A WAVELET THRESHOLDING APPLICATION ON PARTICULATE MATERIAL CONCENTRATION TIME SERIES

UMA APLICAÇÃO DE LIMIAR DE ONDA NA SÉRIE DE TEMPO DE CONCENTRAÇÃO DE MATERIAL DE PARTÍCULADO

Alex Rodrigo dos Santos Sousa[1]

ABSTRACT
Denoising data is extremely important in data analysis to visualize patterns, estimate structural features present in the data and avoid misclassifications. In environmental field, it is of interest to identify days with high air pollutants levels that can impact on population health. Pollutants levels collected data in a certain region, as occur in any data collection procedure, have presence of random noise, which can lead to misclassifications of the real context of pollution at the considered region. In this sense, statistical denoising methods are welcome to reduce noise in the data. The present paper proposes the use of a soft wavelet thresholding with universal threshold policy for denoising particulate materials (PM) time series and identifying days with high concentration levels of these pollutants. The technique is applied on PM10 and PM2.5 time series collected by the São Paulo Environmental Company (CETESB) from Santos station during the period of 2018-2020.

KEYWORDS: Wavelets. Statistics. Particulate material. Air pollution.

RESUMO
A eliminação de ruído é extremamente importante na análise de dados para visualizar padrões, estimar características estruturais presentes nos dados e evitar classificações erradas. Na área ambiental, é de interesse identificar dias com elevados níveis de poluentes suspensos no ar que podem impactar na saúde da população. Entretanto, os dados coletados dos níveis de poluentes em uma certa região, assim como ocorre em qualquer procedimento de coleta de dados, apresentam presença de ruído aleatório, o que pode levar a erros de classificação do real contexto de poluição da região considerada. Nesse sentido, métodos estatísticos de eliminação de ruído são bem-vindos para reduzir o ruído nos dados de níveis de poluição do ar. O presente trabalho propõe o uso do estimador de coeficientes de ondaletas por limiar suave com política de escolha de limiar universal para redução de ruído em séries temporais envolvendo materiais particulados (PM) suspensos no ar e identificação de dias com altos níveis de concentração desses poluentes. A técnica é aplicada em séries temporais de PM10 e PM2.5 coletadas pela Companhia Ambiental de São Paulo (CETESB) na estação de Santos no período de 2018-2020.

PALAVRAS-CHAVE: Ondaletas. Estatística. Material particulado. Poluição do ar

---

[1] Universidade de São Paulo e Centro Universitário Campo Limpo Paulista

1. INTRODUCTION

Wavelet based statistical methods have been extensively applied to analyze data in several areas of science, such as engineering, genetics, ecology, physics and so on. In a nonparametric regression model, the focus of this work, it is possible to expand an unknown square integrable function as a linear combination of wavelet basis functions, in the same way as other basis functions, such as splines, Fourier and polinomials for example. But in opposite to these basis functions ones, wavelet basis expansions are typically sparse, i.e, most of the coefficients of the representation are zero. In fact, due the well localization of the wavelet basis in time and space, the wavelet coefficients of the expansion are nonzero only at positions that contain important features of the function, such as peaks, discontinuities, oscillations and/or minimum and maximum points. It allows the analysis of the underlying function only throw these few nonzero wavelet coefficients. See VIDAKOVIC (1999), NASON (2008) and MALLAT (1998) for excellent overviews about statistical methods based on wavelets.

Although the wavelet coefficients are essentially null in smooth parts of the function to be estimated, in practice, one observes data that are contaminated by noise, which is intrinsic to any data collection process. The impact of the noise presence in the data is noisy coefficients in wavelet domain, called empirical wavelet coefficients, i.e, empirical coefficients obtained after the application of a discrete wavelet transformation (DWT) on the original data also are noisy. Several methods have been proposed in the literature to estimate the true wavelet coefficients of the expansion by denoising the empirical coefficients, most of them are based on thresholding, originally proposed by the seminal papers of DONOHO (1993a), (1993b), (1995a) and (1995b) and DONOHO and JOHNSTONE (1994a), (1994b) and (1995). The main idea of thresholding estimators is reducing the magnitude of an empirical coefficient by application of a threshold, that shrinks sufficiently small empirical coefficients to zero. See VIDAKOVIC (1999) for descriptions of several already proposed thresholding methods in wavelets. See also ANGELINI and VIDAKOVIC (2004), REMÉNYI and VIDAKOVIC (2015), SOUSA (2020), SOUSA et al. (2020) and SOUSA (2021) for some recent developments on wavelet shrinkage procedures.

According to São Paulo Environmental Company (CETESB, 2021a, free translation), "Particulate Material is a set of pollutants consisting of dust, smoke and all types of solid and liquid material that remain suspended in the atmosphere because of their small size. The main sources of particulate emissions to the atmosphere are: automotive vehicles, industrial processes, biomass burning, re-suspension of dust from the ground, among others. Particulate matter can also form in the atmosphere from gases such as sulfur dioxide (SO2), nitrogen oxides (NOx) and volatile organic compounds (VOCs), which are mainly emitted in combustion activities, turning into particles such as result of chemical reactions in the air." There are several types of particulate

materials. We consider in this work the inhalable particles (PM10), that are particulate materials with aerodynamic diameter lesser than 10 μm and "can become trapped in the upper part of the respiratory system or penetrate deeper, reaching the pulmonary alveoli…" (CETESB, 2021a, free translation), and fine inhalable particles (PM2.5), that are particulate materials with aerodynamic diameter lesser than 2.5 μm and "penetrate deeply into the respiratory system and can reach the pulmonary alveoli." (CESTEB, 2021a, free translation). At the Brazilian state of São Paulo, CETESB has stations distributed around the state that measure, among several pollutants, PM10 and PM2.5 daily concentrations. It is of great interest to identify days with high PM10 and PM2.5 concentrations to analyze degree of pollution in the air, since the size of particles is directly associated with their potential to cause health problems, the smaller the greater the effects caused. Particulates can also reduce visibility into the atmosphere.

In this sense, the goal of this work is to identify days with high particulate material PM10 and PM2.5 concentrations measured in a station at Santos city during the period of 2018-2020 by denoising the original time series collected by the São Paulo Environmental Company (CETESB) with the application of a soft thresholding estimator with universal policy proposed by DONOHO and JOHNSTONE (1994b). This paper is organized as follows: some background about wavelets and data description are in Section 2. Wavelet analysis of the PM10 and PM2.5 time series and results are available in Section 3. Conclusions and final considerations are provided in Section 4.

2. METHODOLOGY

2.1. PRELIMINARIES

Wavelets are families of functions $\psi_{jk} = 2^{\frac{j}{2}}\psi(2^j x - k)$ constructed by translations and dilations of a function called mother wavelet ψ. In mathematical point of view, wavelets can be used to model a function f as $f(x) = \sum_{j,k} \theta_{jk} \psi_{jk}(x)$, where j, k are integers (j is called resolution level) and θ_{jk} are the wavelet coefficients. One of the main advantages of using wavelets is that the wavelet coefficients vector is usually sparse, in the sense that most of the coefficients are zero. The few nonzero coefficients concentrate all the interesting features of the function, such as peaks, discontinuities, for example. In a statistical setup, this idea is used to estimate an unknown function by transforming noise observations coming from this function to wavelet domain, which is done by a fast algorithm called discrete wavelet transform (DWT), denoising data in wavelet domain and estimate the unknown function by the inverse wavelet transformation (IDWT). For more details about wavelet basis, see DAUBECHIES (1992) and MALLAT (1998).

Let is suppose one has n ($n = 2^J$, J is integer) observations $y_1, \ldots, y_n$ of a time series coming from the following model,

$$y_i = f(t_i) + \varepsilon_i, \quad i = 1, \ldots, n \quad (1)$$

where f is an unknown function, t_i are locations in time domain and ε_i are zero mean independent normal random variables with unknown variance σ^2. In vector notation, we have

$$y = f + \varepsilon, \quad (2)$$

$y = [y_1, \dots, y_n]'$, $f = [f(t_1), \dots, f(t_n)]'$ e $\varepsilon = [\varepsilon_1, \dots, \varepsilon_n]'$. The goal is to estimate de unknown function f. For this purpose, we apply a discrete wavelet transform (DWT) on (2), which can be represented by a transform matrix W. We obtain the following model, in wavelet domain,

$$Wy = W(f + \varepsilon)$$

$$Wy = Wf + W\varepsilon$$

$$d = \theta + \epsilon, \quad (3)$$

where $d = Wy$, $\theta = Wf$ and $\epsilon = W\varepsilon$. For a specific coefficient of the vector d, we have the simple model

$$d = \theta + \epsilon, \quad (4)$$

where d is the empirical wavelet coefficient, θ is the coefficient to be estimated and ϵ is the normal random error with unknown variance σ^2. Since the method works coefficient by coefficient, we extract the typical resolution level and location subindices of d, θ and ϵ for simplicity. Note that, according to the model (4), d is normal distributed with mean equals θ and then, the problem of estimating a function f becomes a normal location parameter estimation problem in the wavelet domain for each coefficient.

The wavelet coefficients vector θ is typically sparse, i.e, most of its coefficients are zero. One of the most used method to estimate a specific coefficient θ is soft thresholding rule δ, proposed by DONOHO and JOHNSTONE (1994) and DONOHO (1995a), given by

$$\delta(d) = \begin{cases} 0, & if \ |d| \leq \lambda \\ sign(d)(|d| - \lambda), & if \ |d| > \lambda \end{cases}, \quad (5)$$

where $\lambda > 0$ is a chosen threshold value. The idea of the soft thresholding is to shrink sufficiently small empirical coefficient d to zero, according to a chosen threshold value λ. Then, the estimate $\hat{\theta}$ of θ is $\hat{\theta} = \delta(d)$. Figure 1 shows an example of soft thresholding rule with $\lambda = 10$. There are several proposed methods to choose λ, see NASON (2008). In this work, we used the universal thresholding proposed by DONOHO and JOHNSTONE (1994b),

$$\lambda_{univ} = \sigma\sqrt{2\log(n)}. \quad (6)$$

RECISATEC – REVISTA CIENTÍFICA SAÚDE E TECNOLOGIA
ISSN 2763-8405

Once the wavelet coefficients are estimated, we apply the inverse discrete wavelet transform (IDWT) on $\hat{\theta}$ to estimate f, i.e, $\hat{f} = W^t \hat{\theta}$. The full process of denoising is represented in Figure 2, adapted from SOUSA (2021).

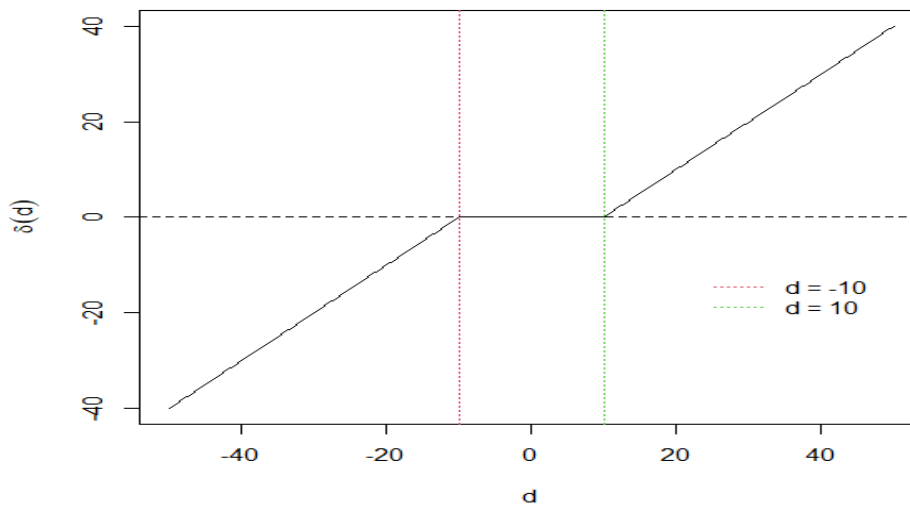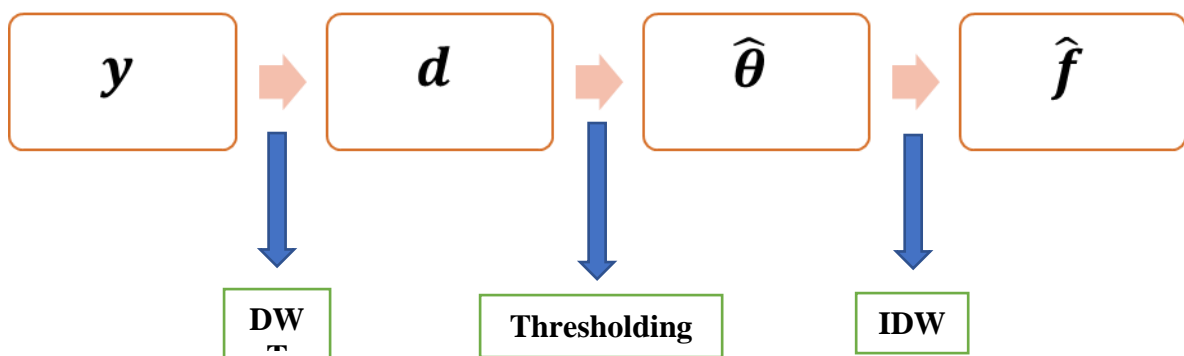**Figure 1: Soft thresholding rule with $\lambda = 10$.**

Figure 2: Denoising data process in wavelet domain by thresholding. Adapted from SOUSA (2021).

A WAVELET THRESHOLDING APPLICATION ON PARTICULATE MATERIAL CONCENTRATION TIME SERIES
Alex Rodrigo dos Santos Sousa

2.2. DATA DESCRIPTION

Our datasets consist in two time series with $n = 1024 = 2^{10}$ observations of daily concentrations (in $\mu g/m^3$) of particulate materials PM10 and PM2.5 collected from March, 14th 2018 to December, 31st 2020 in a station at Santos, SP. São Paulo State Environmental Company (CETESB, 2021b) classifies air quality according to levels of concentrations of several pollutants. Table 1 shows air quality classifications according to PM10 and PM2.5 levels. Note that for PM10, the critical concentration level for classification "Good" is 50 $\mu g/m^3$ and for PM2.5 is 25 $\mu g/m^3$.

Table 1: Air quality classification according to PM10 and PM2.5 concentration levels (in $\mu g/m^3$) used by São Paulo Environmental Company (CETESB).

Quality	PM10	PM2.5
Good	0 – 50	0 – 25
Moderate	> 50 – 100	> 25 – 50
Bad	> 100 – 150	> 50 – 75
Too bad	> 150 – 250	> 75 – 125
Terrible	> 250	> 125

Figures 3 and 4 show the time series of PM10 and PM2.5 respectively, with a horizontal line indicating the critical value for classification "Good" according do Table 1.

RECISATEC – REVISTA CIENTÍFICA SAÚDE E TECNOLOGIA
ISSN 2763-8405

Figure 3: Daily concentration ($\mu g/m^3$) of particulate material with diameter less or equal 10 μm (PM10) from March, 14th 2018 to December, 31st 2020. Red horizontal line indicates the critical value for air quality classification "Good" with relation to this pollutant
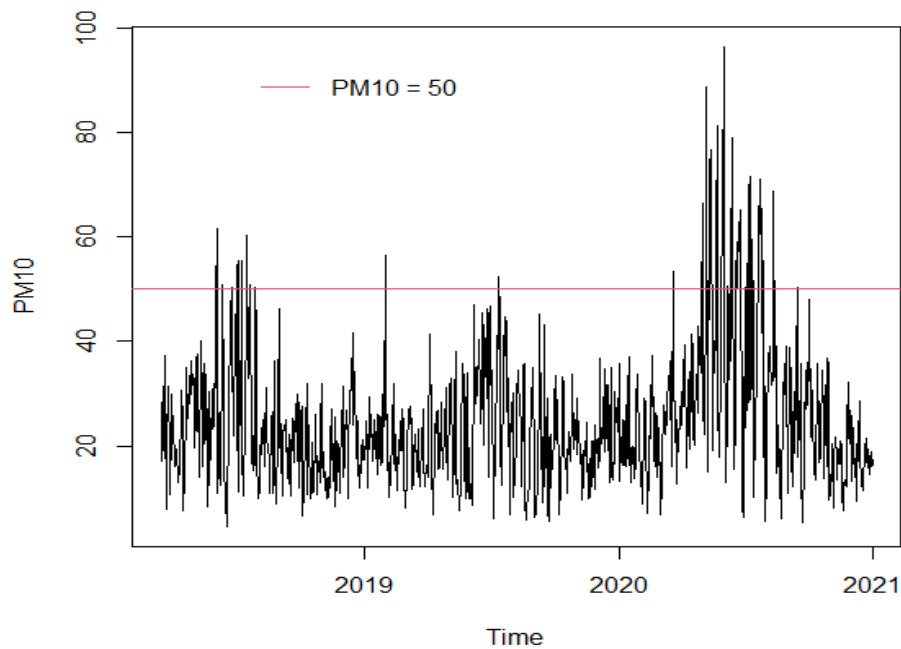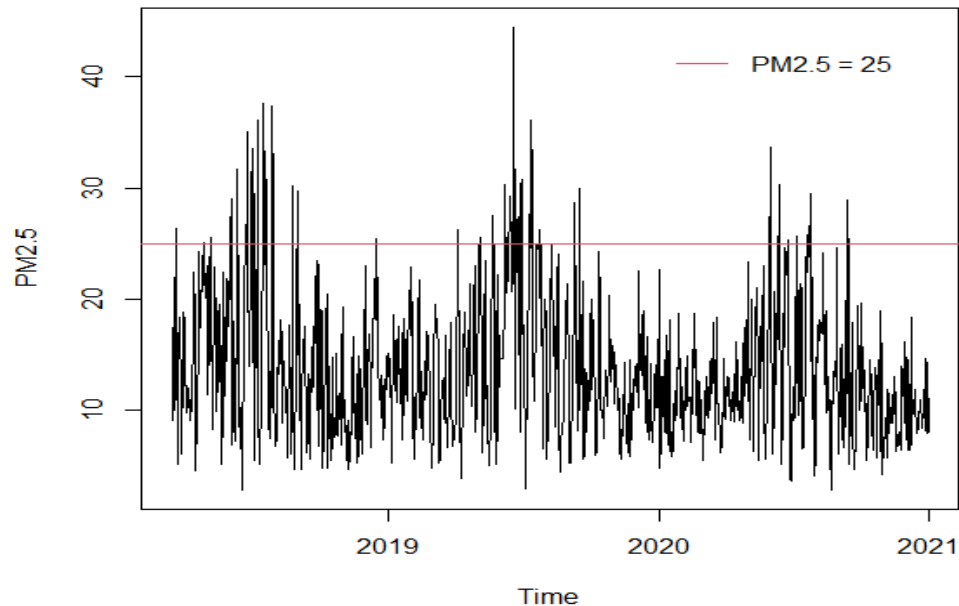
Figure 4: Daily concentration ($\mu g/m^3$) of particulate material with diameter less or equal 2.5 μm (PM2.5) from March, 14th 2018 to December, 31st 2020. Red horizontal line indicates the critical value for air quality classification "Good" with relation to this pollutant.


3. WAVELET ANALYSIS AND RESULTS

We applied a DWT on the time series PM10 and PM2.5 for denoising and analyzing these datasets on wavelet domain. To perform DWT, we chose Daubechies wavelet basis with 10 null moments, see VIDAKOVIC (1999).

3.1. PM 10 TIME SERIES ANALYSIS

After application of a DWT on PM10 time series, we obtained the vector of 1024 empirical wavelet coefficients, which are shown in Figure 5 by resolution level. In the figure, the lengths of the vertical traces are the magnitudes of the associated coefficients. One can observe the presence of noise mainly on higher resolution levels (levels 8 and 9), where typically the coefficients should be zero.

RECISATEC – REVISTA CIENTÍFICA SAÚDE E TECNOLOGIA
ISSN 2763-8405

Figure 5: Empirical wavelet coefficients by resolution level obtained after a DWT application on PM10 time series.
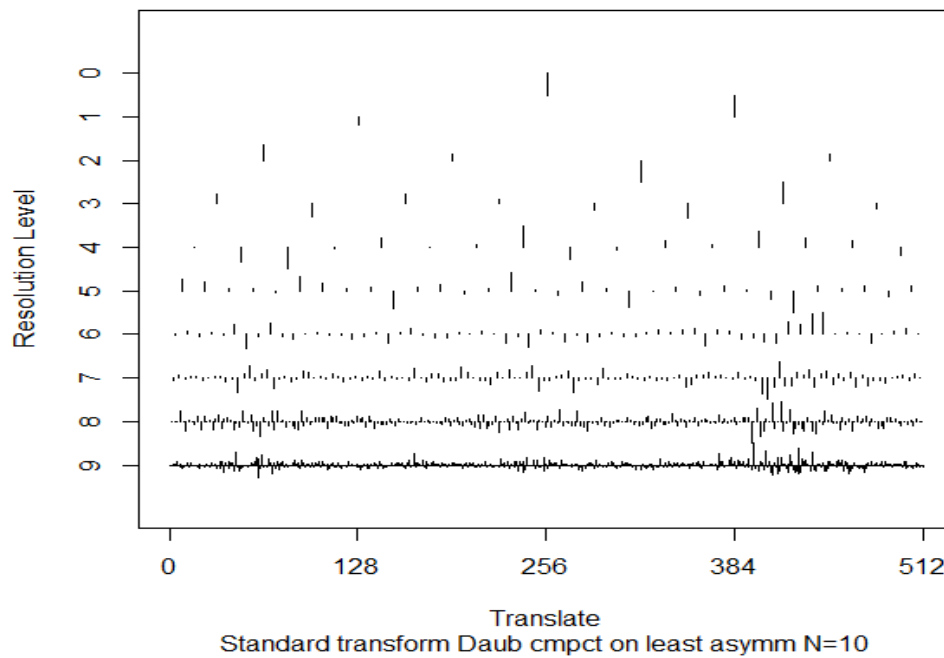

Due the noisy empirical coefficients on the highest resolution level, DONOHO and JOHNSTONE (1994b) suggest the following estimator for the noise standard deviation σ, denoted by $\hat{\sigma}$,

$$\hat{\sigma} = \frac{median\{|d_{J-1,k}| : k = 0, \dots, 2^{J-1}\}}{0.6745}, \quad (7)$$

where $d_{J-1,k}$ are the empirical coefficients of the highest resolution level ($J = 10$ in our case). The estimator (7) is important to perform universal thresholding (6) for denoising empirical coefficients and to obtain the estimates $\hat{\theta}$ of the wavelet coefficients θ. In PM10 time series, the noise standard deviation estimate, and its associated universal threshold are given by, according to equations (7) and (6) respectively, $\hat{\sigma} = 5.16$ and $\lambda_{univ} = 19.21$. The soft thresholding rule (5) shrinks to zero the empirical coefficients less than 19.21, which occurs in 976 of the 1024 PM10 empirical coefficients, or about 95% of them. Thus, after soft thresholding rule application, only 48 coefficients remain nonzero, making identification of the main features of the data easier, once these 48 nonzero coefficients concentrate all the important characteristics to be estimated from the data. Figure 6 shows the estimated wavelet coefficients by resolution level. Observe the significant noise reduction on higher resolution levels, where practically all coefficients were shrunk to zero.

Figure 6: Estimated wavelet coefficients of PM10 (shrunk empirical coefficients versions) by resolution level obtained throw universal thresholding application on empirical coefficients.

Standard transform Daub cmpct on least asymm N=10
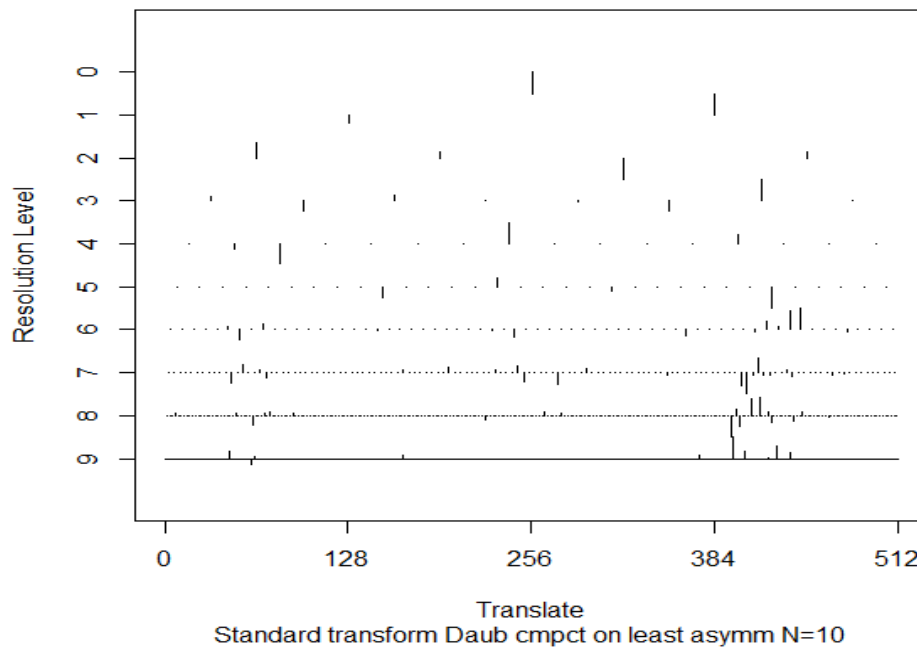
The smooth version of PM10 time series can be obtained after IDWT application on estimated coefficients vector and is presented in Figure 7 and 8. The first one shows the smooth time series and the original one for magnitude comparison. The second one shows only the smooth time series with the critical value for "good" quality according to Table 1, i.e, PM10 concentration equals to 50. In fact, after denoising, just 11 days remained above the critical value for "good" quality, most of them between May and July, 2020. These detected days are in Table 2.

Table 2: Smoothed PM10 concentrations ($\mu g/m^3$) above 50 (critical value for "good" air quality) and their associated days.

Day	Smoothed PM10 concentration
05/06/2020	64.69
05/12/2020	53.36
05/20/2020	55.23
05/21/2020	54.31
05/22/2020	58.14
05/30/2020	59.42
05/31/2020	63.43
06/11/2020	55.75
06/12/2020	58.44
07/05/2020	50.67
07/23/2020	50.09

Figure 7: Original and smooth daily PM10 concentrations ($\mu g/m^3$) time series in dashed black and red line, respectively.

A WAVELET THRESHOLDING APPLICATION ON PARTICULATE MATERIAL CONCENTRATION TIME SERIES
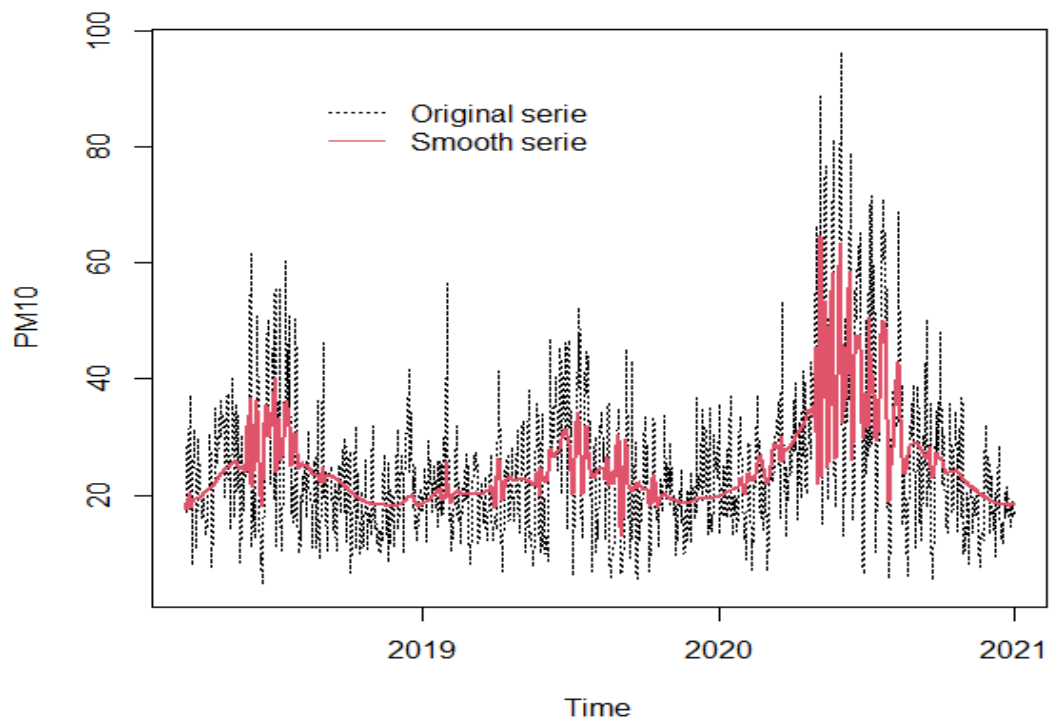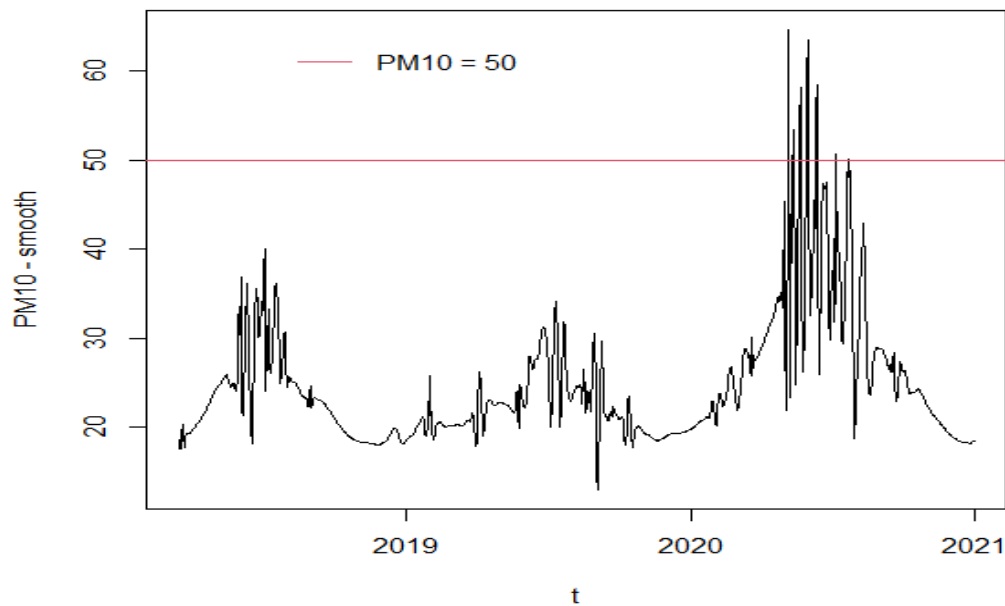Alex Rodrigo dos Santos Sousa

Figure 8: Smooth PM10 time series obtained by soft thresholding rule. Red horizontal line is the critical value for "good" air quality classification according to Table 1.


3.2. PM2.5 TIME SERIES ANALYSIS

Similar wavelet analysis was done in PM2.5 time series. The empirical coefficients are shown in Figure 9 and the estimated noise standard deviation and universal threshold according to (7) and (6) respectively are $\hat{\sigma} = 3.04$ and $\lambda_{univ} = 11.32$. After soft thresholding rule (5), 982 of the 1024 empirical coefficients were shrunk to zero, which represents 96% of them. Thus, the features of the time series became concentrated in only 42 nonzero wavelet coefficients. The estimated wavelet coefficients are available in Figure 10. As expected, most of the coefficients of the higher resolution levels were shrunk to zero due the presence of noise on these levels.

Figure 9: Empirical wavelet coefficients by resolution level obtained after a DWT application on PM2.5 time series.
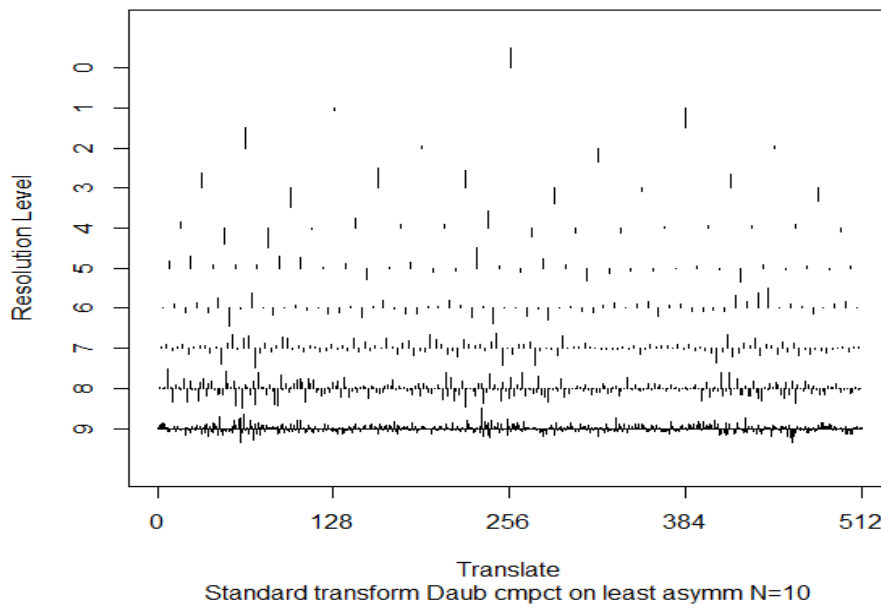
Figure 10: Estimated wavelet coefficients of PM2.5 (shrunk empirical coefficients versions) by resolution level obtained throw universal thresholding application on empirical coefficients.
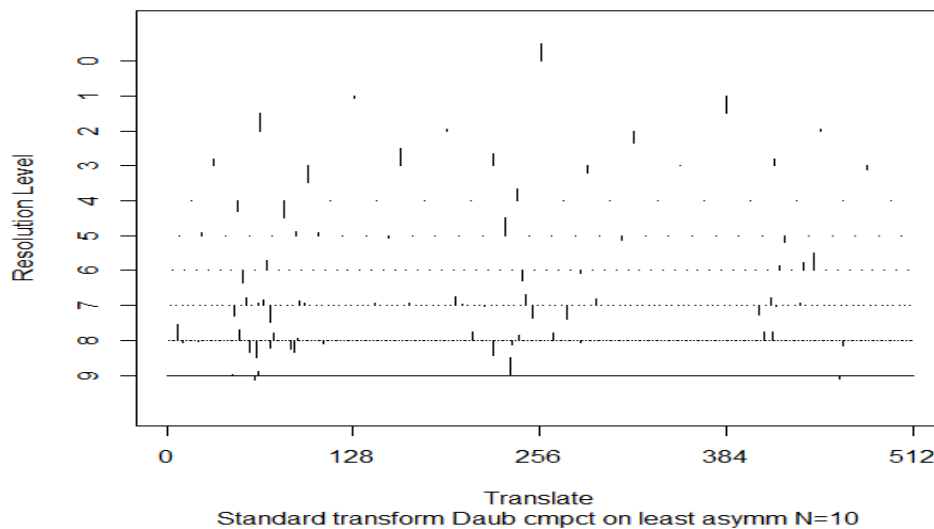
**RECISATEC – REVISTA CIENTÍFICA SAÚDE E TECNOLOGIA**
ISSN 2763-8405

The smooth PM2.5 time series is plotted with the original one in Figure 11 and separately in Figure 12. In the first one, it is possible to observe the smoothness degree applied to the original time series, i.e, the action of the denoising process by thresholding. After smoothness, only three days had PM2.5 concentration above critical value of 25 for "good" air quality classification according to Table 1. These days are set in Table 3.

Figure 11: Original and smooth daily PM2.5 concentrations ($\mu g / m^3$) time series in dashed black and red line, respectively.
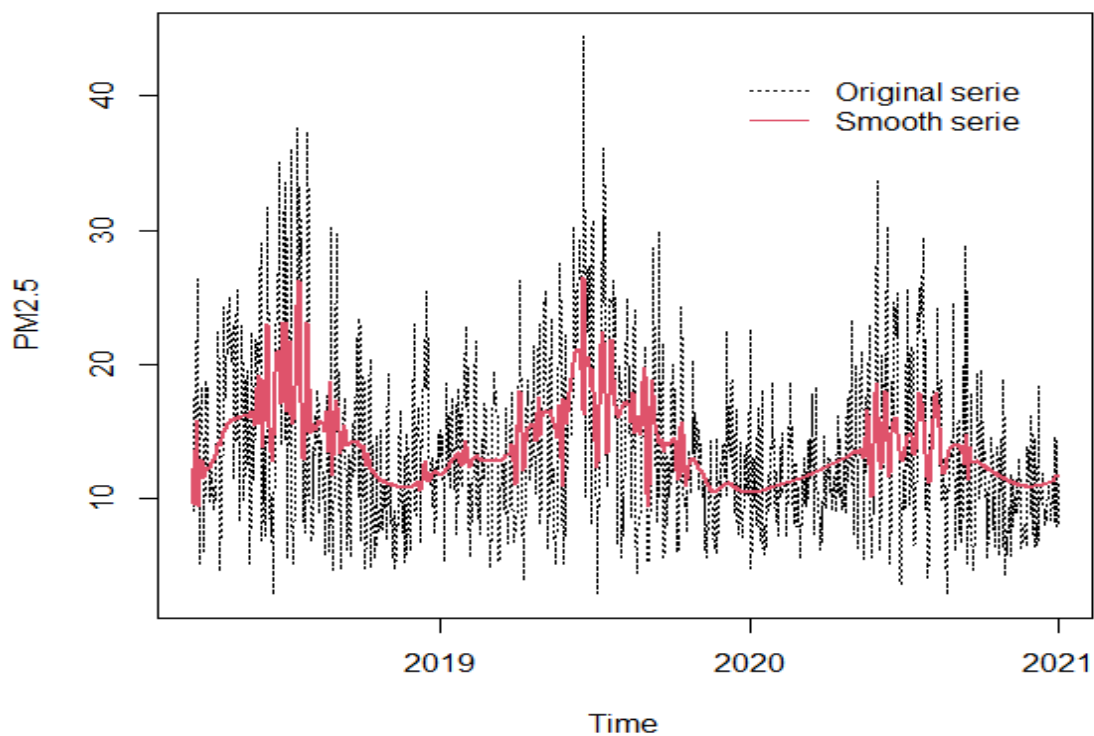
Figure 12: Smooth PM2.5 time series obtained by soft thresholding rule. Red horizontal line is the critical value for "goog" air quality classification according to Table 1.
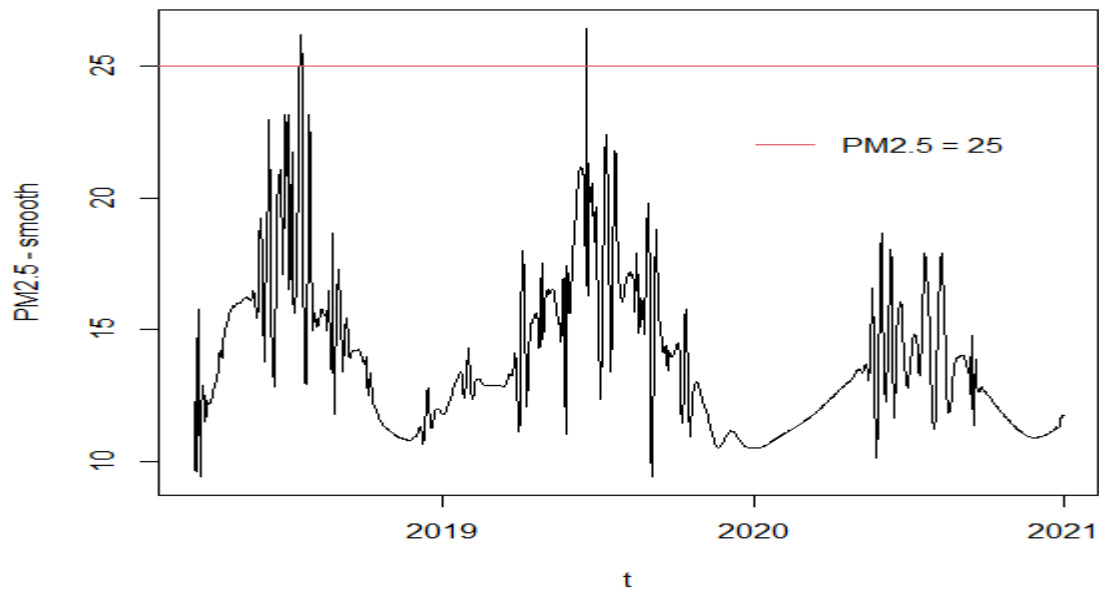
Table 3: Smoothed PM2.5 concentrations ($\mu g/m^3$) above 25 (critical value for "good" air quality) and their associated days.

Day	Smoothed PM2.5 concentration
07/17/2018	25.69
07/18/2018	26.19
06/19/2019	26.46

4. CONCLUSIONS

We applied soft wavelet thresholding with universal threshold policy for denoising wavelet coefficients and identifying days with high PM10 and PM2.5 concentrations levels at Santos station during the period of 2018 to 2020. In fact, the estimation procedure allowed to concentrate the analysis on a few number (about 5% of the sample size) of nonzero wavelet coefficients and smooth the time series, which facilitate visualization and identification of the desirable days.

A WAVELET THRESHOLDING APPLICATION ON PARTICULATE MATERIAL CONCENTRATION TIME SERIES
Alex Rodrigo dos Santos Sousa

Without the denoising process, days with good concentrations of PM10 or PM2.5 can be misclassified as not good concentrations days due the presence of noise in the collected data. In this sense, the application of the thresholding estimator is welcome for denoising data and analyzing them on a sparse set of coefficients.

Applications of the method in other discriminant levels of concentrations can be done. Further, the behavior of the estimator when other wavelet basis and/or threshold policy is adopted should be investigated as future works.

REFERENCES

[1] ANGELINI, C.; VIDAKOVIC, B. Gama-minimax wavelet shrinkage: a robust incorporation of information about energy of a signal in denoising applications. **Stat Sin**, v. 14, p. 103-125, 2004.

[2] CETESB. **Poluentes**. São Paulo: Cetesb, 2021a. Disponível em: https://cetesb.sp.gov.br/ar/poluentes/. Acesso em: 23 ago. 2021.

[3] CETESB. **Padrões de Qualidade do Ar**. Cetesb, 2021b. Disponível em: https://cetesb.sp.gov.br/ar/padroes-de-qualidade-do-ar/. Acesso em: 23 ago. 2021.

[4] DAUBECHIES, I. **Ten Lectures on Wavelets**. Philadelphia: SIAM, 1992.

[5] DONOHO, D. L. Nonlinear wavelet methods of recovery for signals, densities, and spectra from indirect and noisy data. in **Proceedings of Symposia in Applied Mathematics**, American Mathematical Society, v. 47, 1993a.

[6] DONOHO, D. L. Unconditional bases are optimal bases for data compression and statistical estimation. **Appl Comput Harmonic Anal**, v. 1, p. 100-115, 1993a.

[7] DONOHO, D. L. De-noising by soft-thresholding. **IEEE Trans Inf Theory**, v. 41, p. 613–627, 1995a.

[8] DONOHO, D. L. Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. **Appl Comput Harmonic Anal**, v. 2, p. 101–26, 1995b.

[9] DONOHO, D. L.; JOHNSTONE, I. M. Ideal denoising in an orthonormal basis chosen from a library of bases. **Compt Rend Acad Sci Paris A**, v. 319, p. 1317-1322, 1994a.

[10] DONOHO, D. L., JOHNSTONE, I. M. Ideal spatial adaptation by wavelet shrinkage. **Biometrika**, v. 81, p. 425-455, 1994b.

[11] DONHO, D. L, JOHNSTONE, I. M. Adapting to unknown smoothness via wavelet shrinkage. **J Am Stat Assoc**, v. 90, p. 1200–1224, 1995.

[12] MALLAT, S. G. **A Wavelet Tour of Signal Processing**. San Diego: Academic Press, 1998.

[13] NASON, G. P. **Wavelet Methods in Statistics with R**. New York: Springer, 2008.

[14] REMÉNYI, N.; VIDAKOVIC, B. Wavelet shrinkage with double weibull prior. **Commun Stat Simul Comput**, v. 44, n. 1, p. 88-104, 2015.

[15] SOUSA, A. R. S. Bayesian wavelet shrinkage with logistic prior. **Communications in Statistics: Simulation and Computation**, 2020. DOI: 10.1080/03610918.2020.1747076.

[16] SOUSA, A. R. S, GARCIA, N. L.; VIDAKOVIC, B. Bayesian wavelet shrinkage with beta prior. **Computational Statistics**, v. 3, p. 1341-1363, 2020.

[17] SOUSA, A. R. S. Encolhimento bayesiano de coeficientes de ondaletas sob priori de Champernowne com aplicações. **RECIMA21 - Revista Científica Multidisciplinar**, v. 2, n. 2, p. 92-116, 2021. ISSN 2675-6218.

[18] VIDAKOVIC, B. **Statistical Modeling by Wavelets**. New York: Wiley, 1999.