Ingeniería
ISSN: 0121-750X
revista_ing@udistrital.edu.co
Universidad Distrital Francisco José de Caldas
Colombia

Plazas-Nossa, Leonardo; Ávila A., Miguel A.; Torres, Andres
Detection of Outliers and Imputing of Missing Values for Water Quality UV-VIS Absorbance Time Series
Ingeniería, vol. 22, núm. 1, 2017, pp. 111-124
Universidad Distrital Francisco José de Caldas
Bogotá, Colombia

Methodology

REVISTA Ingeniería

Detection of Outliers and Imputing of Missing Values for Water Quality UV-VIS Absorbance Time Series

Detección de Valores Extremos e Imputación de Valores Faltantes para la Calidad de Agua en Series de Tiempo de Absorbancia UV-VIS

Leonardo Plazas-Nossa [*1] ***, Miguel A. Ávila A.*** [1] ***, Andres Torres*** [2]

[1] Universidad Distrital Francisco José de Caldas. Bogotá - Colombia,

[2] Pontificia Universidad Javeriana. Bogotá - Colombia

[*]Correspondence: lplazasn@udistrital.edu.co

Recibido: 05-04-2016. Modificado: 19-09-2016. Aceptado: 03-01-2017

Abstract

Context: The UV-Vis absorbance collection using online optical captors for water quality detection may yield outliers and/or missing values. Therefore, pre-processing to correct these anomalies is required to improve the analysis of monitoring data. The aim of this study is to propose a method to detect outliers as well as to fill-in the gaps in time series.

Method: Outliers are detected using Winsorising procedure and the application of the Discrete Fourier Transform (DFT) and the Inverse of Fast Fourier Transform (IFFT) to complete the time series. Together, these tools were used to analyse a case study comprising three sites in Colombia ((i) Bogotá D.C. Salitre-WWTP (Waste Water Treatment Plant), influent; (ii) Bogotá D.C. Gibraltar Pumping Station (GPS); and, (iii) Itagüí, San Fernando-WWTP, influent (Medellín metropolitan area)) analysed via UV-Vis (Ultraviolet and Visible) spectra.

Results: Outlier detection with the proposed method obtained promising results when window parameter values are small and self-similar, despite that the three time series exhibited different sizes and behaviours. The DFT allowed to process different length gaps having missing values. To assess the validity of the proposed method, continuous subsets (a section) of the absorbance time series without outlier or missing values were removed from the original time series obtaining an average 12 % error rate in the three testing time series.

Conclusions: The application of the DFT and the IFFT using the 10 % most important harmonics of useful values, can be advantageous for its later use in different applications, specifically for time series of water quality and quantity in urban sewer systems. One potential application would be the analysis of dry weather affecting rainy seasons, a feature achieved by detecting values that correspond to unusual behaviour in a time series. Additionally, the results hint at the potential of the method in correcting other hydrologic time series.

Keywords: Imputing missing values, outlier detection, UV-vis absorbance, water quality, winsorizing.

Language: English

Citación: L. Plazas, M. A. Ávila, A. Torres, "Detection of Outliers and Imputing of Missing Values for Water Quality UV-VIS Absorbance Time Series" INGENIERÍA, vol. 22, no. 1, pp. 111-124, 2017.
 DOI: http://dx.doi.org/10.14483/udistrital.jour.reving.2017.1.a01

Resumen

Contexto: El registro de la absorbancia UV-Vis mediante captores ópticos en línea para la detección de la calidad del agua, en donde se pueden presentar valores atípicos o valores faltantes. Por lo tanto, el pre-procesamiento para corregir dichas anomalías es necesario para un mejor análisis de los datos de monitoreo. El objetivo de este estudio es proponer un método para detectar e imputar valores extremos como también completar valores faltantes en series de tiempo.
Método: La detección de valores atípicos utiliza el procedimiento de enventaneo y la aplicación de la Transformada Discreta de Fourier (DFT –Discrete Fourier Transform) y la inversa de la Transformada Rápida de Fourier (IFFT–Inverse of Fast Fourier Transform) para completar las series de tiempo. Estas herramientas fueron utilizadas para un caso de estudio compuesto por tres sitios en Colombia (i) PTAR-Salitre (Planta de Tratamiento de Aguas Residuales) Bogotá D.C., afluente; (ii) Estación Elevadora de Gibraltar Bogotá D.C.; y (iii) PTAR-San Fernando, área metropolitana de Medellín, afluente) analizados mediante espectros UV-Vis (Ultravioleta y Visible).
Resultados: La detección de valores atípicos con el método propuesto obtiene resultados prometedores cuando los valores de los parámetros de la ventana son pequeños y auto-similares, esto a pesar de que las tres series de tiempo utilizadas presentan diferentes tamaños y comportamientos. Para validar la metodología propuesta, sub-conjuntos continuos (una sección) de las series de tiempo de absorbancia sin valores ausentes o atípicos, fueron removidos de las series original obteniéndose tasas de error de 12 % en promedio para todos los tres sitios de estudio.
Conclusiones: La aplicación de la DFT y la IIFT, utilizando el 10 % de los harmónicos más importantes de los valores útiles es crucial para su posterior uso en diferentes aplicaciones, específicamente para series de tiempo de calidad y cantidad de agua en sistema de saneamiento urbano. Una posible aplicación podría ser la comparación de los efectos de clima seco respecto a temporadas de lluvia, mediante la detección de valores que corresponden a comportamiento inusual en una serie de tiempo. Además, los resultados indican potencial aplicación futura en la corrección de otras series de tiempo hidrológicas.
Palabras clave: Absorbancia UV-Vis, calidad de agua, detección de valores extremos, enventaneo, imputación de valores faltantes.

1. Introducción

Continuous on-line measurements such as UV-Vis spectrometry is increasingly applied technique for water quality measurement in sewer systems [1]–[5]. These continuous time series [6]–[8] help to estimate pollutant concentrations in sewer systems and offer real-time control applicability [9]–[13]. Absorbance (essentially surrogate measurements for TSS – Total Suspended Solids or COD – Chemical Oxygen Demand) time series can be used to estimate the dry weather contribution to total TSS and COD loads measured during storm events [14], [15]. In fact, most models of storm weather pollutant loads in combined sewer systems are based on the assumption that total storm event load is the sum of dry and wet weather contributions, with the latter including surface runoff plus possible erosion of deposits accumulated in sewers [16]–[19]. Therefore, during rain events, it is important to distinguish the fraction of flow rate corresponding to dry weather and that corresponding to wet weather. This process, however, is hindered by the fact that the time series may present outliers. Johnson and Wichern [20] define an outlier as "an observation in a data set which appears to be inconsistent with the remainder of that set of data" [21]–[23], [25], [28]. Researches

had been used different methodologies to detect the isolated and extremely peak values as PCA by [23], also using statistical procedures based on prior knowledge of the system that produces the data [21], [24]–[28]. Others researches have been applied artificial intelligence and machine learning methodologies [22], [29]–[31]. In addition, lost values, due to obstructions on the sensors themselves or the occasional sensors removal required for maintenance, would be also present; researches have been using different methodologies for infilling missing values in data sets as Artificial Neural Networks (ANN) by [32], Evolutionary Algorithms by [33], Genetics Algorithms by [34], clustering by [35], using PCA by [36], Bayes' theorem [37], linear and non-linear regressions [38]–[40]. Several detection outlier techniques require to obtain a fitted model [41] to assess which are the extremely peaks to be removed and the infilling methods require multivariate information to obtain a model to fill the gaps in data sets.

The methodology laid out in this article detects and removes outliers and fill gaps making the missing values imputation in time series. In theory, the proposed procedure combines statistical values, outlier detection [42], [43], DFT and IFFT to complete time series. Thus, it is used the combination of the Winsorising and DFT procedures to perform the task of detection, removal of outliers and imputing of missing values to maintain the observed periodic behaviour of these time series. In practice, this methodology was applied to three time series at the same number of study sites in Colombia for UV-Vis spectra. Locations of the study sites are as follows: (i) Bogotá D.C. Salitre-WWTP (Waste Water Treatment Plant), influent; (ii) Bogotá D.C. Gibraltar Pumping Station (GPS); and, (iii) Itagüí, San Fernando-WWTP, influent (Medellín metropolitan area).

2. Materials and Methods

The spectro::lyser^{TM} UV-Vis sensors deployed at the three Colombian sites are submersible probes with a length of approximately 65 cm and a diameter of approximately 44 mm. Functionally speaking, they register light attenuation (absorbance) on-line in relative continuous time (one signal per minute). These sensors work with a light source provided by a Xenon lamp for wavelengths ranging from 200 nm to 750 nm, with intervals of 2.5 nm [1], [44].

With regard to the duration of the study, i.e. data gathered, the three time series are displayed in Figure 1: (i) Bogotá D.C. Salitre-WWTP, influent (5705 records, one per minute, from June 29th, 2011 at 9:03 h to July 3rd, 2011 at 17:33 h); (ii) Bogotá D.C. Gibraltar Pumping Station (GPS) (35684 records, one per minute, from October 18th, 2011 at 16:17 h to November 11th, 2011 at 11:20 h); and (iii) Itagüí, San Fernando-WWTP, influent (Medel- lín) (107204 records, one every two minutes, from September 24th, 2011 at 11:08 h to February 20th, 2011 at 10:18 h.

As Figure 1 aptly demonstrates, outliers accompanied our data collection. Outliers, those data that deviate significantly from the majority of observations, may be caused by different mechanisms in relation to normal data [21], [24]–[28]. These different mechanisms include, but are not limited to, factors such as sensor noise, process disturbance and instrument degradation. Relying on a time series tainted by outliers proves to be a source of frustration, as it often leads to misinterpretation and model misspecification. To help offset this aspect of data collection, data pre-processing is a necessary pre-requisite to monitoring data processing [45]. In order to detect and imputing these

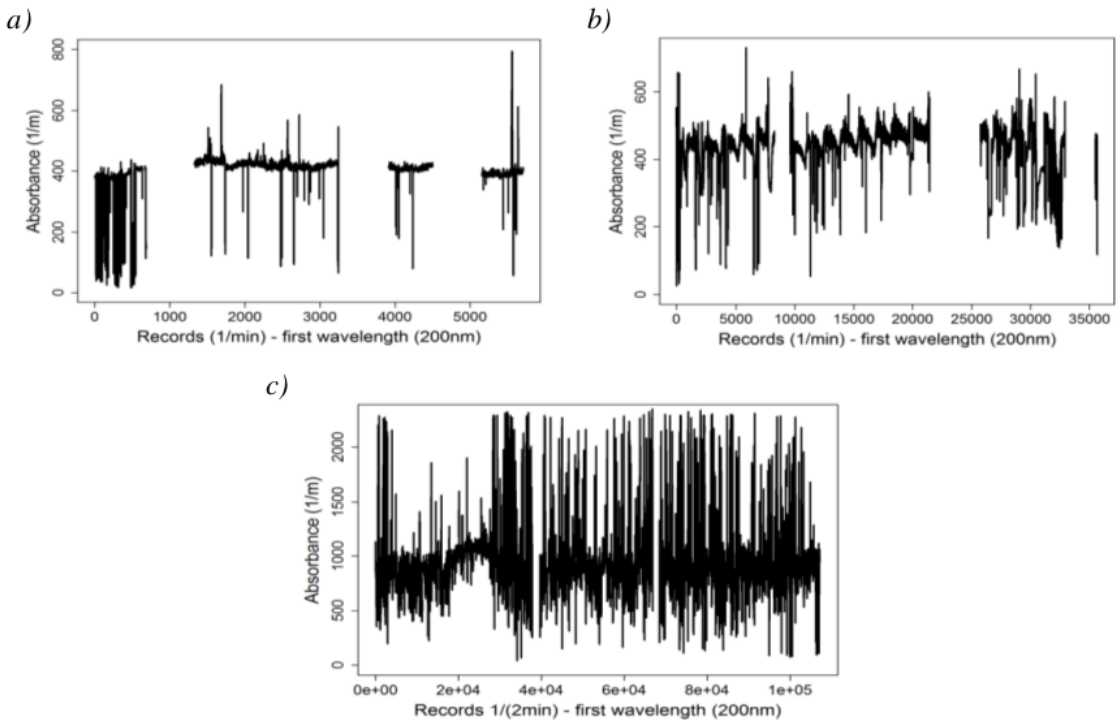

Figure 1. Original time series with outliers and missing values for Salitre-WWTP (a), GPS (b) and San Fernando-WWTP (c) (self- authorship)

pesky outliers, the Winsorising procedure, which consists of transforming the original data by limiting extreme values, reduces the effect of possibly spurious outliers [46]. Moreover, Winsorising is known for promoting a smoother filter, allowing researchers to place more weight on the central value of each time window by using a mobile window of values (N) [47], [48]. Such values depend on r, the number of values to be modified, where $2r + 1$ equals window size (N); yet, here it is incumbent upon researchers to set the range m of values to be analysed, for m is required to establish the minimum and maximum values for the window range and allow the removal of values beyond the upper or lower limits with regard to m in order to obtain the winsorised data [47]. The process carried out in this investigation can be explained in terms of two multi-part steps.

The first consists of sorting the data in the selected window from lowest to highest; several visual reviews were done to test values of r and m parameters. These values were defined once was observed the removal of outliers and before the shape of the signal has lost. Then, values below the minimum acceptable value (Min_{AV}) are removed. Min_{AV} is understood as the value of $(m+1)th$ place for all values in the window. Min_{AV} is represented by Equation 1.

$$Z_i = \begin{cases} X_i & if \quad X_i > Min_{AV} \\ Min_{AV} & otherwise \end{cases} \tag{1}$$

where X_i is the ith window value, Min_{AV} the minimum acceptable value and Z_i the resulting values after the removal of values below Min_{AV}. As a result, any value lower (outlier) than Min_{AV} will be replaced. A similar procedure is followed to remove outliers above a threshold Max_{AV}, which is equal to the value of the $(N - m)th$ place of all values in the Z_i window (see Equation 2).

$$Y_i = \begin{cases} Z_i & if \quad Z_i < Max_{AV} \\ Max_{AV} & otherwise \end{cases} \tag{2}$$

where Z_i is the ith window value, Max_{AV} the maximum acceptable value and Y_i the resulting values after removal. It is worth mentioning here that the above, Equation 1 and Equation 2, adapts the method proposed by [46].

Once the corrected time series (without outliers) is obtained, a DFT procedure to complete the time series can be applied. With DFT, the time series is calculated to facilitate the switch from the time domain to the frequency domain. This technique involves the conversion of a finite number of equally spaced samples (discrete points) into a number of coefficients that stem from a finite combination of complex sinusoid components. Doing so ensures that the frequency domain is comprised of the same number of sample values as the previous time domain [48], [49].

Equation 3 details this conversion. DFT likewise ranks components based on their importance, with importance determined by component amplitude. After ranking, components are eliminated from lower to higher importance, leaving only the most significant harmonics (10 %) in the resulting values. Finally, the Inverse Fast Fourier Transform (IFFT) is utilized to convert complex sinusoids (harmonics) into a finite number of discrete points. This movement represents a return to the time domain (as opposed to the time to frequency shift found in the initial DFT process), as shown in Equation 4. For more details, see [50].

$$X_k = \sum_{n=0}^{N-1} x_n \, e^{-j2\pi \frac{k}{N} n} \qquad k = 0, 1, \ldots, N-1 \tag{3}$$

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k \, e^{-j2\pi \frac{k}{N} n} \qquad n = 0, 1, \ldots, N-1 \tag{4}$$

where n is the current sample that is analysed, N is the total number of time samples taken, k is current harmonic (frequency) that is considered (0 Hertz up to $N-1$ Hertz), x_n is the value of the time series at time n and X_k is the amount of frequency k in the signal (complex number) [49].

The Winsorising process (first step) is applied to the absorbance time series. However, outlier values may still remain; these values must be dealt with. Having outlined the first step, we can proceed to the second step. The procedure obtains the median value from the entire dataset; next, any upper and lower values within median value +/- one standard deviation range will be discarded (referred to round one). In fact, one standard deviation is used because the time series is highly affected by outlier values that remain after the first round and before DFT is applied. Consequently, the dataset must be switched from the time to the frequency domain. This switch, done with values present in the data set prior to the first missing values gap, naturally leads to the selection of the two most important harmonics. These two harmonics are selected "naturally" insofar as they reproduce the pattern and dynamic of the events. In effect, it is possible to reproduce the outliers if more than two harmonics are included. By virtue of IFFT, the data is translated from the frequency domain back into the time domain. Afterwards, the resulting time series is used to complete the first gap of missing values. Then, the median value from the entire dataset is again obtained and any upper and

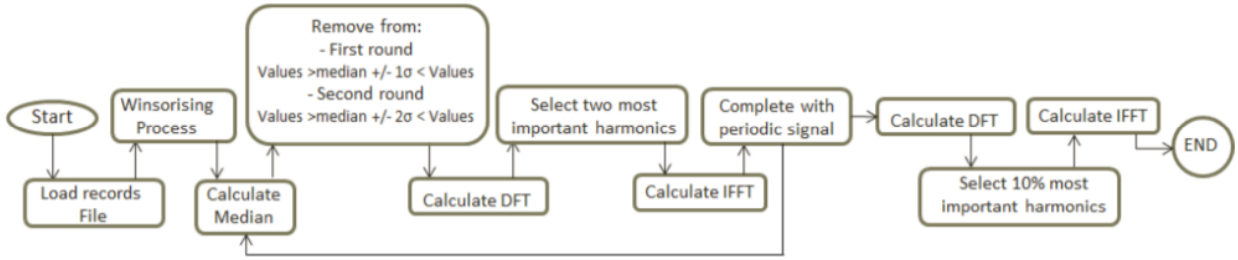

Figure 2. Flow diagram depicting detection and removal of outliers and replacement of missing values (self-authorship)

lower value within median value +/- two standard deviations range will be taken out (referred to as round two).

On account of Winsorising, along with the two rounds of standard deviation refinement, the effect of outliers at this stage has been attenuated but not fully counteracted. So, the DFT process is run again with only the two most important harmonics applied, followed by another run of IFFT. The resultant time domain values are then able to facilitate the completion of the first missing values gap. When this process is completed, the first gap (missing values) will have been complemented. Therefore, the process described up to this point is brought to bear on the corrected data ranging from the starting point (including the first "completed" gap) to the beginning of the second gap. The same is done with all the corrected data from the initial point to the start of the third gap (including the "completed" first and second gaps) and so on.

Finally, DFT is applied to the resulting time series and the 10 % of the most important harmonics are used. Going from the frequency domain back to the time domain is once again done with the IFFT process. At this stage, new values imputing either outliers or missing values have the same, or almost the same, shape as the original time series, granting the "macro" vision of the time series coherence. To follow the intricacies of this process, readers can consult the flow diagram shown in Figure 2.

However, as would be expected, the proposed procedure was tested to determine its validity. To this end, continuous subsets of absorbance time series with no outliers and no missing values were intentionally removed from the original time series. In order to assess the accuracy of the method proposed, the mean absolute percentage of error (MAPE) was used as shown in Equation 5, where Val real is the original time series value and Val rep is the replaced time series value.

$$MAPE = \frac{\sum_{i=1}^{n} \frac{|Val_{i-real} - Val_{i-rep}|}{Val_{i-real}}}{n} \times 100\,\% \tag{5}$$

3. Results and Discussion

For all the study sites, the Winsorising process, coded with R statistical language [51], was applied using a value of ten (10) for r and m parameters to build the window size and find the minimum and maximum acceptable values. Several value combinations of r and m parameters were verified, and the best results using r and m proved to be a value of ten (10). Although

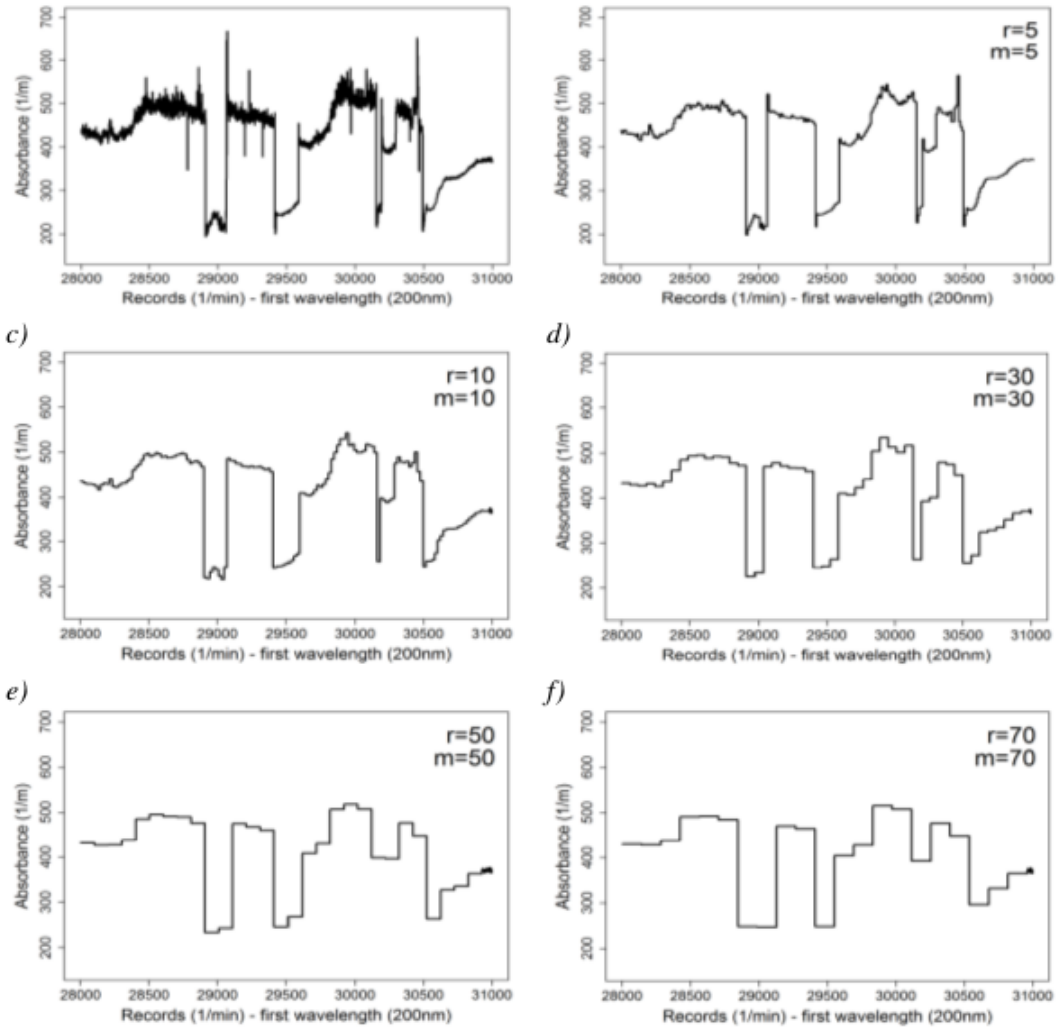

Figure 3. Comparison of different r and m Winsorising parameter values for the GPS Absorbance time series (section spanning values 28000 to 31000) with original (a), $r = m = 5$ selected value (b), $r = m = 10$ selected value (c), $r = m = 30$ selected value (d) and $r = m = 50$ selected value (e), $r = m = 70$ selected value (f) (self-authorship)

using a bigger window size with 20, 30, 50 or 100 as values for the m parameter, in addition to a lower r value compared with m, effectively eliminates outliers, but the shape of the resulting time series is reminiscent of stair steps (thereby losing its resemblance to the original shape of the time series—Figure 3).

Figure 3 shows a select section of the GPS absorbance time series, which includes values from 28000 to 31000 for the time series. Five different combinations for r and m parameter values were used, which led to the observation of a time series that mirrors the original time series analysed.

Figure 3 demonstrates that although large window size values (30, 50 or 70) for r and m parameters translate into less outlier values, the time series' shape becomes stair-like, losing its similarity to the shape of the original times series (Figure 3, graph a). Thus, using these large window size values would be disadvantageous because an analysis of the last part of the time series (roughly values 30950 until 31000) becomes largely impossible, as observed in Figure 3 (graphs d, e and f).

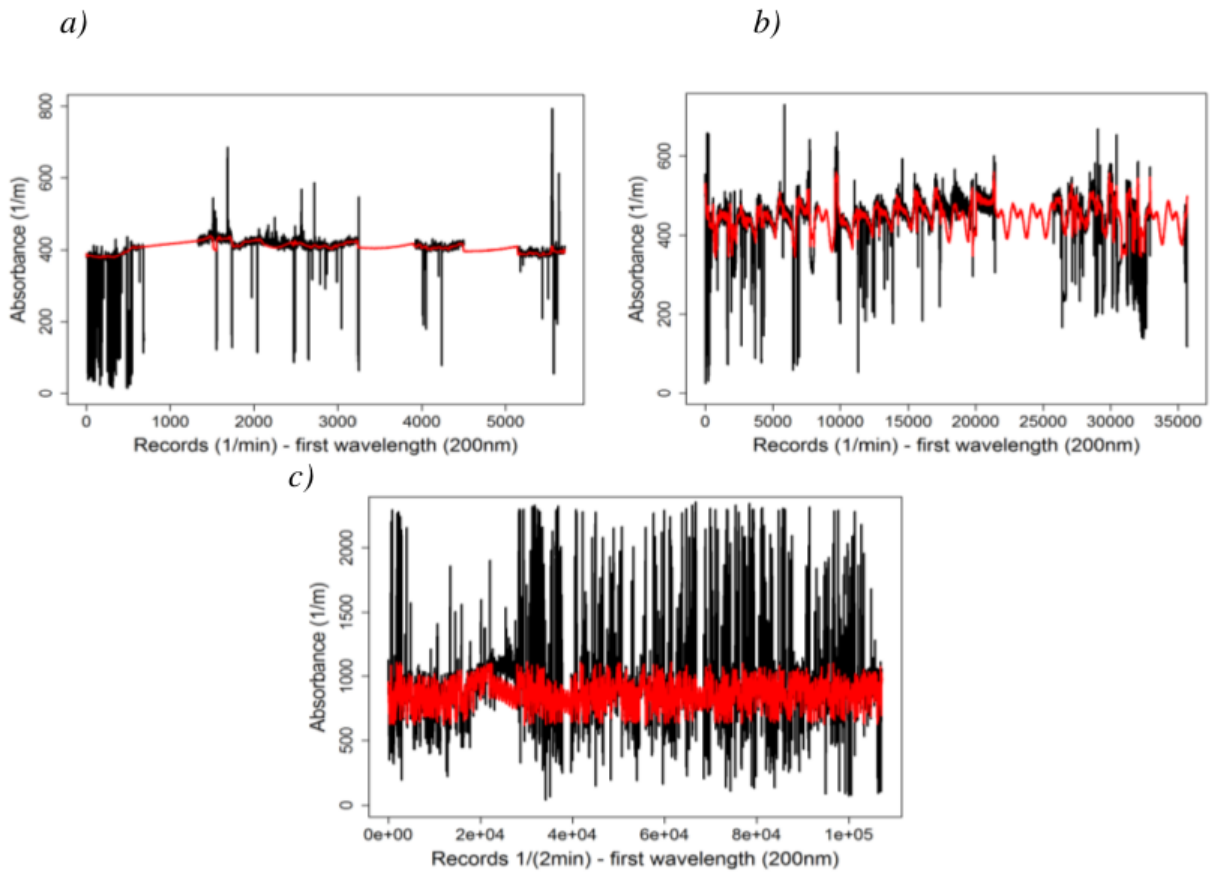

Figure 4. Resulting time series for all three study sites after applying the proposed procedure to Salitre-WWTP (a), GPS (b) and San Fernando- WWTP (c) (self-authorship)

However, Figure 3 (graph b) shows the results with some outliers. Therefore, 10 was selected as the optimal value for r and m parameters to strike a balance between the analysed window size and the number of outlier values requiring removal. However, given the fact that some outlier values still remain after applying the Winsorising process, the process of applying one and two standard deviations and DFT is repeated to expunge these outliers from the dataset.

As a result, after the final parameter values are selected for window size using the Winsorising process, the last part of the processes indicated in the flow diagram are undertaken (Figure 2). The time series for all three study sites after applying this complete process can be seen in Figure 4.

Figure 4's black curve represents the original times series (absorbance) and its red curve the resulting time series. The results obtained are as follows:

a) Salitre-WWTP: three large gaps of missing values. Although DFT could not be used for imputing missing values in the first gap, it was successfully used for the other two large gaps. Instead of DFT for the first gap, a linear interpolation was used.

b) GPS: three large gaps of missing values imputed with DFT.

c) San Fernando-WWTP: longest absorbance time series and greatest presence of outliers (of the three study sites). The entire filter procedure (Winsorising, one and two standard deviations,

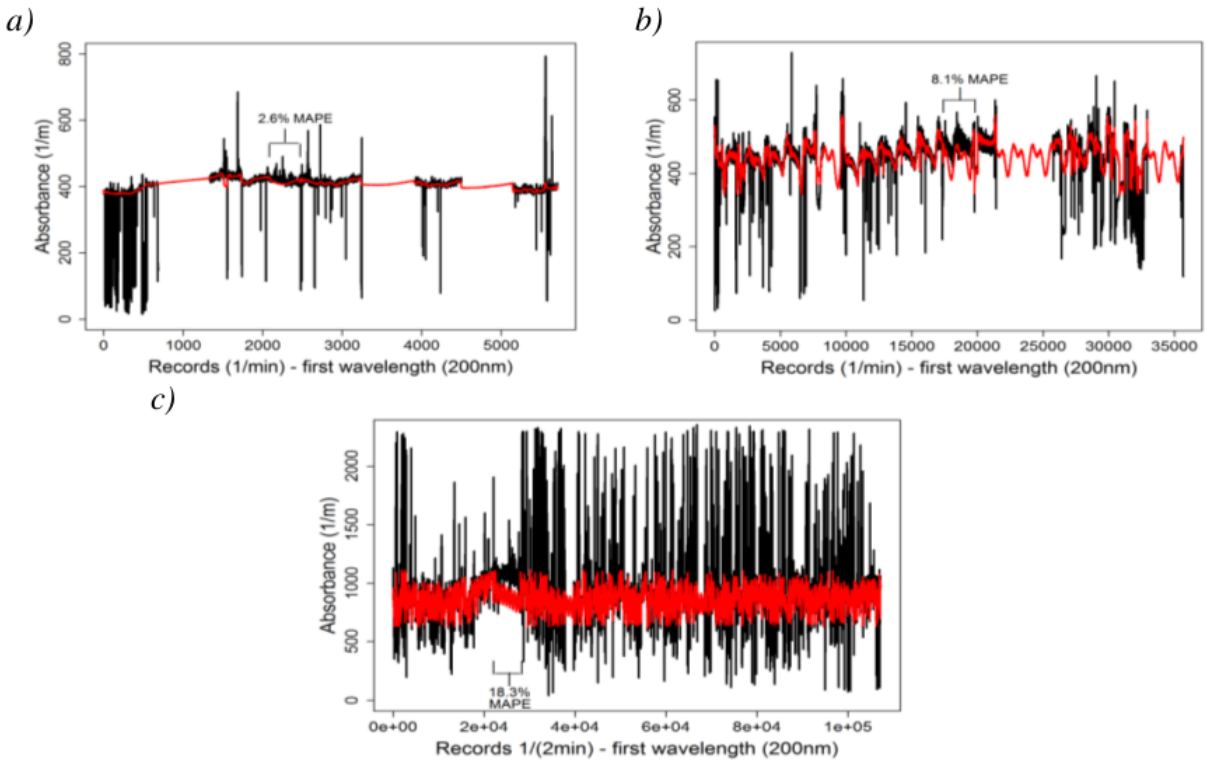

Figure 5. Resulting time series and MAPE values for all three study sites after removal a selecting range values and applying the proposed procedure to Salitre-WWTP (a), GPS (b) and San Fernando-WWTP (c) (self-authorship)

DFT, IFFT) detected and removed outliers, with DFT reapplied for imputing missing values (see Figure 4, graph c).

To assess the validity of the results presented by the proposed methodology, continuous subsets (a section) of the absorbance time series without outlier or missing values were removed from the original time series. Similarly, the MAPE (see Equation 5) was used, as shown in Figure 5. Thus, DFT is applied to the resulting time series and according to preliminary test varying from 5 % to 20 % of the most important harmonics (see the appendix). Therefore, it was decided to use the 10 % of the most important harmonics as a sufficient number of harmonics in order to capture de variability of the time series.

Figure 5 shows the time series (black curve) and the modified-filtered time series (red curve) for all three study sites. Below are the results.

a) Salitre-WWTP: values 2100 to 2449 were removed, which gives a 2,6 % MAPE.

b) GPS: values 17500 to 19499 were removed, which gives an 8,1 % MAPE.

c) San Fernando-WWTP: values 22500 to 27999 were removed, which gives an 18,3 % MAPE.

As the above paragraphs a-c witness, different combinations of values were intentionally eliminated from the original time series (absorbance) at each of the three study sites, in the line with the aforementioned criteria for value exclusion. Overall, a 12 % average of MAPE was obtained.

4. Conclusions

The main contribution of this work is the proposal of a new methodology in order to complete time series with missing values and different characteristics (absorbance), and consists of Winsorising as a step in outlier removal. Also, the application of the DFT and the IFFT, using the 10 % most important harmonics of useful values, which is crucial for its later use in different applications, specifically for time series of water quality and quantity in urban sewer systems. The imputing of missing values process (DFT and IFFT) can be applied to any missing value gap with different length. It was applied the combination of the Winsorising and DFT procedures to perform the task of detection, removal of outliers and imputing of missing values to maintain the observed periodic behaviour of these time series.

The process laid out in this paper consists of Winsorising as a first step in outlier removal. One and two standard deviations were then removed (with the concomitant applications of DFT and IFFT) before repeating DFT and IFFT. As a whole, this process effectively removes all outliers and fills the gaps in the time series.

Likewise, the proposed process relies on small r and m parameter values (where $r = m$) for window sizes. The three different time series (absorbance) to which the process was applied each exhibited different behaviours and had different sizes. Nevertheless, good results were obtained in terms of what researchers would expect the time series to look like even though there is no dataset with which to compare the time series. Additionally, these results bode well for future applicability in that they offer themselves for application in correcting other hydrologic time series.

DFT allows for the completion of the time series based on the fact that it includes various gap sizes, removes outliers and imputing of missing values. DFT meant lower error percentages at all three study sites, with an error average of 12 %. This reflects what would have likely been the shape or pattern of the time series behaviour had there not been outliers or missing values in the first place.

This work is the starting point to continue with the spectral density estimation process by means of modified rectangular window averaging periodograms with a 50 % overlapping to estimate the periodic behaviour that could be hidden. Also, apply the proposed methodology to different time series with different behaviour and different length as rainfall information, pH, conductivity, temperature, etc. On the other hand, it is necessary to compare the performance of the DFT with other techniques, as machine learning technics (ANN, SVM, AG, etc.) to detect, remove outlier values and impute missing values for time series.

5. Acknowledgement

Authors acknowledge Bogotá Water and Sewage Company (Empresa de Acueducto y Alcantarillado de Bogotá – EAAB, under the Administrative Contract No. 9-07-25100-0763-2010) and Medellín Water and Sewage Company (Empresas Públicas de Medellín – EPM) for providing the information used in this research. Also, thank the reviewers and editors for suggestions and observations that have improved the quality of the manuscript.

Referencias

[1] Langergraber, G., Fleischmann, N., ofstaedter, F. and Weingartner A., "Monitoring of a paper mill waste water treatment plant using UV/VIS spectroscopy". *IWA Water Science and Technology*, 49(1), 2004, pp. 9-14. ↑112, 113

[2] Youquan, Z., Yuchun, L., Yang, Z. and Yanjun, F., "A Novel Monitoring System for COD Using Optical Ultraviolet Absorption Method". *Procedia Environmental Sciences*, 10, 2011, pp. 2348-2353. ↑112

[3] Storey, M., van der Gaag, B. and Burns, B., "Advances in on-line drinking water quality monitoring and early warning systems". *Water Research*, 45, 2011, pp. 741-747. ↑112

[4] Sempere-Payá, V. and Santonja-Climent, S., "Integrated sensor and management system for urban waste water networks and prevention of critical situations". *Computers, Environment and Urban Systems*, 36, 2012, pp. 65-80. ↑112

[5] Xu, Z., Liu, B., Dong, Q., Lei, Y., Li, Y., Ren, J., and McCutcheon, J., "Flat microliter membrane-based microbial fuel cell as "on-line sticker sensor" for self-supported in situ monitoring of wastewater shocks". *Bioresource Technology*, 197, 2015, pp. 244-251. ↑112

[6] Bowerman, B., O'Conell, R. and Koehler, A., *Forecasting, Time Series, and Regression: An Applied Approach*. Fourth Edition. Thomson Learning. USA 2006. ↑112

[7] Gujarati, D. and Porter, D., *Basic Econometrics*. Fifth Edition. McGraw-Hill Higher Education/Irwin New York-USA. 2008. ↑112

[8] Lind, D., Marchal, W. and Wathen, S., *Statistical techniques in business & economics*. Fifteenth Edition. McGraw-Hill/Irwin. New York-USA 2012. ↑112

[9] Drolc, A. and Vrtovšek, J., "Nitrate and nitrite nitrogen determination in waste water using on-line UV spectrometric method". *Bioresource Technology*, 101, 2010, pp. 4228-4233. ↑112

[10] Al-Monami, F. and Örmeci, B., "Measurement of polyacrylamide polymers in water and wastewater using an in-line UV–vis spectrophotometer". *Journal of Environmental Chemical Engineering*, 2, 2014, pp. 765-772. ↑112

[11] Bollmann, U., Vollertsen, J., Carmelier, J. and Bester, K., "Dynamics of biocide emissions from buildings in a suburban stormwater catchment – Concentrations, mass loads and emission processes". *Water Research*, 56, 2014, pp. 66-76. ↑112

[12] Altmann, J., Massa, L., Sperlich, A. and Gnirss, R., "UV254 absorbance as real-time monitoring and control parameter for micropollutant removal in advanced wastewater treatment with powdered activated carbon". *Water Research*, 94, 2016, pp. 240-245. ↑112

[13] Murla, D., Gutierrez, O, Martinez, N., Suñer, D., Malgrat, P. and Poch, M., "Coordinated management of combined sewer overflows by means of environmental decision support systems". *Science of the Total Environment*, 550, 2016, pp. 256-264. ↑112

[14] Lacour, C., Joannis, C. and Chebbo, G., "Assessment of annual pollutant loads in combined sewers from continuous turbidity measurements: Sensitivity to calibration data". *Water Research*, 43, 2009, pp. 2179-2190. ↑112

[15] Becouze-Lareure, C. Thiebaud, l. Bazin, C., Namour, P., Breil, P. and Perrodin, Y., "Dynamics of toxicity within different compartments of a peri-urban river subject to combined sewer overflow discharges". *Science of the Total Environment*, 539, 2016, pp. 503-514. ↑112

[16] Gasperi, J. Gromaire, M. Kafi, M. Moilleron, R. and Chebbo, G., "Contributions of wastewater, runoff and sewer deposit erosion to wet weather pollutant loads in combined sewer systems". *Water Research*, 44, 2010, pp. 5875-5886. ↑112

[17] Métadier, M., & Bertrand-Krajewski, J.-L., "Assessing dry weather flow contribution in TSS and COD storm events loads in combined sewer systems". *Water Science and Technology*, 63(12), 2011, pp. 2983-2991. ↑112

[18] Bi, E., Monette, F. and Gasperi, J., "Analysis of the influence of rainfall variables on urban effluents concentrations and fluxes in wet weather". *Journal of Hydrology*, 523, 2015, pp. 320-332. ↑112

[19] Saagi, R., Flores-Alsina, X., Fu, G., Butler, D. and Gernaey, K., "Catchment & sewer network simulation model to benchmark control strategies within urban wastewater systems". *Environmental Modelling & Software*, 78, 2016, pp. 16-30. ↑112

[20] Johnson, R. and Wichern, D., *Applied Multivariate Statistical Analysis*. 6th ed. Pearson Prentice Hall, USA, 2007. ↑112

[21] Díaz, C., García, P., Alonso, J., Torres, J. and Taboada, J., "Detection of outliers in water quality monitoring samples using functional data analysis in San Esteban estuary (Northern Spain)". *Science of the Total Environment*, 439, 2012, pp. 54–61. ↑112, 113

[22] Cucina, D., di Salvatore, A. and Protopapas, M., "Outliers detection in multivariate time series using genetic algorithms". *Chemometrics and Intelligent Laboratory Systems*, 132, 2014, pp. 103-110. ↑112, 113
[23] Gharibnezhad, F., Mujica, L. and Rodellar, J., "Applying robust variant of Principal Component Analysis as a damage detector in the presence of outliers". *Mechanical Systems and Signal Processing*, 50-51, 2015, pp. 467-479. ↑112, 113
[24] Grané, A. and Veiga, H. "Wavelet-based detection of outliers in financial time series". *Computational Statistics and Data Analysis*, 54, 2010, pp. 2580-2593. ↑113
[25] Piñeiro, J., Martínez, J. García, P., Alonso, J., Díaz, C. and Taboada, J., "Analysis and detection of outliers in water quality parameters from different automated monitoring stations in the Miño river basin (NWSpain)". *Ecological Engineering*, 60, 2013, pp. 60-66. ↑112, 113
[26] Gumedze, F. and Chatora, T., "Detection of outliers in longitudinal count data via overdispersion". *Computational Statistics and Data Analysis*, 79, 2014, pp. 192-202. ↑113
[27] Sangeux, M. and Polak, J., "A simple method to choose the most representative stride and detect outliers". *Gait & Posture*, 41, 2015, pp. 726-730. ↑113
[28] Qi, M., Fu, Z. and Chen, F., "Outliers detection method of multiple measuring points of parameters in power plant units". *Applied Thermal Engineering*, 85, 2015, pp. 297-303. ↑112, 113
[29] Martínez, J., Saavedra, Á. García-Nieto, P., Piñero, J. Iglesias, C., Taboada, J., Sancho, J. and Pastor, J., "Air quality parameters outliers detection using functional data analysis in the Langreo urban area (Northern Spain)". *Applied Mathematics and Computation*, 241, 2014, pp. 1-10. ↑113
[30] Maciá-pérez, F., Berna-Martinez, J., Fernández, A. and Abreu, M., "Algorithm for the detection of outliers based on the theory of rough sets". *Decision Support Systems*, 75, 2015, pp. 63-75. ↑113
[31] Song, X., Liu, Z., Yang, J. and Qi, Y., "Extended semi-supervised fuzzy learning method for nonlinear outliers via pattern discovery". *Applied Soft Computing*, 29, 2015, pp. 245-255. ↑113
[32] Dumedah, G., Jeffrey, P. and Li, W., "Assessing artificial neural networks and statistical methods for infilling missing soil moisture records". *Journal of Hydrology*, 515, 2014, pp. 330-344. ↑113
[33] Figueroa, J., Kalenatic, D. and Lopez, C., "Missing data imputation in multivariate data by evolutionary algorithms". *Computers in Human Behavior*, 27(5), 2011, pp. 1468-1474. ↑113
[34] Figueroa-García, J., Kalenatic, D. and Lopez, C., "Incomplete Time Series: Imputation through Genetic Algorithms". *Time Series Analysis, Modeling and Applications*, 47, 2013, pp. 31-52. ↑113
[35] de França, F., Coelho, G. and von Zuben, F., "Predicting missing values with biclustering: A coherence-based approach". *Pattern Recognition*, 46, 2013, pp. 1255-1266. ↑113
[36] Folch-Fortuny, A., Arteaga, F. and Ferrer, A., "PCA model building with missing data: New proposals and a comparative study". *Chemometrics and Intelligent Laboratory Systems*, 146, 2015, pp. 77-88. ↑113
[37] Carvajal, C., Bayona, D. and Ortiz, Z., "Taxonomy extension and missing-values treatment over an informatics-security incident repository". *Ingeniería*, 18(1), 2013, pp. 24-49. ↑113
[38] Dumedah, G. and Coulibaly, P., "Evaluation of statistical methods for infilling missing values in high-resolution soil moisture data". *Journal of Hydrology*, 400, 2011, pp. 95-102. ↑113
[39] Haworth, J. and Cheng, T., "Non-parametric regression for space–time forecasting under missing data". *Computers, Environment and Urban Systems*, 36, 2012, pp. 538-550. ↑113
[40] Junger, W. and Ponce, A., "Imputation of missing data in time series for air pollutants". *Atmospheric Environment*, 102, 2015, pp. 96-104. ↑113
[41] Avellaneda, J., Ochoa, C., and Figueroa-García, J., Çomparison between a self organizing neural fuzzy system and an ARIMAX model to forecasting volatile economic series". *Ingeniería*, 17(2), 2012, pp. 26-34. ↑113
[42] Tukey, J., *Exploratory data analysis*. Addison-Wesely. ↑113
[43] Acuña, E. and Rodríguez, C., *On Detection of Outliers and Their Effect in Supervised Classification*. Department of Mathematics University of Puerto Rico at Mayaguez, Mayaguez, Puerto Rico. 2013. [Online]. Available http://academic.uprm.edu/ eacuna/vene31.pdf ↑113
[44] s::can, *Manual ana::pro Version 5.3 September 2006 Release*. Messtechnik GmbH, Vienna, Austria, 2006. ↑113
[45] Liu, H., Shah, S. and Jiang, W., "On-line outlier detection and data cleaning". *Computers and Chemical Engineering*, 28, 2004, pp. 1635-1647. ↑113
[46] Ko, S-J and Lee, Y., "Theoretical analysis of winsorizing smoothers and their applications to image processing". *Acoustics, Speech, and Signal Processing, ICASSP-1991*, pp. 3001-3004. ↑114, 115
[47] Pearson, R., "Outliers in process modelling and identification". *IEEE Transactions on Control Systems Technology*, 10, 2002, pp. 55-63. ↑114

[48] Kontaki, M., Gounaris, A., Papadopoulos, A., Tsichlas, K. and Manolopoulos, Y., "Efficient and flexible algorithms for monitoring distance based outliers over data streams". *Information Systems*, 55, 2015, pp. 37–53. ↑ 114, 115

[49] Proakis, J. and Manolakis, D., *Digital signal processing principles, algorithms, and applications*. Fourth Ed. New Jersey: Pearson Prentice Hall, 2007. ↑115

[50] Plazas-Nossa, L. and Torres, A., "Fourier analysis as a forecasting tool for absorbance time series received by UV-Vis probes installed on urban sewer systems". *Proceedings of 8th International Conference Novatech*, 2013, Lyon, France, 23-27 June 2013. ↑115

[51] R Core Team, "R: A language and environment for statistical computing". *R Foundation for Statistical Computing*, Vienna, Austria, 2014. [Online]. Available URL http://www.R-project.org/ ↑116

Leonardo Plazas Nossa

Ingeniero Electrónico, Magister en Teleinformática, Doctorado en Ingeniería, Universidad Distrital Francisco José de Caldas, Facultad de Ingeniería, Carrera 7 No. 40-53, 3239300 Ext. 2405. Bogotá, Colombia.
Contacto: lplazasn@udistrital.edu.co.

Miguel Antonio Ávila Angulo

Ingeniero Catastral y Geodesta, Magister en Teleinformática, Universidad Distrital Francisco José de Caldas, Facultad de Ingeniería, Carrera 7 No. 40-53, 3239300 Ext. 2405. Bogotá, Colombia.
Contacto: maavila@udistrital.edu.co

Andrés Torres

Ingeniero Civil, Especialización en Sistemas Gerenciales de Ingeniería, Maestría en Ingeniería Civil, Doctorado en Ingeniería Civil, Grupo de Investigación Ciencia e Ingeniería del Agua y el Ambiente, Facultad de Ingeniería, Pontificia Universidad Javeriana, Carrera 7 No. 40 -62, 3208320 Ext. 5553. Bogotá, Colombia.
Contacto: andres.torres@javeriana.edu.co

Appendix

The DFT is applied to the resulting time series, several test were done varying from 5 % to 20 % of the most important harmonics. As an example, Figure 6 shows the results obtained for San Fernando-WWTP applied to the values between 22500 and 27999, which were removed the same process was done for Salitre-WWTP and GPS. Table I shows the MAPE values for 5 %, 10 % and 20 % of the most important harmonics results respectively for all three study sites. Therefore, it was decided to use the 10 % of the most important harmonics as a sufficient number of harmonics in order to capture de variability of the time series.

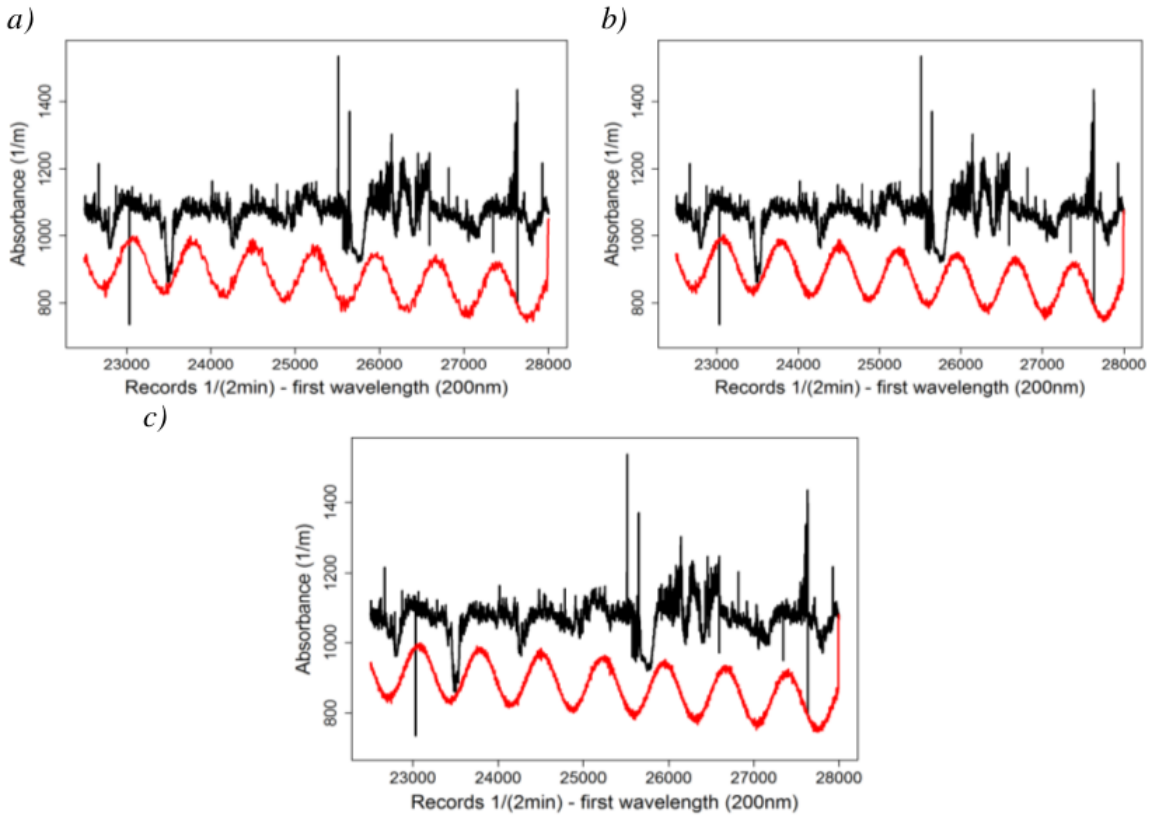

Figure 6. Resulting time series and MAPE values for San Fernando WWTP study site after removal a selecting range values and applying the proposed procedure for 5 % (a), 10 % (b) and 20 % (c) of most relevant harmonics respectively (self-authorship)

Table I. MAPE values for 5 %, 10 % and 20 % of the most important harmonics results respectively for all three study sites

Study	MAPE Value		
	5 %	10 %	20 %
Salitre-WWTP	2.59044	2.59056	2.58611
GPS	8.06975	8.07084	8.07012
San Fernando-WWTP	18.22815	18.23058	18.23014