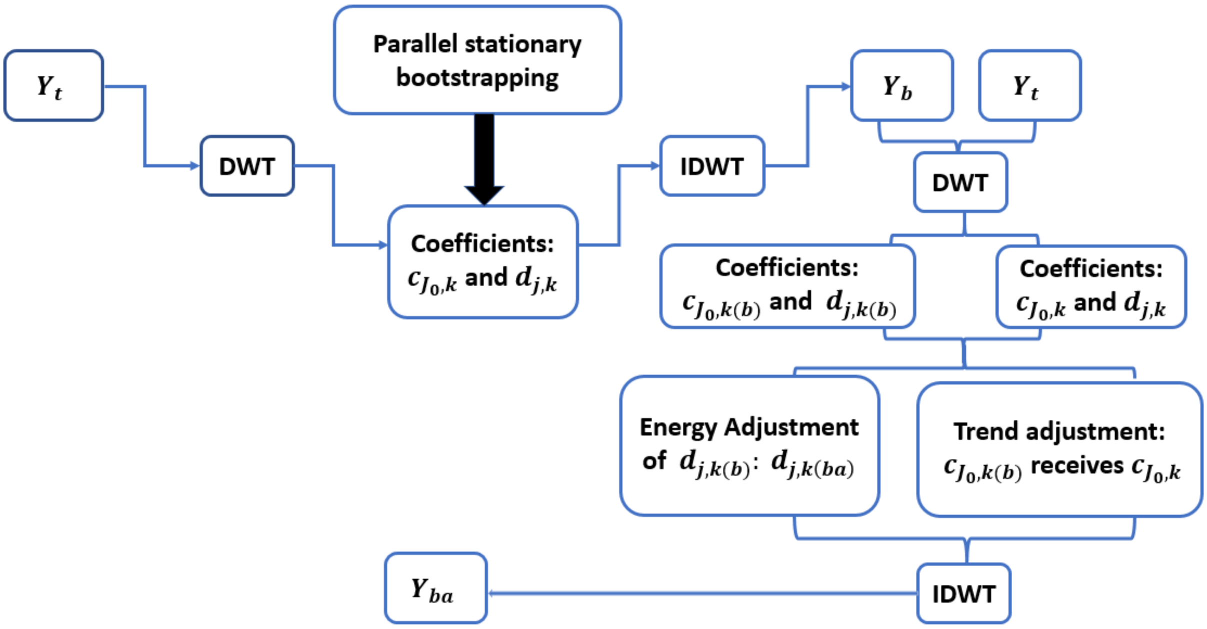

Figure 12 – Two step wavestrap (TSW) algorithm

Figure 12 summarizes the TSW algorithm. An important contribution of this methodology is the idea of an energy adjustment in the levels to preserve the inherent variability of the original data, even after the resampling. Furthermore, each realization of this procedure provides a surrogate time series with the same feature of the original time series. Another important point, is that the vertical correlation of wavelet coefficients among scale levels was taken into account, since the scaling and wavelet coefficients were resampled together.

In the next section, we present the proposed bootstrap methods, that are based on NDWT, SB, and TSW of the wavelet coefficients.

4.3 Proposed Methods

Using the statistical language R (R Core Team, 2016), we implemented five methods to generate the proposed confidence interval for μ_t.

The first bootstrap technique consists in performing a decomposition of the time series using NDWT, applying the naive bootstrap to detail coefficients and then generating a surrogate time series using INDWT. Figure 13 presents the NDWT naive algorithm.

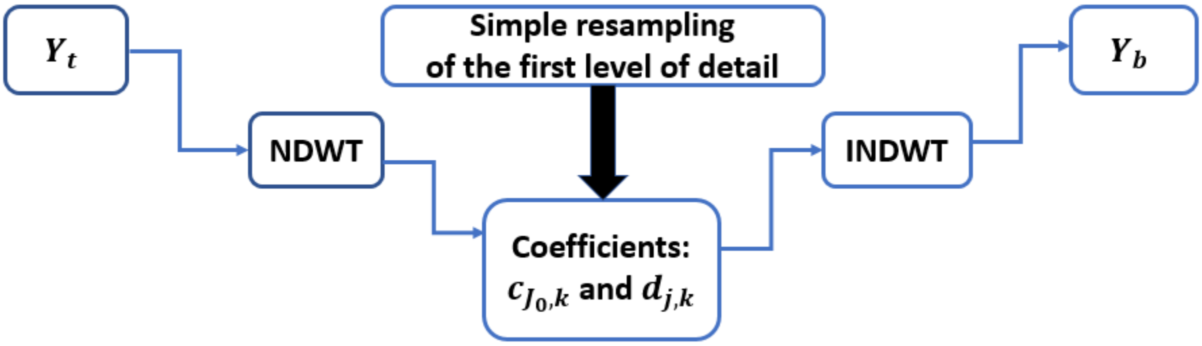

Figure 13 – Naive bootstrap based on NDWT

The next approaches were implemented following the same steps as TSW. The first one, called of TSWNDWT, follows the same steps of TSW but replacing the DWT by NDWT. In the second step, we work only with NDWT coefficients that comprise the first level of details. As in TSW we also developed the trend and energy adjustment as described in subsection 4.2.1. Figure 14 describes the TSWNDWT algorithm.

The method called of TSWDWT follows the same steps of TSWNDWT, but the decomposition and reconstruction of the time series is performed using DWT. The latest approaches consists in reinflating the surrogate time series obtained from TSWDWT and TSWNDWT. In the literature, reinflation means multiplying the correlation factor correction $\sqrt{1.1}$ to the surrogate time series (TANG; WOODWARD; SCHUCANY, 2008).

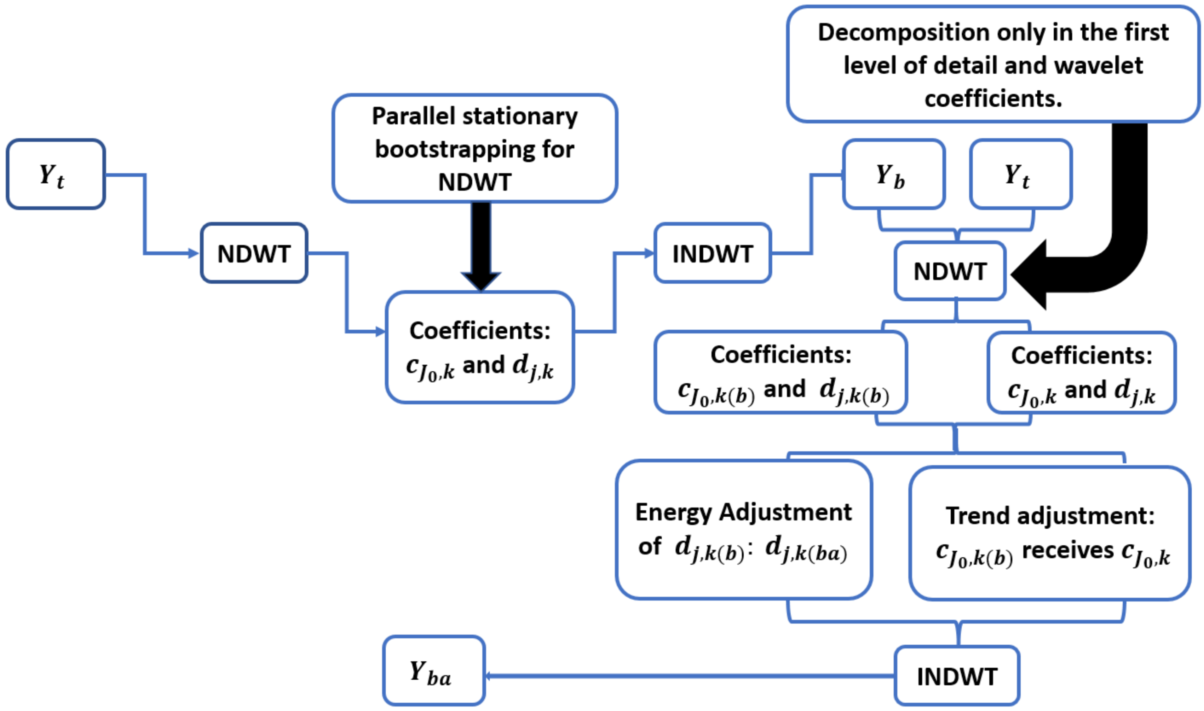

Figure 14 – TSWNDWT algorithm

To illustrate the confidence interval of the evolutionary stochastic process mean we used the month rate of bronchiolitis hospitalizations time series from the Paraná State - BR, in the period from 2000 to 2014. This time series was collected from DATASUS database and contains 180 observations.

The resampling methods based on DWT require data of power-of-two size. So, we extend the time series by reflection to 256 observations.

For each one of the proposed method, we fixed the orthonormal Daubechies' wavelet (DAUBECHIES, 1992), with 2 vanishing moments $(d4)$. This family of wavelets has been frequently used in similar works (GOLIA, 2002; TANG; WOODWARD; SCHUCANY, 2008). Furthermore, to obtain the mean of the sthocastic process μ_t, the level mean of the time series, standard errors and bias we resampled the time series 5000 times for each one of the proposed methods.

4.4 Results

Figure 15 presents the time series of the rate of monthly hospitalizations for bronchiolitis (Y), and the mean of the group of surrogate time series for Y from each presented bootstrap method.

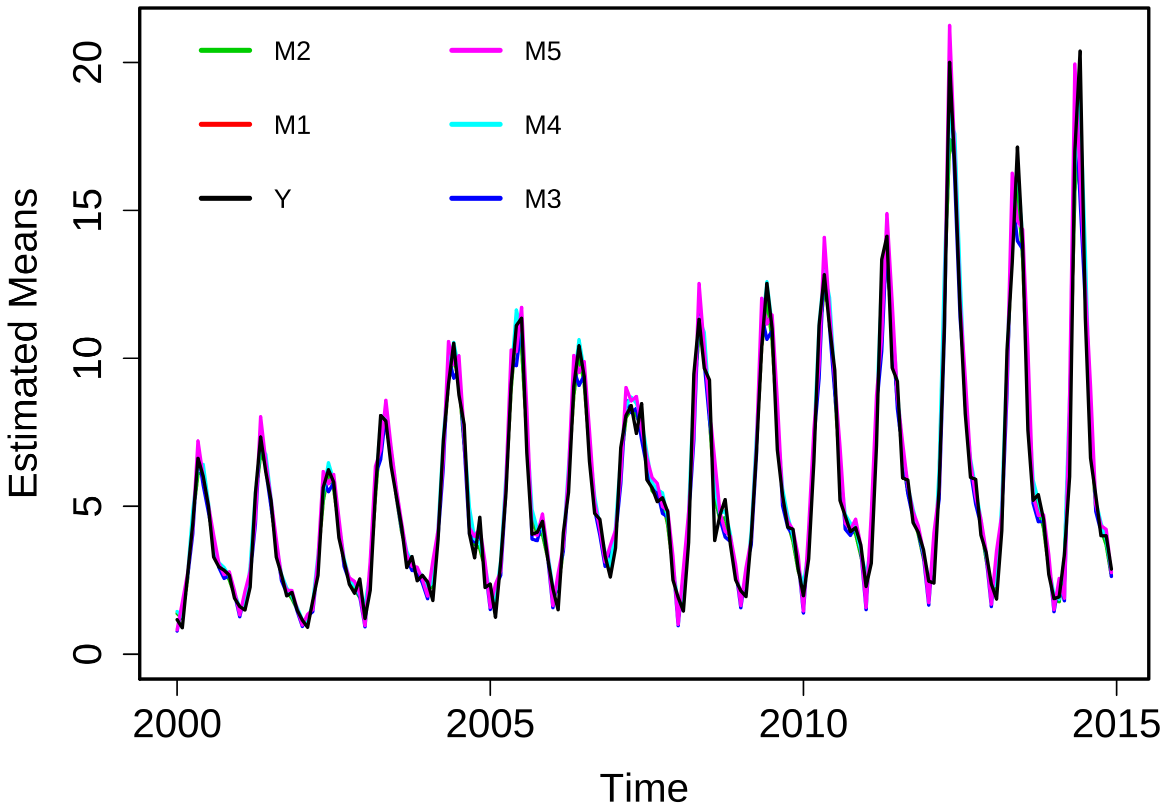

Figure 15 – Averages of the surrogate time series: Y - rate of bronchiolitis time series, M1 - naive bootstrap based on NDWT, M2 - TSWNDWT, M3 - TSWDWT, M4 - reinflated TSWNDWT and M5 - reinflated TSWDWT.

From Figure 15 all the bootstrap means seems to be similar to the observed time series. We can also generate the standard errors of the surrogate time series, as presented in Figures 16 and 17.

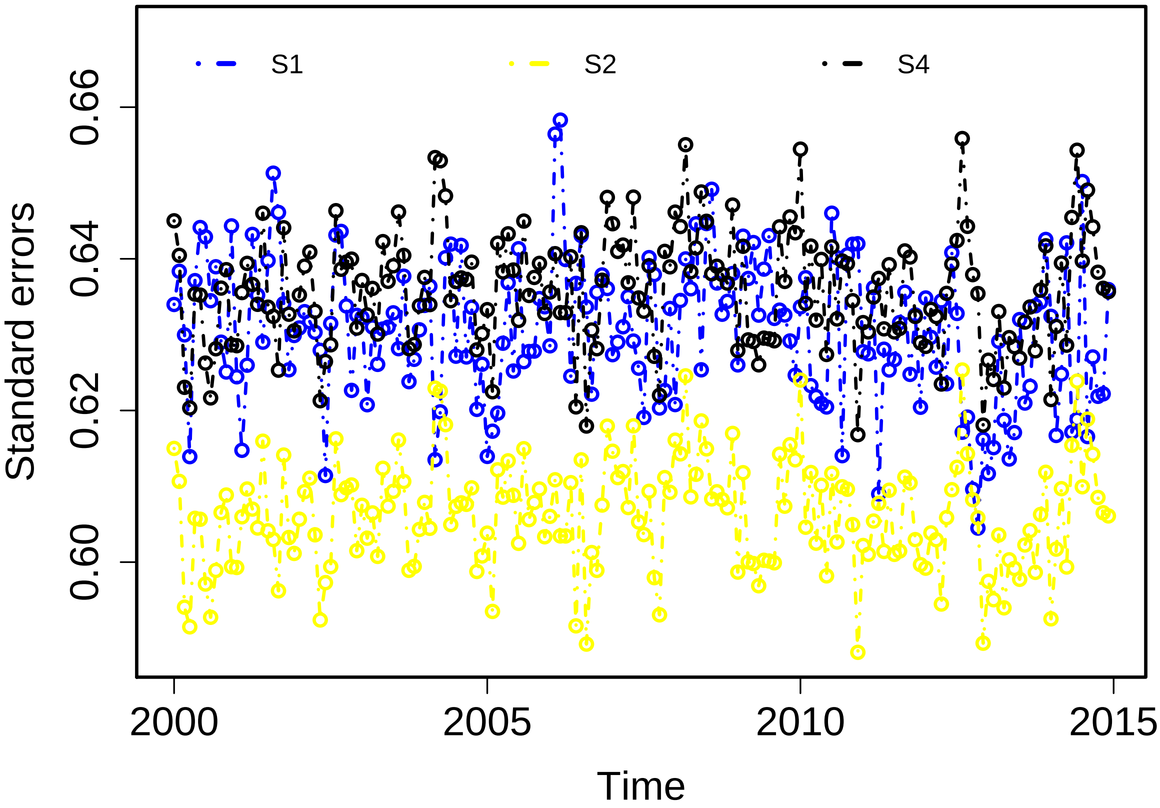

Figure 16 – Standard errors of surrogate time series: M1 - naive bootstrap based on NDWT, M2 - TSWNDWT, and M4 - reinflated TSWNDWT.

We can see that the naive bootstrap based on wavelet, TSWNDWT, and reinflated TSWNDWT methods presented low variability, whereas TSWDWT, and reinflated TSWDWT standard errors present a high level of oscillation. Possibly this behavior is related to the number of coefficients in each decomposition level. While the number of coefficients in each level of NDWT remains the same as the observed time series, in DWT, the number of coefficients decreases by half in each level. In general, the TSWNDWT presented the best standard errors and coefficient of variation.

The averages of the standard errors, coefficient of variation, and bias are represented in Table 5. The results corroborate with the graphical analyzes, pointing the TSWNDWT as the method with the smallest variability.

Table 5 – Average of standard errors (SE), coefficient of variation (CV) and bias of the surrogate time series

Classes	NDWT bootstrap	TSWNDWT	TSWDWT	Reinflated TSWNDWT	Reinflated TSWDWT
SE	0.63	0.61	0.90	0.64	0.95
CV	11.23	10.79	16.05	10.79	16.05
Bias	2.74×10^{-5}	-7.72×10^{-6}	6.15×10^{-5}	-2.74×10^{-1}	-2.74×10^{-1}

The results in Table 5 also corroborates with the Figure 17 indicating the largest variability for the methods that use DWT.

We also can see that the naive bootstrap based on wavelet, TSWNDWT, and reinflated TSWNDWT average bias are smaller than those for TSWDWT and reinflated TSWDWT methods. The TSWNDWT presented the best average of bias, which is about -0.000008.

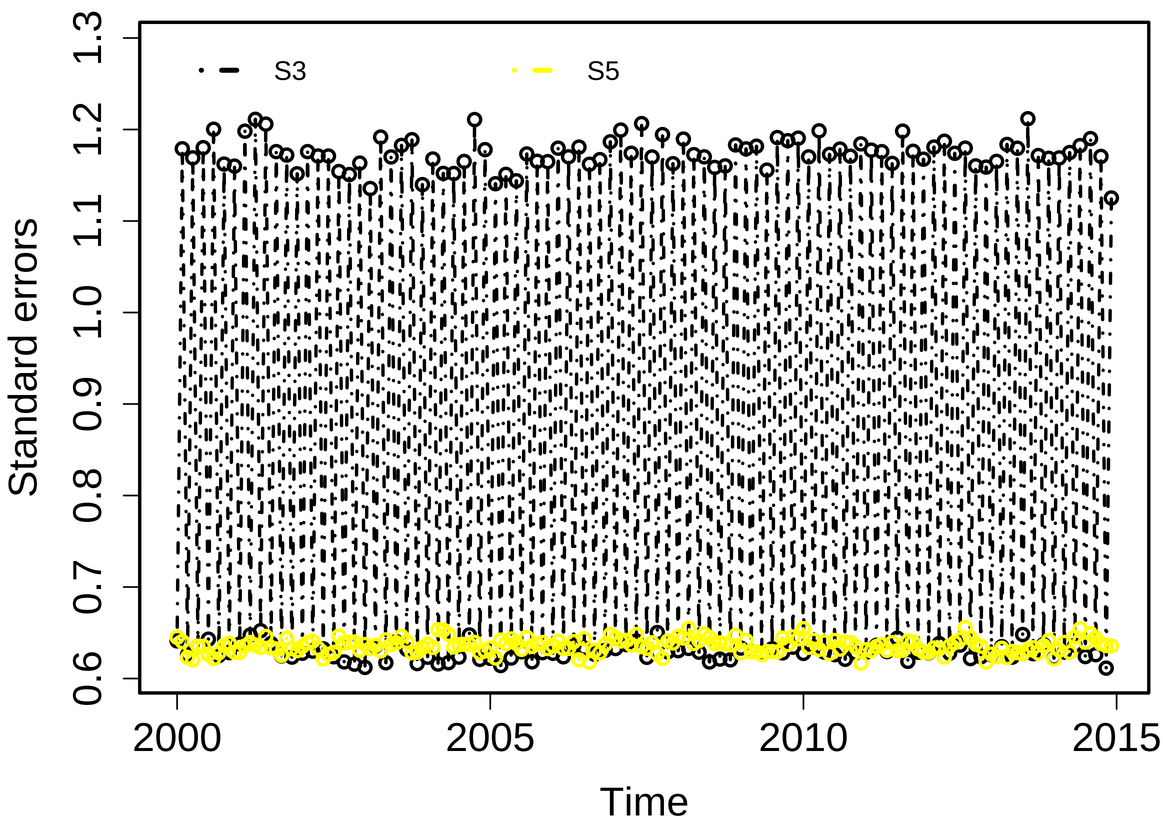

Figure 17 – Standard errors of surrogate time series: M3 - TSWDWT, and M5 - reinflated TSWDWT.

Figures 18 and 19 present the bias of the mean of the surrogate time series for each one of the evaluated methods. As in the standard errors analyze, the naive bootstrap based on wavelet, TSWNDWT, and reinflated TSWNDWT methods presented best results. On the other hand, TSWDWT and reinflated TSWDWT bias reached largest values.

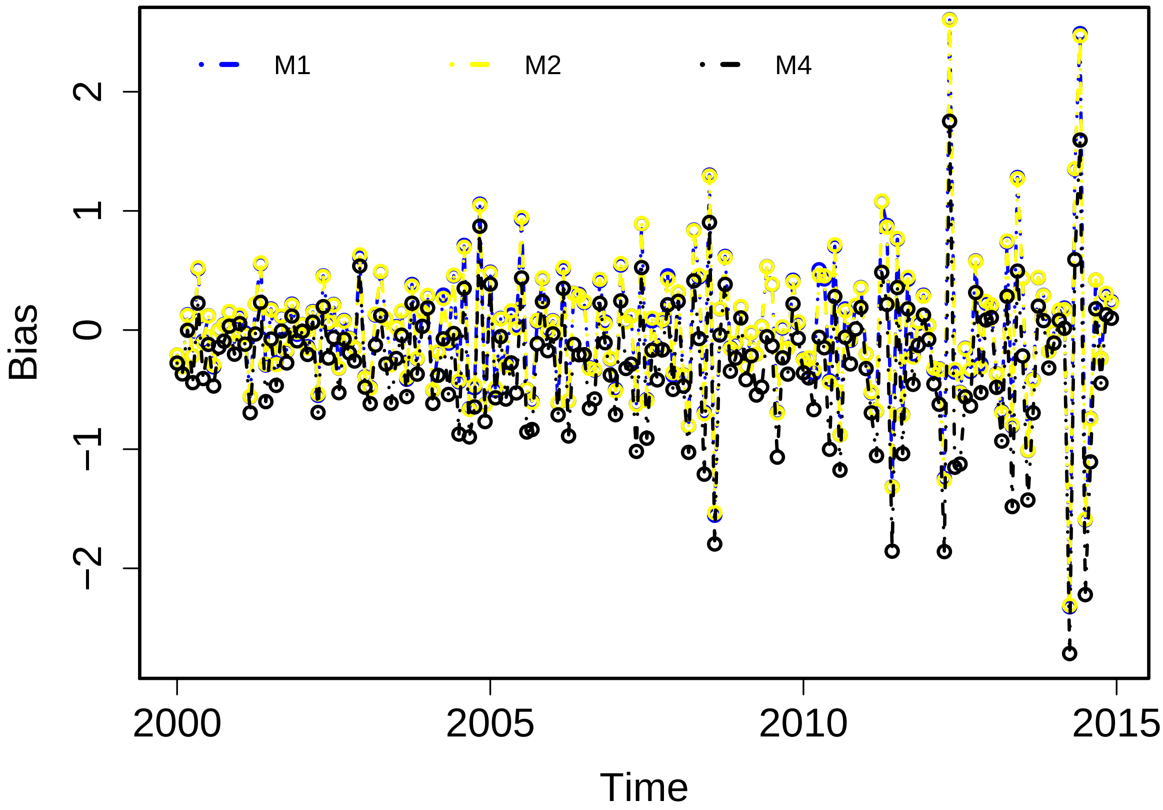

Figure 18 – Bias: M1 - naive bootstrap based on NDWT, M2 - TSWNDWT, and M4 - reinflated TSWNDWT.

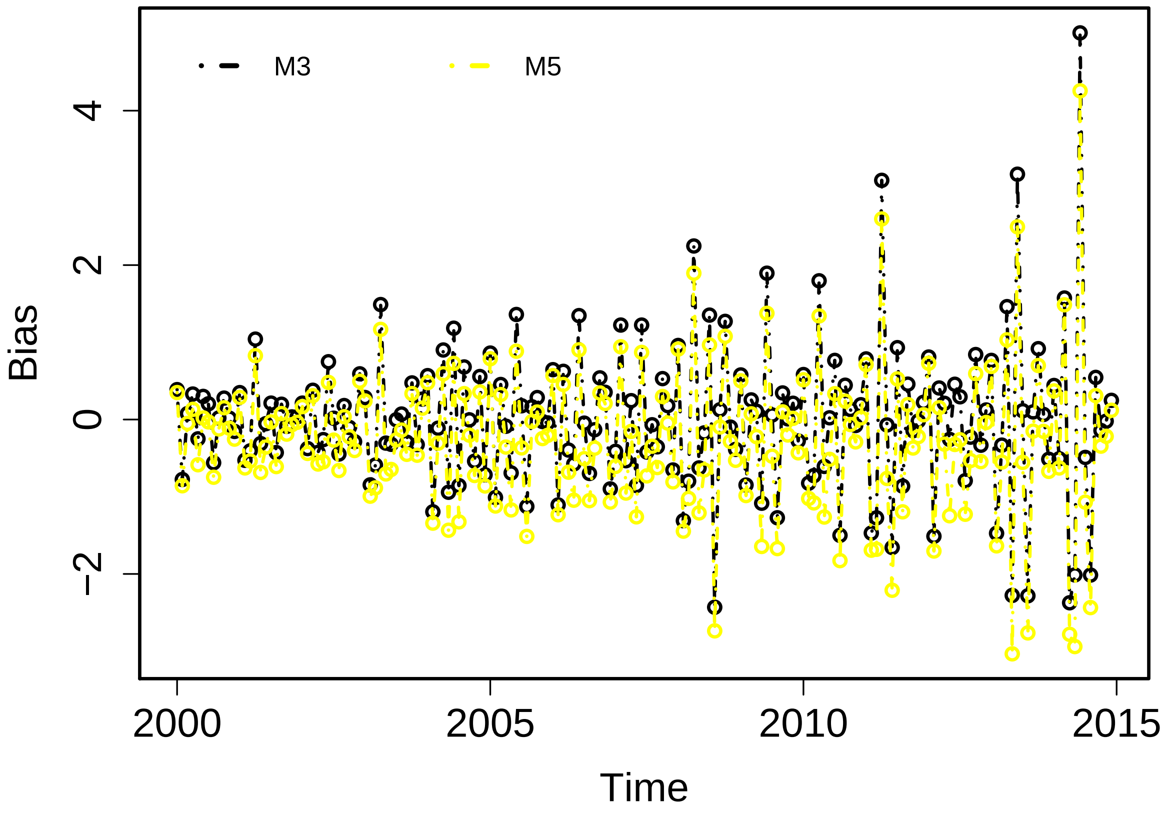

Figure 19 – Bias: M3 - TSWDWT, and M5 - reinflated TSWDWT.

In the Figures 20, 21, 22, 23, and 24 the confidence interval for the rate bronchiolitis hospitalizations time series obtained from each discussed method are presented. In all graphs, the time series Y is represented as a black dotted line and the confidence interval are a red line.

Figure 20 presents the confidence interval generated from the naive bootstrap based on NDWT. The CI constructed from this method included almost all the values of the observed time series, which represents the mean of the stochastic process.

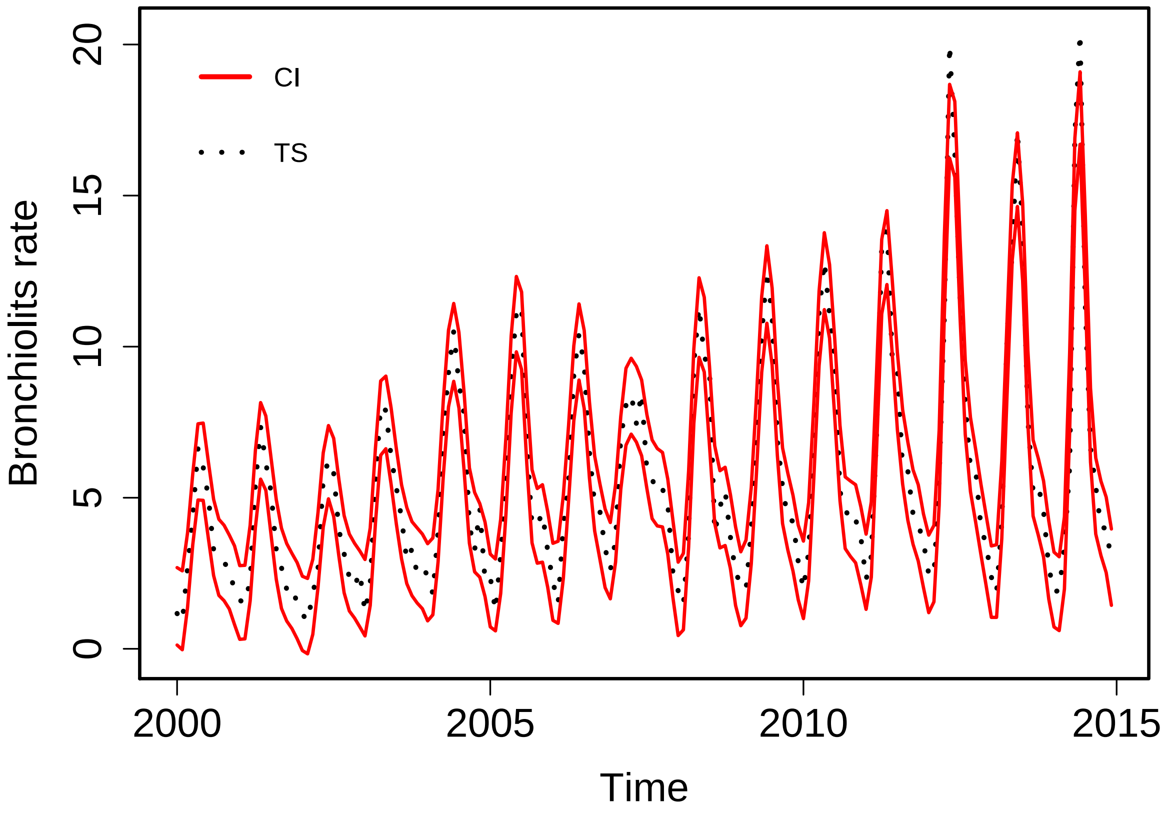

Figure 20 – Confidence interval (CI) obtained from naive bootstrap for the rate of bronchiolitis hospitalizations time series (TS)

From Figure 21 one can observe the confidence interval generated from TSWNDWT. This method also includes almost all the values of that observed time series, but we can observe that this interval is little more narrow than the in the confidence interval using only naive NDWT bootstrap.

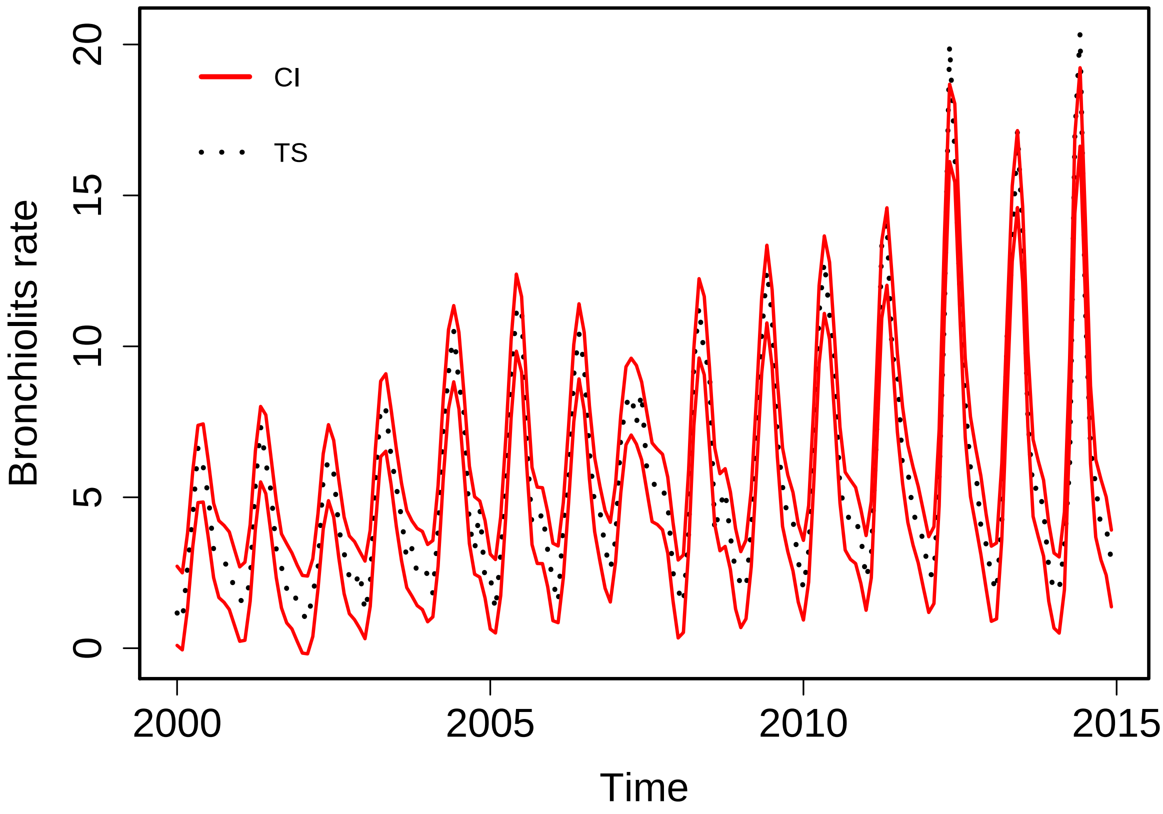

Figure 21 – Confidence interval obtained from TSWNDWT

Figure 22 presents the confidence interval generated from TSWDWT. The CI obtained from this method contains the most part of the observed values, and it seems to have less point out of the interval than the two methods already analyzed. However, each one of the time series that compose the confidence interval has more noise than those obtained from naive NDWT bootstrap and TSWNDWT.

The presence of more noise in the confidence interval generated from TSWDWT is expected since this method based on DWT seems to present more variability, bias, besides fewer wavelet coefficients in each multiresolution level to be resampled.

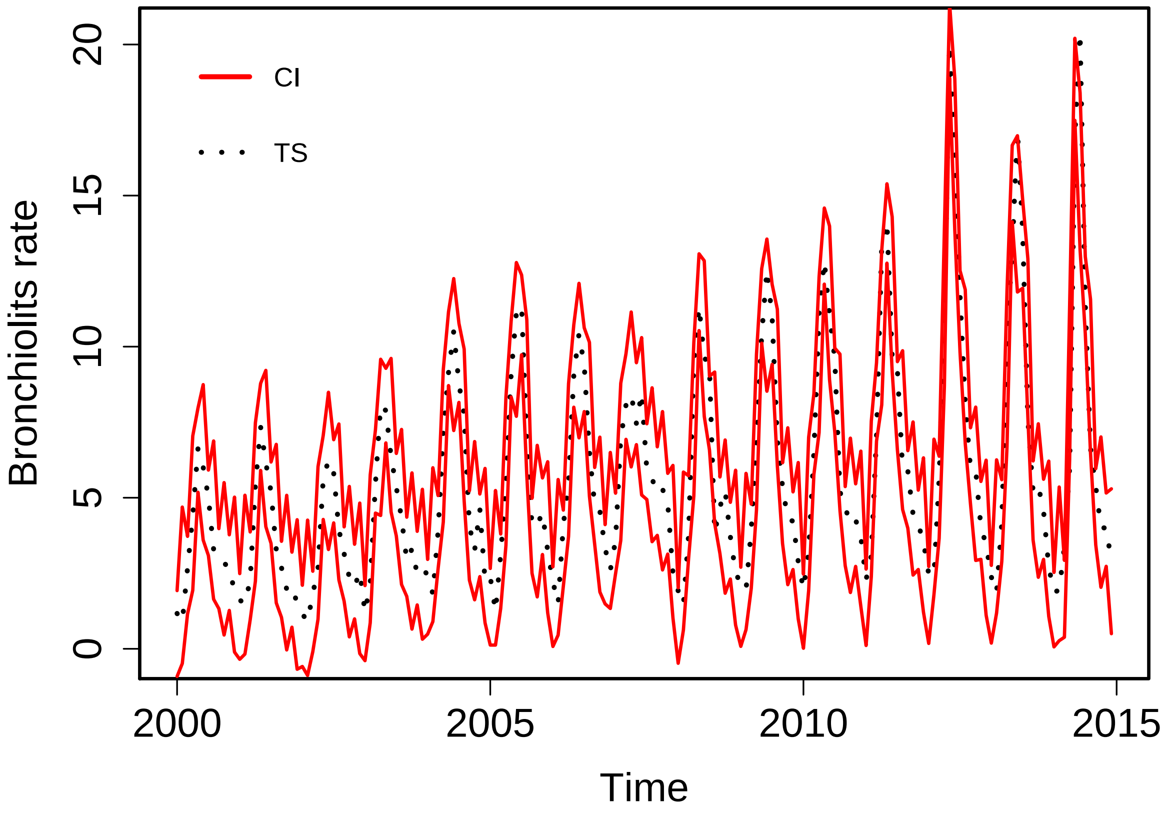

Figure 22 – Confidence interval obtained from TSWDWT

Figure 23 presents the confidence interval generated from reinflated TSWNDWT. The CI obtained from this method contains almost all the observed process values, and a few outside points. In general, we observe that the methods based on NDWT have a similar behavior.

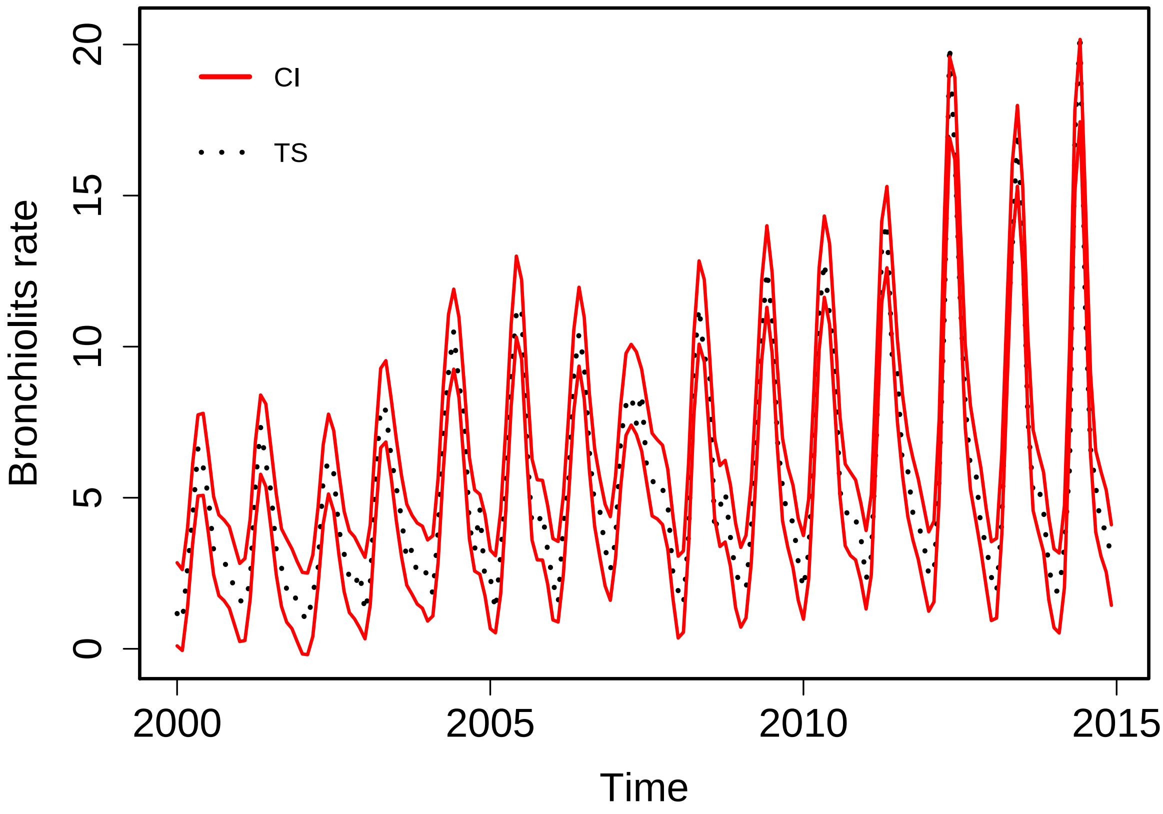

Figure 23 – Confidence interval obtained from reinflated TSWNDWT

In Figure 24, the confidence interval generated from the reinflated TSWNDWT are represented. The CI obtained from this method also contains almost all the observed process values, but as in TSWDWT, the time series that compose the confidence interval are more noisy than those NDWT based methods.

In general, the built confidence intervals include almost all the time series values that represent the mean of the stochastic process. But, when the bronchiolitis hospitalizations is hight producing spikes in the time series, mainly in May of 2012 and June 2014 the CI does not contain the time series values.

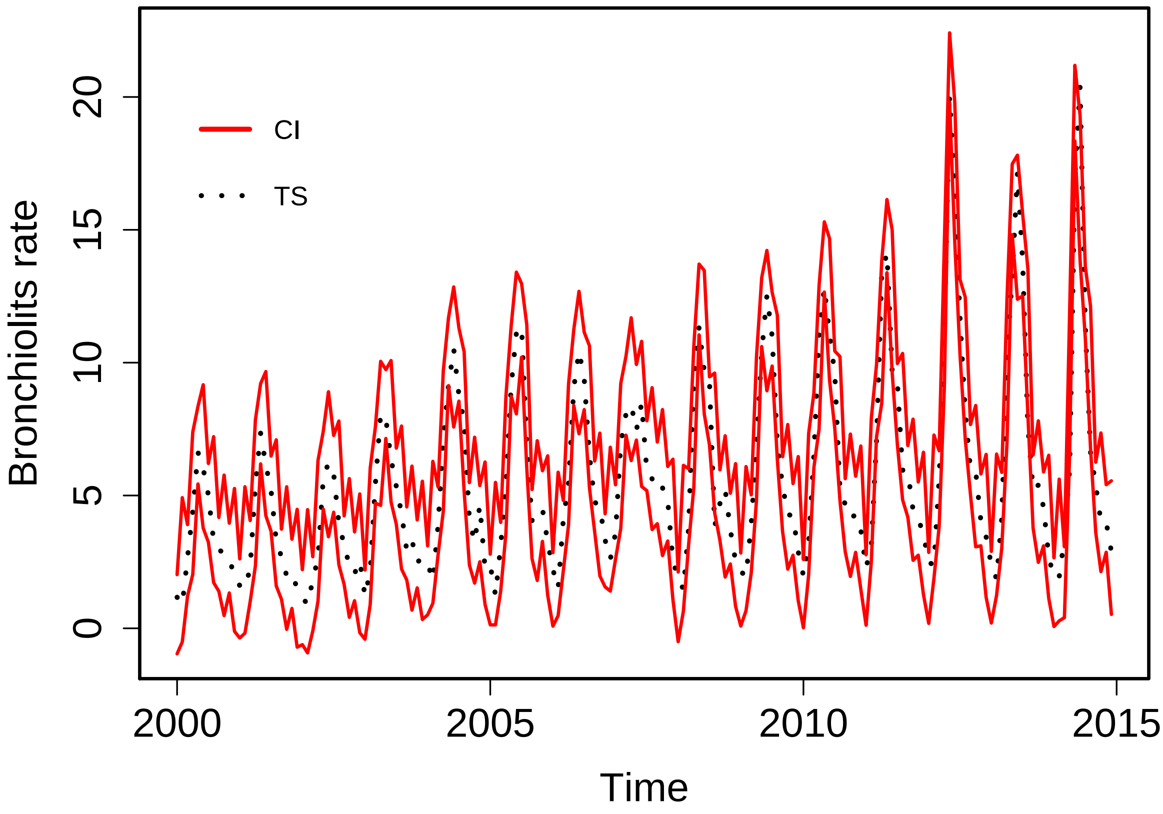

Figure 24 – Confidence interval obtained from reinflated TSWDWT

Although all the methods contain observed points that are not inside of the confidence intervals, those based on NDWT have less outside points. The TSWNDWT share more interesting results presenting low variability and the smaller bias.

4.5 Final considerations

The difficulty or impossibility in accessing more than one trajectory in a stochastic process such as the monthly rate of bronchiolitis hospitalizations is well known. Providing a method to estimate the uncertainty associated with the evolutionary stochastic process mean without considering the presupposition of normality is a challenging problem. With the presented possibilities, this problem can be taken into account from wavelet-based bootstrapping.

All the evaluated methods provide a measure of the confidence interval of the mean μ_t for the monthly hospitalization rate for bronchiolitis via wavelet decomposition using the Daubechies' wavelet $d4$. At the moment, we are analyzing these methods considering different wavelet families and vanishing moments, as well as other time series with diverse behaviors and lengths. In the literature, the usual methods for resampling time series are based on DWT. In this work, we observed that NDWT provides good estimates, with the smallest standard errors and coefficient of variations.

The generation of the confidence interval for μ_t can also be used to estimate the uncertainty for wavelet regression models, since they also represents of the mean of a stochastic process.

Chapter 5

Final considerations

In the development of this dissertation, firstly, it was possible to access many important time series characteristics. In a second moment, concepts of wavelet analysis and their applications to time series analysis were reviewed. And finally, the combination of these two areas in order to perform the resampling of time series.

In chapter 2 some considerations were made on the estimation of the Hurst exponent. It shows how important is to know the various arguments involved in the use of empirical estimation methods.

In order to estimate the uncertainty associated with the evolutionary mean of a stochastic process, five methods of resampling were investigated, they are naive bootstrap based on NDWT, TSWDWT, TSWNDWT, and the reinflated versions of TSWDWT, and TSWNDWT. All these methods combines the wavelet transform with bootstrap methods. Furthermore, these methods were used for resampling simulated time series, and the preservation of the Hurst statistic, MSE and bias were investigated. The estimated values of the Hurst exponent of the time series obtained by the bootstrap simulation were very close to those observed in the original time series. In this sense, the values of the bias and MSE were close to zero, especially for larger time series.

Considering the five evaluated methods, in chapter 3, a confidence interval for the time series of bronchiolitis hospitalization rate in Paraná State from 2004 to 2014 was estimated. The bronchiolitis is an infectious respiratory disease more frequent in children, in the first year of life (AMANTÉA; SILVA, 1998). Due to lack of particular vaccines and the fact the available drugs are of high cost and/or short effectiveness, estimating the uncertainty of this data is very important to understand indirectly the behavior of respiratory syncytial virus (RSV), the main causer of bronchiolitis, which may help the public health managers in taking the best decisions related, e. g., in medicine administration. Furthermore, because the non-stationary of these time series is important to emphasize that classical approaches as

SARIMA model could not be applied (GIMENES, 2015). On the other hand, wavelet-based methods have been shown an adequate methodology for dealing with this feature.

The estimated mean for the stochastic process, which has bronchiolitis hospitalization rate time series as the evolutionary mean, by using the bootstrap methods, presented similar behavior than the observed bronchiolitis time series. Furthermore bias, standard errors are close to zero. The CI constructed for bronchiolitis hospitalization rate time series, in general, contemplated the main features of this time series. In addition, only few outside points have been observed, and TSWNDWT presented the best average of bias.

Analyzing all methods we have observed that the choice of wavelet function and number of vanishing moments have developed important aspects. Changing the wavelet function or the vanishing numbers we can obtain more or less extrapolations in the confidence interval. Another point is to analyze more current and efficient methods for estimating the Hurst exponent as Beran et al. (2013), Kang and Vidakovic (2017), Feng and Vidakovic (2017). These choices are topics for future studies.

Bibliography

ABOUFADEL, E.; SCHLICKER, S. *Discovering wavelets*. [S.l.]: John Wiley & Sons, 2011.

AMANTÉA, S. L.; SILVA, F. d. A. Bronquiolite viral aguda-um tema ainda controvertido. *J. pediatr.(Rio J.)*, v. 74, n. sup. 1, p. 37–S47, 1998.

ANGELINI, C. et al. Resampling hierarchical processes in the wavelet domain: A case study using atmospheric turbulence. *Physica D: Nonlinear Phenomena*, Elsevier, v. 207, n. 1, p. 24–40, 2005.

BERAN, J. *Statistics for long-memory processes*. New York: CHAPMAM & HALLI/CRC PRESS, 1994. v. 61.

BERAN, J. et al. *Long-Memory Processes.* New York: Springer, 2013.

BOX, G. E. et al. *Time series analysis: forecasting and control*. New Jersey: John Wiley & Sons, 2015.

BREAKSPEAR, M.; BRAMMER, M.; ROBINSON, P. A. Construction of multivariate surrogate sets from nonlinear data using the wavelet transform. *Physica D: Nonlinear Phenomena*, Elsevier, v. 182, n. 1, p. 1–22, 2003.

BÜHLMANN, P. et al. Sieve bootstrap for time series. *Bernoulli*, Bernoulli Society for Mathematical Statistics and Probability, v. 3, n. 2, p. 123–148, 1997.

CHATFIELD, C. *The Analysis Of Time Series An Introduction*. 6. ed. Boca Raton: Chapman & Hall CRC, 2013.

CLEGG, R. G. A practical guide to measuring the hurst parameter. *arXiv preprint math/0610756*, 2006.

DAUBECHIES, I. Ten lectures on wavelets || 1. the what, why, and how of wavelets. In: ____. [S.l.: s.n.], 1992. v. 10.1137/1.9781611970104. ISBN 978-0-89871-274-2,978-1-61197-010-4.

DONOHO, D.; CANDES, E. Continuous curvelet transform: Ii. discretization and frames. *Applied and Computational Harmonic Analysis*, Citeseer, v. 19, n. 2, p. 198–222, 2005.

DONOHO, D. L.; JOHNSTONE, I. M. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, JSTOR, p. 425–455, 1994.

DONOHO, D. L.; JOHNSTONE, I. M. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association*, Taylor & Francis Group, v. 90, n. 432, p. 1200–1224, 1995.

EFRON, B.; GONG, G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, Taylor & Francis, v. 37, n. 1, p. 36–48, 1983.

FENG, C.; VIDAKOVIC, B. Estimation of the hurst exponent using trimean estimators on nondecimated wavelet coefficients. *arXiv preprint arXiv:1709.08775*, 2017.

FENG, H.; WILLEMAIN, T. R.; SHANG, N. Wavelet-based bootstrap for time series analysis. *Communications in Statistics - Simulation and Computation*, Taylor and Francis Group, v. 34, 04 2005.

GIMENES, E. *Análise e Modelagem de Séries Temporais Epidemiológicas no Domínio do Tempo e Frequencia*. Tese (Doutorado) — Universidade Estadual de Maringá, 2015.

GOLIA, S. Evaluating the gph estimator via bootstrap technique. In: SPRINGER. *Compstat*. [S.l.], 2002. p. 343–348.

HEINLEIN, P. Discretizing continuous wavelet transforms using integrated wavelets. *Applied and Computational Harmonic Analysis*, Elsevier, v. 14, n. 3, p. 238–256, 2003.

HIGUCHI, T. Approach to an irregular time series on the basis of the fractal theory. *Physica D: Nonlinear Phenomena*, Elsevier, v. 31, n. 2, p. 277–283, 1988.

HURST, H. E. Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civil Eng.*, v. 116, p. 770–808, 1951.

KANG, M.; VIDAKOVIC, B. Medl and medla: Methods for assessment of scaling by medians of log-squared nondecimated wavelet coefficients. *arXiv preprint arXiv:1703.04180*, 2017.

KANG, M. K. *Non-decimated wavelet transform in statistical assessment of scaling: Theory and applications*. Tese (Doutorado) — Georgia Institute of Technology, 2016.

KENDERDINE, R. D. Wavelet-based resampling techniques. *Bulletin of the Australian Mathematical Society*, Australian Mathematical Society, v. 85, 4 2012.

KEYLOCK, C. Improved preservation of autocorrelative structure in surrogate data using an initial wavelet step. *Nonlinear Processes in Geophysics*, v. 15, n. 3, p. 435–444, 2008.

KEYLOCK, C. J. Constrained surrogate time series with preservation of the mean and variance structure. *Physical Review E*, The American Physical Society, v. 73, 3 2006.

KEYLOCK, C. J. A wavelet-based method for surrogate data generation. *Physica D: Nonlinear Phenomena*, Elsevier, v. 225, n. 2, p. 219–228, 2007.

KHAREL, P. Rescaled range (r/s) analysis of the stock market returns. The University of the South, 2010.

KRAKOVSKÁ, H.; KRAKOVSKÁ, A. Fractal dimension of selfaffine signals: Four methods of estimation. *arXiv preprint arXiv:1611.06190*, 2016.

LAHIRI, S. N. *Resampling methods for dependent data*. New York: Springer Science & Business Media, 2013.

MALLAT, S. G. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, Ieee, v. 11, n. 7, p. 674–693, 1989.

MANDELBROT, B. B. *The fractal geometry of nature*. New York: Henry Holt and Company, 1982.

MEDEIROS, A. E.; SOUZA, E. M. Estimating the confidence interval of evolutionary stochastic process mean from wavelet based bootstrapping. In: *Frontiers of Biostatistics an Bioinformatics*. New York: Springer, 2018. To appear.

MORETTIN, P. A. Ondas e ondaletas. *Ondas e Ondeletas*, EdUSP São Paulo, v. 1, 1999.

MORETTIN, P. A.; TOLOI, C. *Análise de séries temporais*. 2. ed. São Paulo: Blucher, 2006.

NASON, G. *Wavelet methods in statistics with R*. New York: Springer Science & Business Media, 2010.

NASON, G.; SAPATINAS, T.; SAWCZENKO, A. Statistical modelling of time series using non-decimated wavelet representations. *submitted for publication*, 1997.

NASON, G. P.; SACHS, R. V.; KROISANDT, G. Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 62, n. 2, p. 271–292, 2000.

NASON, G. P.; SILVERMAN, B. W. The stationary wavelet transform and some statistical applications. In: *Wavelets and statistics*. [S.l.]: Springer, 1995. p. 281–299.

PERCIVAL, D.; SARDY, S.; DAVISON, A. Wavestrapping time series: Adaptive wavelet-based bootstrapping. *Nonlinear and nonstationary signal processing*, Cambridge, UK: Cambridge Univ. Press, p. 442-471, 2000.

PERCIVAL, D. B.; WALDEN, A. T. *Wavelet methods for time series analysis*. Cambridge: Cambridge university press, 2006. v. 4.

PICARD, D.; TRIBOULEY, K. Adaptive confidence interval for pointwise curve estimation. *Annals of statistics*, JSTOR, p. 298–335, 2000.

POLITIS, D. N.; ROMANO, J. P. The stationary bootstrap. *Journal of the American Statistical association*, Taylor & Francis Group, v. 89, n. 428, p. 1303–1313, 1994.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2016. Disponível em: <https://www.R-project.org/>.

RESTA, M. Hurst exponent and its applications in time-series analysis. *Recent Patents on Computer Science*, Bentham Science Publishers, v. 5, n. 3, p. 211–219, 2012.

RIZZO, M. L. *Statistical computing with R*. London: CRC Press, 2007.

TANG, L.; WOODWARD, W. A.; SCHUCANY, W. R. Undercoverage of wavelet-based resampling confidence intervals. *Communications in Statistics—Simulation and Computation®*, Taylor & Francis, v. 37, n. 7, p. 1307–1315, 2008.

TAQQU, M. S.; TEVEROVSKY, V.; WILLINGER, W. Estimators for long-range dependence: an empirical study. *Fractals*, World Scientific, v. 3, n. 04, p. 785–798, 1995.

VEDOVATTO, T. Medidas de memória longa em séries temporais: comparação de métodos de estimação do coeficiente de hurst. 2015.

VIDAKOVIC, B. [wiley series in probability and statistics] statistical modeling by wavelets || wiley series in probability and statistics. In: ____. [S.l.: s.n.], 1999. v. 10.1002/9780470317020. ISBN 9780470317020,9780471293651.

WEI, W. W.-S. *Time series analysis*. USA: Addison-Wesley publ Reading, 1994.

WORNELL, G.; OPPENHEIM, A. V. *Signal processing with fractals: a wavelet-based approach*. [S.l.]: Prentice Hall Press, 1996.

WUERTZ, D.; SETZ, T.; CHALABI, Y. *fArma: ARMA Time Series Modelling*. [S.l.], 2013. R package version 3010.79. Disponível em: <https://CRAN.R-project.org/package=fArma>.

YI, J. S. et al. Parallel wavestrap: Simulating acceleration data for mobile context simulator. *Current Developed in Theory and Applications of Wavelets*, v. 1, p. 251–272, 2007.