IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 22, 2020, accepted July 19, 2020, date of publication August 7, 2020, date of current version August 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3015038

Robust Analysis of the Information Obtained From a Set of 12 Years of SO_2 Concentration Measurements

WILMAR HERNANDEZ[1], (Senior Member, IEEE), ALFREDO MENDEZ[2], VICENTE GONZÁLEZ-POSADAS[3], AND JOSÉ LUIS JIMÉNEZ-MARTÍN[3]

[1]Facultad de Ingenieria y Ciencias Aplicadas, Universidad de Las Americas, Quito 170504, Ecuador
[2]Departamento de Matematica Aplicada a las Tecnologias de la Informacion y Comunicaciones, ETSIS de Telecomunicacion, Universidad Politecnica de Madrid, 28031 Madrid, Spain
[3]Departamento de Teoria de la Señal y Comunicaciones, ETSIS de Telecomunicacion, Universidad Politecnica de Madrid, 28031 Madrid, Spain

Corresponding author: Wilmar Hernandez (wilmar.hernandez@udla.edu.ec)

This work was supported in part by the Corporación Ecuatoriana para el Desarrollo de la Investigación y la Academia (CEDIA)-Ecuador under Project CEPRA XII-2018-13, in part by the Universidad de Las Americas, Quito, Ecuador, under Project ERa.ERI.WHP.18.01, and in part by the Universidad Politecnica de Madrid, Spain.

ABSTRACT In this paper, a robust analysis of SO_2 concentration measurements taken at the Belisario air quality monitoring station, Quito, Ecuador is carried out. The analyzed data contain 12 years of measurements, from January 2008 to December 2019. In addition, this set of measurements was decomposed into variables that represent each year, month, day of the week, and hour of the day in groups of two hours. For the analysis, classic, nonparametric and robust statistical methods were used, and the data were classified based on criteria established by the Quiteño Air Quality Index, taking confidence intervals into account. The results showed that the level of air pollution at the Belisario station due to the SO_2 concentration is acceptable. In addition, the trend in the level of SO_2 concentration decreased over the years studied, with a sharp drop from 2008 to 2012, then a small rise in 2013 and another fall until 2019, presenting decreasing oscillations that tend toward a desirable level of pollution. In this paper, it was shown that the air pollution at the Belisario station due to the concentration of SO_2 in the last 12 years is not harmful to humans, with the measurement precision provided by robust statistical methods. Therefore, it can be concluded that the measures that have been taken by the Quito city council over the last few years are yielding good results.

INDEX TERMS Central tendency estimation, L-estimators, M-estimators, nonparametric confidence interval, robust confidence interval, sulfur dioxide, Wilcoxon rank sum test.

I. INTRODUCTION

Sulfur dioxide (SO_2) is an invisible toxic gas that is dangerous to human health when inhaled [1]. According to [1], sulfate particles, sulfurous acid and sulfuric acid are harmful compounds created by the reaction of SO_2 with other substances. Additionally, approximately ten minutes is enough time to feel the effects of being exposed to SO_2, and the people most at risk are those with asthma and other illnesses associated with breathing difficulties [1].

According to [2], the burning of fossil fuels emits SO_2, and power plants and motor vehicles are among the sources that produce the SO_2 in the air. In addition, volcanic activity produces SO_2. Furthermore, in accordance with the air quality guidelines of the World Health Organization [3], the SO_2 concentration should not exceed the following levels: $500\mu g/m^3$ over a 10-minute period (short-term exposure) and $20\mu g/m^3$ over a 24-hour period (long-term exposure). In [4], it is noted that exposure to SO_2 for short periods of time obstructs breathing. Additionally, the United States Environmental Protection Agency (EPA) [4] says that SO_2 in the air leads to the formation of sulfur oxides that, by reacting with other compounds, contribute to the formation of particles that reduce visibility and cause health problems. Moreover, SO_2 is among the substances responsible for acid rain [4].

These observations show the need to carry out an analysis and interpretation of the information obtained from SO_2 measurement systems. Several relevant works have appeared in the scientific literature.

In [5], an RGB-based oximetry for the retina was proposed, a measuring system for a rat fundus was built, and SO_2 was

The associate editor coordinating the review of this manuscript and approving it for publication was Nikhil Padhi.

estimated in the rat fundus from the RGB image of the fundus. Additionally, in [6], the strengths and limitations of a linear fit SO_2 (LFSO2) algorithm were assessed. The LFSO2 algorithm was described, and the operational LFSO2 retrievals were compared with the principal component analysis (PCA) retrievals. The measurements were made by the Suomi NPP Ozone Mapping and Profiler Suite [7], and it was shown that the LFSO2 algorithm presented in [6] was in good agreement with the PCA algorithm used by the National Aeronautics and Space Administration (NASA) of the United States of America.

In [8], a land-use regression model was developed for SO_2 concentrations. To produce valid models, taking into consideration the spatial structure of the ordinary least-squares regression model error terms, a spatial error model was presented in [8]. The data used for the analysis performed in [8] correspond to mobile monitoring data that were collected from 2005 to 2010 in Hamilton, Ontario, Canada.

In [9], a method of estimating exposure to SO_2 was presented. The Gaussian plume atmospheric transport model used in [9] was introduced by [10] and [11], and it obtained the average ground-level concentration of SO_2 over long periods of time. In [9], uncertainty was incorporated into the abovementioned model by multiplying by a lognormally distributed factor. Additionally, the geometric bias expressed in terms of the predicted-to-observed ratio was used to compare the predicted and observed SO_2 concentrations.

In [12], a study of the influence of processing phosphorite in an industrial sector in the Republic of Kazakhstan was performed. In [12], the air pollution due to the release of toxic substances associated with the processing of phosphate rock was analyzed. To that end, the authors of [12] used correlation and regression methods to obtain linear equations that described the relationship between the volume of phosphorus produced and the amount of SO_2 and phosphorus pentoxide (P_2O_5) emissions in the region under study.

Two stochastic models for air pollution due to SO_2 in the meteorological season of winter in Vienna were presented in [13]. Specifically, regression was used to both explain the influence of meteorological factors on the SO_2 concentration in the region under study and give an approximate measure of the SO_2 concentration in the city center, which was the most polluted area due to industrial sources in the vicinity, that is, in the southeastern part of the area. The stochastic modeling performed in [13] was carried out by using autoregressive models with exogenous inputs [14], [15].

The temporal and spatial dynamics of SO_2 concentration in the Beijing-Tianjin-Hebei (Jing-Jin-Ji) region of China from 2007 to 2016 were studied in [16]. Additionally, the temporal and spatial distributions of SO_2 were shown along with annual changes and trends, and a correlation analysis was used to explain the relationship between the SO_2 concentration and SO_2 emissions in the region.

A supervisory control and data acquisition (SCADA) system for monitoring SO_2, particulate matter ($PM_{2.5}$), carbon monoxide (CO), carbon dioxide (CO_2), air temperature, and relative humidity (RH) was presented in [17]. This SCADA system was based on smart technologies, and the SO_2 concentration was estimated by using an autoregressive moving average model [14], [15]. Furthermore, in [17], an Android mobile application was designed to allow users to obtain the estimated values of some air pollutants.

Moreover, an estimation of anthropogenic SO_2 emissions from 1850 to 2005 at both the global and country levels is presented in [18], and to carry out this estimation, an uncertainty analysis was performed considering both random and systematic uncertainties, with 95% confidence intervals for both types of uncertainties.

In [19], a geochemical and statistical analysis of SO_2, heavy metal pollutants, and total organic carbon was carried out. The data for the analysis came from soil samples taken from major oil refineries, industries, and residential areas in the Al-Ahmadi governorate, State of Kuwait. In [19], the minimum, maximum, mean, and standard deviation were used to describe the data. Additionally, variance analysis was used to test whether there were significant differences among representative soil samples taken from three major oil refineries (the Mina Al-Ahmadi Refinery, Mina Abdullah Refinery, and Al-Shuaiba Refinery), the metals emitted by the industry and anthropogenic sources, and the interactions among them.

In the research presented in this paper, 12 years of measurements of SO_2 concentrations taken at the Belisario air-quality monitoring station (Quito, Ecuador) [20] are analyzed by using robust statistics techniques [21]–[23]. Belisario station is an important part of the metropolitan atmospheric monitoring network of Quito [24]. The data used in this paper correspond to a set of measurements carried out from January 2008 to December 2019. Here, each year was considered a variable, and the distributions of these variables were heavy tailed [22], [25]. Additionally, robust estimations of the central tendency and scale were carried out, and both nonparametric confidence intervals [26], [27] and robust confidence intervals [21]–[23] were calculated. Some previous research aiming at using nonparametric and robust tools to analyze measurements taken by particulate matter ($PM_{2.5}$) sensors are described in [28]–[31].

In addition, other works focusing on the statistical analysis of measurements of air pollution variables that support the above work and complement the examples given in previous paragraphs include the following:

In [32], to process high-dimensional data with the aim of predicting the $PM_{2.5}$ concentration at 35 air quality monitoring stations in Beijing, China over the subsequent 24 hours, a LightGBM model [33] was proposed. Additionally, to capture the trend of the $PM_{2.5}$ concentration in a time series and reduce its dimensions, correlation analysis and PCA were used. In [34], a methodological framework combining a bidirectional long short-term memory network, a deep learning network, and the inverse distance weighting technique to conduct spatiotemporal predictions of air pollutants at different time granularities was presented. Moreover, forecasting of

the $PM_{2.5}$ concentration at Guangdong, China, was carried out in [34]. Furthermore, in [35] a PM2.5 remote sensing retrieval method was presented, and to establish the relationship between moderate-resolution imaging spectroradiometer images and ground observations of $PM_{2.5}$, an ensemble random forest machine learning method was used.

The present paper has the following objectives: (1) to obtain groupings of the variables (i.e., years) under study, comparing the results of SO_2 concentration measurements by years, months, days of the week, and hours of the day; (2) to find the differences that exist between air pollution categories due to the SO_2 concentration [36]; and (3) to quantify the abovementioned differences by using confidence intervals. Here, a study of the trends in the concentration of SO_2 at Belisario station [20] over the last 12 years is exhaustively developed, and it is concluded that the concentration is downward, measuring this decrease in SO_2 concentration with the precision provided by robust statistical methods, which is part of the novelty of this research.

Furthermore, in accordance with the aims and scope of this journal, this paper can be classified as an application-oriented paper aimed at analyzing and interpreting the information from a set of SO_2 measurements using robust statistical methods. A description of the characteristics of the data is given in Section II. The aim of Section III is to perform data classification by using nonparametric techniques. The robust estimation of the central tendency of the data and scale is performed in Section IV. Moreover, Section IV is devoted to classifying the data by building robust confidence intervals. Finally, Section V presents the conclusions of the paper.

II. DESCRIPTION OF THE DATA

The air-quality monitoring station used to perform the SO_2 concentration analysis is called Belisario, and it is located in one of the most important parts of Quito [20]; the characteristics of all the air pollution variables that are currently measured at Belisario station can be found in [36]. In accordance with [36], SO_2 concentration measurements were carried out using the Thermo Scientific Model 43i SO_2 Analyzer [37]. The sampling time was 10 minutes, and each datum used for the analysis was the average of the set of samples corresponding to one hour. To cover 75% of the valid records, these averages were calculated in accordance with international criteria [36]. Furthermore, the results of the analysis carried out in this paper focus on studying the data that were collected over the period of time from January 2008 to December 2019.

In this paper, the variations recorded in the data are analyzed with the aim of verifying whether they are due to random variations or whether the samples taken from the variables under study show that these variables are different from each other. Here, the characteristics of the variables under study are established in order to capture the differences among the variables and analyze how the air pollution at the Belisario station changed due to the SO_2 concentration during the years of study. This is verified by using robust and nonparametric statistical techniques to build confidence intervals that reveal the similarities and differences in the variables considered.

Regarding the construction of the confidence intervals, in the rest of the paper, the results obtained for the analysis by years, months, days of the week, and hours will be discussed in parallel. The types of these confidence intervals are classic, nonparametric, bootstrap, and robust.

In this paper, the SO_2 concentration is given in $\mu g/m^3$, and the variables under study are as follows:

1) X_k, $k = 1, \ldots, 12$, is the SO_2 concentration in year $2007 + k$.
2) Y_k, $k = 1, \ldots, 12$, is the SO_2 concentration in each month of the year.
3) Z_k, $k = 1, \ldots, 7$, is the SO_2 concentration on each day of the week.
4) W_k, $k = 1, \ldots, 12$, is the SO_2 concentration pooled for each group of two hours. That is, W_1 stands for the SO_2 concentration at 0:00 and 1:00, W_2 stands for the SO_2 concentration at 2:00 and 3:00, and so on until W_{12}, which stands for the SO_2 concentration at 22:00 and 23:00.

To summarize the set of measurements of SO_2 concentration and provide a description of the observations, a statistical summary of the data is shown in Table 1. Table 1 shows that the mean is greater than the median for all the variables, that all the values of skewness are greater than 1.9, and that all the values of kurtosis are greater than 9. Additionally, the value of kurtosis in 2019 was 533.8117. Furthermore, a large number of the observations from all years are outliers. Specifically, between 4.71% and 8.01% of the observations are outliers. Therefore, these data do not form a Gaussian distribution. Instead, this indicates that the variables follow a heavy-tailed distribution [22], [25].

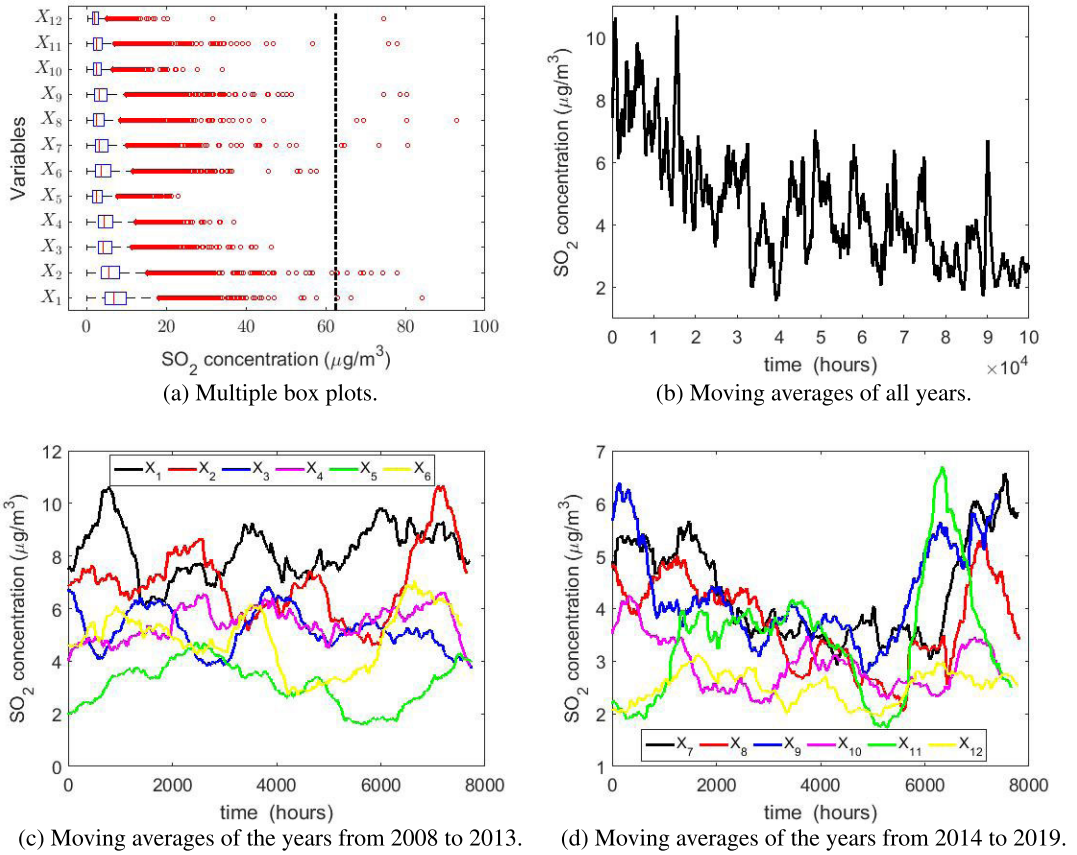
The box plot of the data is shown in Fig. 1 along with three graphs of moving averages (MAs) [14], [15], where one graph shows all the years and two other graphs show only half of the years. In the case under study, the size that was preferred to perform the MA smoothing was 720 because this is the number of samples collected in a 30-day month.

At this point, it is important to mention that the MA functions are used to detect and eliminate the trend of a series. Therefore, the application of the MA functions to the series values produces a new series with certain characteristics, which suppresses the non-systematic alterations and highlights the main aspects of the time series.

The box plot shown in Fig. 1(a) confirms that the variables follow heavy-tailed distributions. In addition, a discontinuous straight line was included in the box plot. This line was drawn to represent the separation between having a desirable level of air pollution (that is, a concentration of SO_2 in the interval $[0, 62.5\mu g/m^3)$) and an acceptable level of air pollution (that is, an SO_2 concentration in the range $[62.5\mu g/m^3, 125\mu g/m^3)$), according to the Quiteño Air Quality Index (QAQI) [36]. Therefore, according to the QAQI, the level of air pollution at the Belisario station is, in the worst case,

TABLE 1. Summary statistics of the SO_2 concentration measurements.

Year	Count	Mean	Median	Standard Deviation	Skewness	Kurtosis	Minimum	Maximum	Outliers %
2008 (X_1)	8423	8.0830	6.8700	5.3320	2.5700	17.8613	0.01	84.2600	4.89
2009 (X_2)	8372	6.9080	5.7250	5.2932	3.9706	33.5676	0	77.9900	5.28
2010 (X_3)	8471	5.1788	4.2400	3.7815	2.6538	15.6548	0	46.2400	6.54
2011 (X_4)	8432	5.2712	4.3500	3.6296	1.9293	9.0941	0	37.0600	5.24
2012 (X_5)	8452	3.1074	2.4500	2.5231	2.0748	9.6371	0.01	22.9200	5.55
2013 (X_6)	8277	4.8171	3.6600	4.1870	3.1259	21.9440	0	57.7500	6.13
2014 (X_7)	8505	4.4210	3.2900	4.0050	4.8574	55.6759	0	80.5800	6.96
2015 (X_8)	8530	3.7672	2.5900	3.8379	5.9199	85.2278	0.03	92.7700	8.01
2016 (X_9)	8126	4.3526	3.2400	4.1706	5.1628	57.5357	0	80.3000	6.74
2017 (X_{10})	8172	2.9068	2.4600	2.0238	2.8375	22.4814	0	33.9800	4.71
2018 (X_{11})	8375	3.3227	2.5300	3.4137	6.4378	85.6959	0	77.8300	7.28
2019 (X_{12})	8513	2.4893	2.1000	1.8681	15.1990	533.8117	0.23	74.5600	5.57
Total	100648	4.5530	3.3700	4.1379	3.9041	37.8520	0	92.7700	6.08



(a) Multiple box plots.
(b) Moving averages of all years.
(c) Moving averages of the years from 2008 to 2013.
(d) Moving averages of the years from 2014 to 2019.

FIGURE 1. Box plot diagram and moving averages of all years and of half of the years.

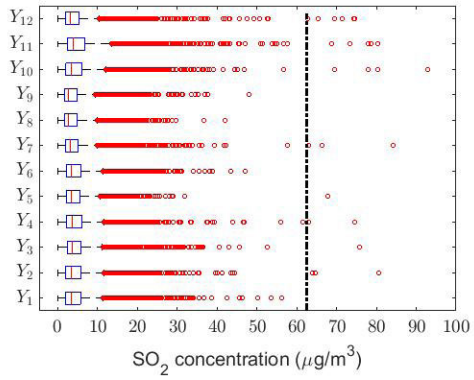
acceptable. Thus, this variable does not represent a health risk in the surroundings of the monitoring station.

Figure 1(b) indicates that the SO_2 concentration at the Belisario station decreases continuously over time. Likewise, the MA graphs of the years (see Figs. 1(c) and 1(d)) show clear oscillations, with increases at the beginning and end of each year. Furthermore, these graphs indicate that there is a decrease in the concentration of SO_2 in the third quarter of each year. However, in none of the cases is there an evident tendency.

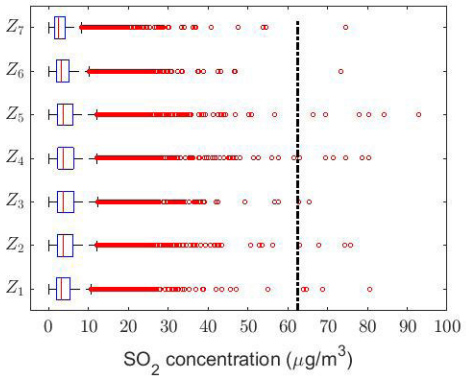
The results shown above for the analysis of the SO_2 concentration by year are also confirmed when the study is carried out for the months of the year, the days of the week, and the hours of the day.

In Fig. 2, the box plot diagrams (Fig. 2(a) - 2(c)) indicate that the variables taken into account follow heavy-tailed distributions, and the MA graphs (Fig. 2(d) - 2(f)) show a clear tendency to decrease throughout the years of study.

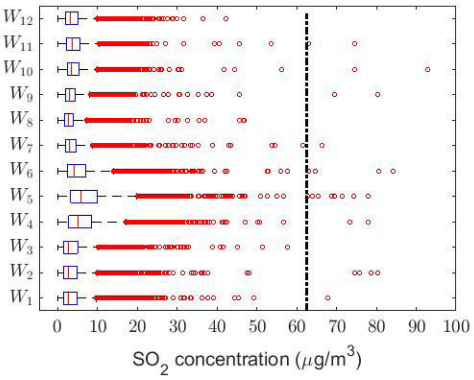
From the box plot diagram shown in Fig. 2(a), in the central months of the year, the SO_2 concentration is lower than in the



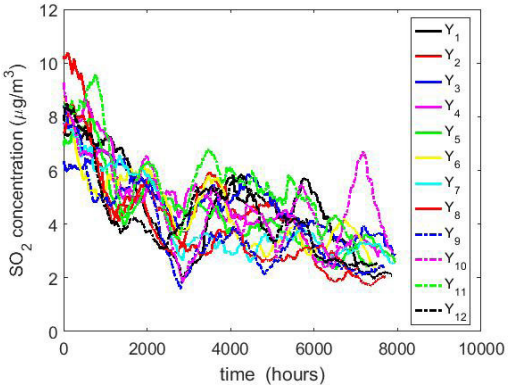
(a) Box plot: Months; Y_k, $k = 1, \ldots, 12$.



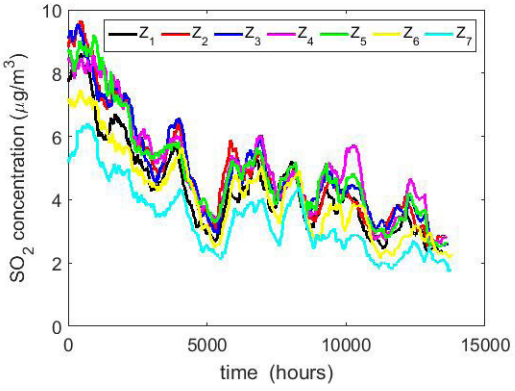
(b) Box plot: Days; Z_k, $k = 1, \ldots, 7$.



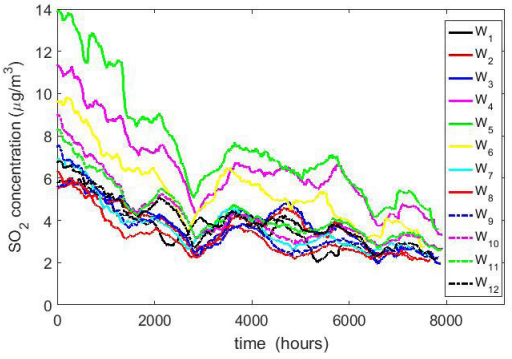
(c) Box plot: Hours of the day in groups of two hours; W_k, $k = 1, \ldots, 12$.



(d) Moving averages of months; Y_k, $k = 1, \ldots, 12$.



(e) Moving averages of days; Z_k, $k = 1, \ldots, 7$.



(f) Moving averages of the time of day in groups of two hours; W_k, $k = 1, \ldots, 12$.

FIGURE 2. Box plot diagrams and moving averages of all years for months, weeks and every two hours.

rest of the year. On weekends (see Fig. 2(b)), the SO_2 concentration also seems to decrease. In addition, observing the box plot of the hours of the day (see Fig. 2(c)), the concentration of SO_2 increases from dawn until the first hours of the day, then gradually decreases, rebounds in mid-afternoon and then decreases until the following dawn.

The distinguishing characteristics of the MA graphs shown in Fig. 2(d) - 2(f) are that the behavior of the months, weeks, and hours are similar throughout the years. That is, if they rise or fall in a certain period of a year, a similar pattern occurs in the rest of the years.

In Fig. 2(d), there is a gradual decrease in the concentration of SO_2 from January to December, but this decrease occurs with many fluctuations. Additionally, some of these fluctuations have greater amplitudes. Because the first observations in each series have been removed to apply the moving averages, the continuity between the last observations in each series and the first observations in the next cannot be seen.

TABLE 2. Confidence interval limits for the median of each variable and the confidence interval lengths, with $\alpha = 0.05$.

Variable	Lower limit	Upper limit	Length	Variable	Lower limit	Upper limit	Length
X_1	6.76	6.96	0.20	X_7	3.22	3.34	0.12
X_2	5.64	5.81	0.17	X_8	2.55	2.64	0.09
X_3	4.18	4.30	0.12	X_9	3.17	2.28	0.11
X_4	4.29	4.43	0.14	X_{10}	2.42	2.50	0.08
X_5	2.39	2.50	0.11	X_{11}	2.50	2.58	0.08
X_6	3.59	3.74	0.15	X_{12}	2.07	2.12	0.05

In the graph of the moving averages of each day of the week (see Fig. 2(e)), the behavior is very similar for all days of the week. Specifically, the series is in a band approximately $2\mu g/m^3$ wide, the trend is clearly descending from Monday to Sunday, and there are many wide-ranging oscillations compared to the values taken by the variable. This also happens in Fig. 2(d). Furthermore, in Fig. 2(e), the continuity between series cannot be perceived due to the observations that are not considered.

Unlike the graphs in Fig. 2(d) and 2(e), Fig. 2(f) shows that although the behavior of the SO_2 concentration is parallel, the SO_2 concentration is greater between 6:00 and 11:00. In addition, the SO_2 concentration behaves similarly during these hours. Moreover, the tendency for the SO_2 concentration to decrease linearly is more evident than in the two preceding graphs, although there appears to be a breakpoint in the slope in the first third of the observations.

To what has been said previously, it should be added that in Fig. 2(f), fluctuations are still observed, but in this case, the fluctuations are not very wide. There are also no abrupt changes. Moreover, because the first observations in each series were removed, the continuity between series is not noted.

In this paper, different variable changes [38] were made in order to use classical statistical inference methods. Nevertheless, the data could not be fitted to parametric distributions other than heavy-tailed distributions. The adjustments achieved, at best, had a p-value [27] of less than 0.005. This justified the use of nonparametric statistics and robust statistics in this paper. The aforementioned was also attempted for the months, days of the week and hours of the day, but at best, heavy-tailed distributions were achieved in very few variables.

III. DATA CLASSIFICATION BY USING NONPARAMETRIC STATISTICAL INFERENCE

To check whether the variables come from distributions with the same median, the Wilcoxon rank sum test [26], [27] was used. This idea was also used in [28]–[30]. Here, the null hypothesis was H_0 : Median $= M_0$, and the alternative hypothesis was H_0 : Median $\neq M_0$. For stable data, half of the observations will be less than M_0 and the rest of the observations will be greater than M_0 if H_0 is true. For the median, the confidence intervals must verify (1). Here, the significance level is $\alpha = 0.05$ and the confidence level is $(1-\alpha)$.

$$P\left(X_{(k'_{\frac{\alpha}{2}})} < M_e < X_{(k_{\frac{\alpha}{2}})}\right) = 1-\alpha \tag{1}$$

where M_e is the median, $X_{(.)}$ is the vector consisting of the sample of the specific variable being analyzed with its elements in ascending order, and $k'_{\frac{\alpha}{2}}$ and $k_{\frac{\alpha}{2}}$ are based on the binomial distribution. This is justified by the central limit theorem [38]. Additionally, $k'_{\frac{\alpha}{2}}$ and $k_{\frac{\alpha}{2}}$ are given by (2) and (3), respectively.

$$k'_{\frac{\alpha}{2}} = \frac{N}{2} + 0.5 - z_{\frac{\alpha}{2}}\sqrt{\frac{N}{4}} \tag{2}$$

$$k_{\frac{\alpha}{2}} = \frac{N}{2} + 0.5 + z_{\frac{\alpha}{2}}\sqrt{\frac{N}{4}} \tag{3}$$

where N stands for the sample length and $z_{\frac{\alpha}{2}}$ is the value of the inverse cumulative distribution function of the standard normal distribution evaluated at $(1-\alpha)/2$ [27], [38].

For a confidence level of 95%, Table 2 shows the limits and lengths of the confidence intervals, and Fig. 3 shows the graphs of the confidence intervals. Once again, there is a decreasing tendency in the amount of SO_2 across the years at the Belisario station because as the median decreases, the interval shifts to lower values. There is a very sharp decrease from 2008 to 2012, especially notable in 2012, and then a rebound in 2013. Furthermore, from that year on, the amount of SO_2 is balanced with a slight downward trend.

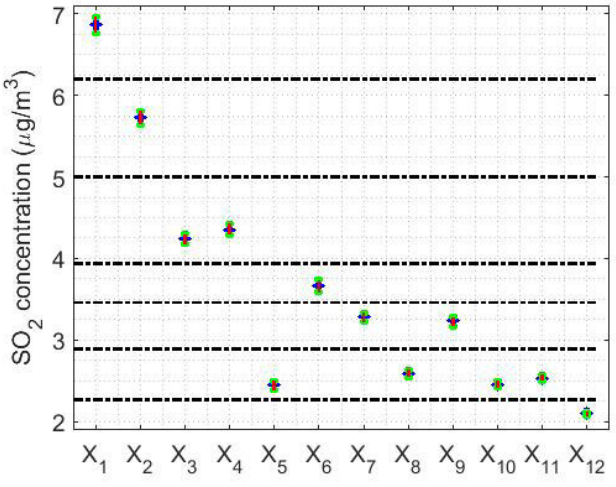
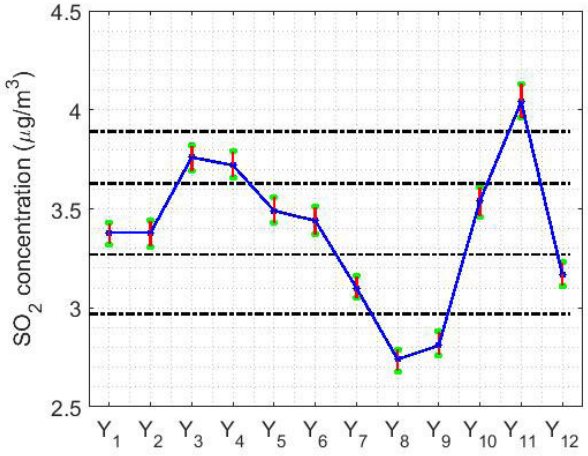
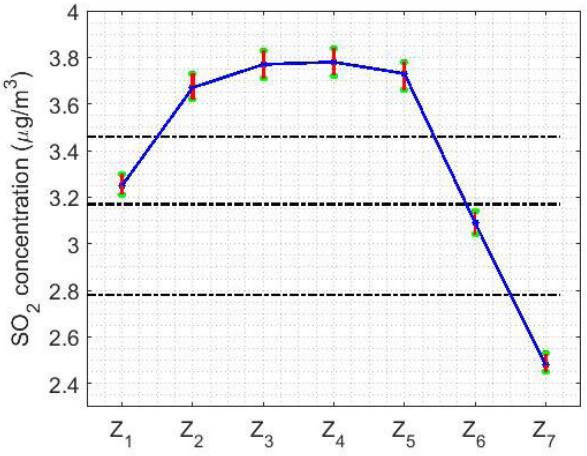


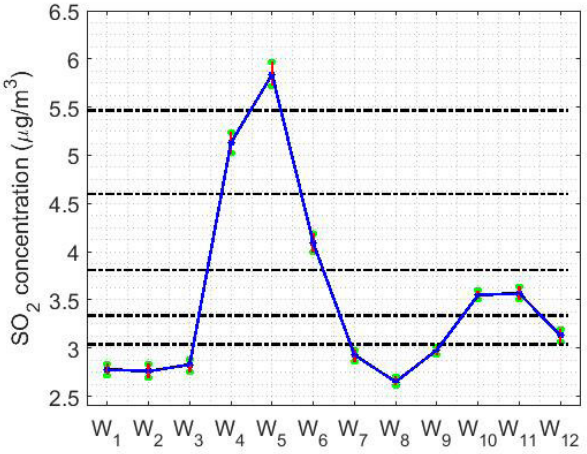
FIGURE 3. 95% confidence intervals for the median of each variable.



(a) Confidence intervals for the months.



(b) Confidence intervals for the weeks.



(c) Confidence intervals for the groups of every two hours of the day.

FIGURE 4. 95% confidence intervals for the median of each month, each week and every two hours.

Likewise, as the median value decreases, the width of the confidence interval also decreases. Note that all confidence intervals are at the desirable air quality level and that none are close to the acceptable level.

In the case under study, as shown in Fig. 3, the groupings obtained are the following seven categories: 1)

1) X_3 and X_4,
2) X_7 and X_9,
3) X_5, X_8, X_{10} and X_{11},
4) X_1,
5) X_2,
6) X_6,
7) X_{12}.

On the other hand, the groupings made with the Wilcoxon rank sum test [26], [27] coincide with those made using nonparametric confidence intervals, although with slight differences. Specifically, the grouping of variables X_5, X_8, X_{10} and X_{11} is now a set of three groups, with the variables X_5 and X_{10} grouped together and the variables X_8 and X_{11} separate.

The nonparametric confidence intervals for the months, weeks, and groups of every two hours of the day are shown in Fig. 4. For the months, five categories are established using the nonparametric intervals, which coincide with the results of the Wilcoxon rank sum test. The level of SO_2 concentration according to the months seems to vary periodically throughout the years. Low levels of SO_2 concentration occur at the end of summer, and the highest value is reached in November; in the second half of the year, there is a rebound in the values. The amplitudes of these intervals are smaller, in general, than in the analysis by year, suggesting less variability by month than by year. Similar to the analysis carried out for the years, the larger the median is, the greater the width of the nonparametric confidence intervals.

Fig. 4(b) shows that in the analysis of the weeks, the SO_2 concentration decreases very noticeably on the weekends and remains on the working days, except on Monday, because the drop achieved on the weekend is still maintained. For this variable, the same results are obtained by the

nonparametric intervals as by the Wilcoxon rank sum test. Additionally, the behavior seems to be similar for the weeks in all the years, and the amplitudes of the intervals are smaller the lower the median.

Finally, Fig. 4(c) indicates that the SO_2 concentration in the study of the variables every two hours reaches its maximum between 10:00 and 11:00. In the other hours of the day, there are abrupt drops to concentrations below $3\mu g/m^3$, although there is a rebound between 20:00 and 23:00. In this case, the number of categories is greater because there are transit variables between a state with a high SO_2 concentration and one with a low SO_2 concentration.

IV. DATA CLASSIFICATION BY USING ROBUST STATISTICS

In this paper, robust statistics [21]–[23] are used to obtain measurements of the central tendency and scale that are as immune as possible to the effect on the analysis of data that show unusual observations [30], [31]. These statistics are characterized by the influence curve [39], which is a continuous, bounded curve for robust estimators that prevents the estimators from being affected by observations that stray from the data set.

Here, for the analysis, the sample order statistics [27] will be used. That is, given $X_1, \ldots, X_n$, the ordered sample is given by $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$, where $X_{(1)}$ is the observation with the lowest value and $X_{(n)}$ is the observation with the highest value.

A. CENTRAL TENDENCY AND SCALE ESTIMATORS

The statistics used in this section of the paper can be found in [21]–[23]. The L-location estimators used for the analysis were the following:

1) Trimean [21], [40]:

$$TM = \frac{Q_1 + 2Q_2 + Q_3}{4} \tag{4}$$

where Q_i stands for the i-th quartile [30].

2) α-trimmed mean [21]–[23]:

$$T(\alpha) = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{(i)} \tag{5}$$

where $0 \leq \alpha \leq 0.5$, n is the sample length, $[.]$ is the integer part, and the i-th order statistic of the sample is given by $X_{(i)}$.

3) α-winsorized mean, $W(\alpha)$, with $0 \leq \alpha \leq 0.5$ [23]. To find this estimator, the first step was to build a new vector by replacing the sample values that were above the $(1-\alpha)$ percentile with $X_{(1-\alpha)}$ and replacing the sample values that were below the α percentile with $X_{(\alpha)}$. Then, the mean value of this new vector was found.

Furthermore, the M-location estimators [21]–[23] used for the analysis were the following:

1) Andrew's wave [21], [23]:

$$T_{wa} = M_e + (cMAD) \arctan\left(\frac{\sum_{|u_i|<1} \sin(\pi u_i)}{\pi \sum_{|u_i|<1} \cos(\pi u_i)}\right) \tag{6}$$

where $c = 2.4\pi$, M_e stands for the median of the sample,

$$u_i = \frac{x_i - M_e}{cMAD} \tag{7}$$

where x_i is the i-th observed value, for $i = 1, \ldots, n$, and MAD is the median absolute deviation [21]–[23],

$$MAD = Median\{|x_1 - M_e|, \ldots, |x_n - M_e|\}. \tag{8}$$

2) Biweight [21], [22]:

$$T_{bi} = M_e + \frac{\sum_{|u_i|<1} (x_i - M_e)(1 - u_i^2)^2}{\sum_{|u_i|<1} (1 - u_i^2)^2} \tag{9}$$

where u_i is given by (7) with $c = 9$.

Table 3 shows the estimates of the abovementioned statistics for the 0.2-trimmed mean, 0.3-trimmed mean, 0.2-winsorized mean, and 0.3-winsorized mean.

Additionally, scale estimators [21], [22], [39], [41] were used to analyze the variability of the data. The scale estimators were as follows:

1) Sample standard deviation:

$$s_x = \left(\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2\right)^{\frac{1}{2}} \tag{10}$$

where $\bar{X}$ is the sample mean, the i-th element of the sample is given by X_i, and n is the sample length.

2) Mean absolute deviation:

$$MAD_{mean} = \frac{1}{n} \sum_{i=1}^{n} |X_i - \bar{X}| \tag{11}$$

3) Median absolute deviation (MAD), given by (8).

4) Half of the interquartile range [21], [42]:

$$SRH = \frac{1}{2}(H_2 - H_1) \tag{12}$$

where $H_1 = X_{(h_1)}$, $h_1 = \left[\frac{\left[\frac{n+1}{2}\right] + 1}{2}\right]$, $H_2 = X_{(h_2)}$, and $h_2 = n + 1 - h_1$.

5) Least median squares (LMS):

$$LMS = \frac{1}{2} \min_{i=1,\ldots,[n/2]} (|X_{(i+[n/2])} - X_{(i)}|) \tag{13}$$

6) Winsorized standard error of order α [23], $0 \leq \alpha \leq 0.5$:

$$s^W(\alpha) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(W_i - \overline{W}\right)^2} \tag{14}$$

TABLE 3. Point estimates of the location.

Year	Mean	Median	Trimean	0.2-trimmed mean	0.3-trimmed mean	0.2-winsorized mean	0.3-winsorized mean	Andrew's wave	Biweight
		M_e	TM	$T(0.2)$	$T(0.3)$	$W(0.2)$	$W(0.3)$	$T_{wa}(2.4\pi)$	$T_{bi}(9)$
2008 (X_1)	8.0830	6.8700	7.0925	7.8619	7.7835	7.2760	7.0478	7.2987	7.3027
2009 (X_2)	6.9080	5.7250	5.9025	6.4120	6.6638	6.0737	5.8806	6.0567	6.0674
2010 (X_3)	5.1788	4.2400	4.4025	5.3135	5.1451	4.5487	4.3702	4.4573	4.4755
2011 (X_4)	5.2712	4.3500	4.5450	5.5908	5.6849	4.7158	4.5260	4.6960	4.6972
2012 (X_5)	3.1074	2.4500	2.5850	3.1399	3.3475	2.7021	2.5778	2.6843	2.6857
2013 (X_6)	4.8171	3.6600	3.8875	4.3247	4.1889	4.0803	3.8525	4.0209	4.0261
2014 (X_7)	4.4210	3.2900	3.5125	3.7448	3.5885	3.7057	3.4787	3.6065	3.6146
2015 (X_8)	3.7672	2.5900	2.8275	3.3536	3.2117	3.0379	2.7922	2.8460	2.8613
2016 (X_9)	4.3526	3.2400	3.4425	3.7941	3.5712	3.6036	3.4128	3.5226	3.5314
2017 (X_{10})	2.9068	2.4600	2.5500	2.6856	2.7448	2.6246	2.5469	2.6196	2.6205
2018 (X_{11})	3.3227	2.5300	2.6300	3.5483	3.2439	2.7192	2.6169	2.6549	2.6689
2019 (X_{12})	2.4893	2.1000	2.1775	2.5154	2.3586	2.2449	2.1675	2.2196	2.2216

TABLE 4. Point estimates of the scale.

Year	s_x	MAD_{mean}	MAD	SRH	LMS	$s^W(0.2)$	$s_{wa}(2.4\pi)$	$s_{bi}(9)$	$C_n^{0.2}$
2008 (X_1)	5.3320	3.7352	2.5500	2.6950	2.3600	2.6020	3.9450	4.0442	3.5348
2009 (X_2)	5.2932	3.3742	2.1850	2.3000	2.0150	2.2301	3.3185	3.4250	3.0246
2010 (X_3)	3.7815	2.5606	1.5600	1.6850	1.4650	1.6470	2.4511	2.5602	2.2047
2011 (X_4)	3.6296	2.6284	1.7500	1.8800	1.6400	1.8469	2.7521	2.8367	2.5327
2012 (X_5)	2.5231	1.8001	1.1800	1.2700	1.0950	1.2435	1.8355	1.8948	1.6581
2013 (X_6)	4.1870	2.8184	1.7000	1.8850	1.5100	1.8644	2.6519	2.7733	2.2958
2014 (X_7)	4.0050	2.5409	1.4200	1.6050	1.2100	1.6221	2.2291	2.3555	1.8767
2015 (X_8)	3.8379	2.3445	1.1200	1.3850	0.9450	1.4091	1.7670	1.9175	1.4577
2016 (X_9)	4.1706	2.5271	1.4000	1.5650	1.2150	1.5416	2.1709	2.2816	1.8767
2017 (X_{10})	2.0238	1.3994	0.9600	1.0100	0.8900	0.9755	1.4841	1.5204	1.3483
2018 (X_{11})	3.4137	1.8909	1.0200	1.0900	0.9500	1.0710	1.5444	1.6319	1.4394
2019 (X_{12})	1.8681	1.0461	0.6500	0.7050	0.5950	0.6902	1.0288	1.0651	0.8928

where W_i is the i-th element of the α-winsorized sample and $\overline{W}$ is the unbiased estimate of the α-winsorized mean.

7) Andrew's wave:

$$s_{wa} = (cMAD)\frac{\sqrt{n\sum_{|u_i|<1}\sin^2(\pi u_i)}}{\pi\left|\sum_{|u_i|<1}\cos(\pi u_i)\right|} \tag{15}$$

where $c = 2.4\pi$, MAD is given by (8), and u_i is given by (7).

8) Biweight:

$$s_{bi} = \frac{\sqrt{n\sum_{|u_i|<1}(x_i - M_e)^2(1-u_i^2)^4}}{\left|\sum_{|u_i|<1}(1-u_i^2)(1-5u_i^2)\right|} \tag{16}$$

where $c = 9$ and u_i is given by (7).

9) C_n^α, for $0 < \alpha < 0.5$, given by [41]:

$$C_n^\alpha = \frac{subrange}{\Phi^{-1}(0.75) - \Phi^{-1}(0.75-\alpha)} \tag{17}$$

where

$$subrange = \left|X_{(i+[\alpha n]+1)} - X_{(i)}\right|_{([\frac{n}{2}]-[\alpha n])}$$

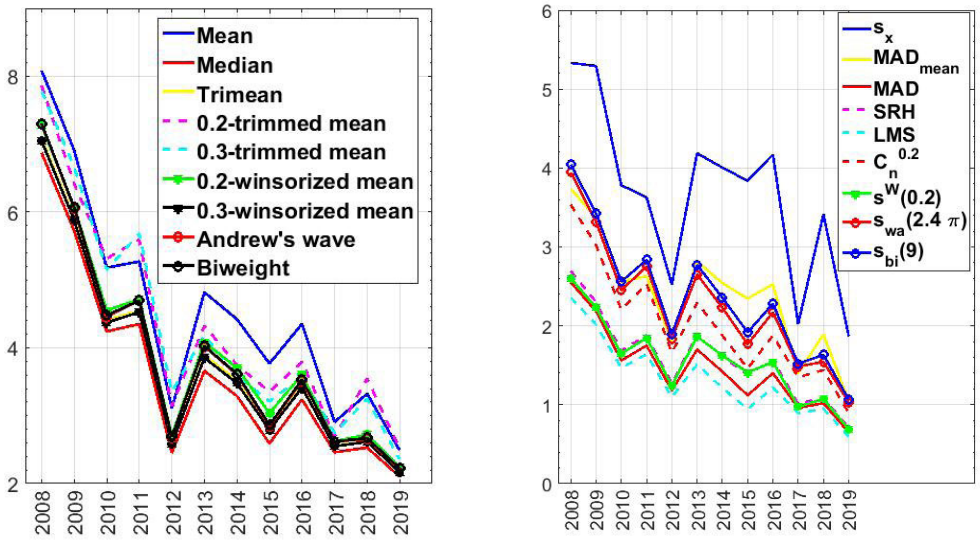
and Φ^{-1} stands for the inverse standard normal distribution cumulative distribution function [38].

Table 4 shows the point estimates of the scale found in this paper. Additionally, Fig. 5 shows the location and scale estimates by year for all variables under study. With the location estimates (see Fig. 5(a)), a pronounced decrease is verified again from 2008 to 2012. After a rebound, between 2013 and 2019, there is a stabilization with a slight decrease in all the estimates found. Note that, in general, all measures of centralization for each variable fluctuate between the median and the mean.

On the other hand, the scale estimate graph (Fig. 5(b)) shows that the standard deviation is an estimator that is well above all other estimates. In addition, the rest of the estimates are bounded lower by the LMS point estimator and higher by the MAD mean. However, in this case, there are several estimates that are similar. Additionally, it should be noted that there is a parallel between the scale estimates and the location estimates with respect to the fact that growth and decrease occur in the same years. This indicates that the increase in SO_2 concentration produces an increase in its variability.

In this paper, the abovementioned location and scale estimators were also used to analyze the variables that collect the SO_2 concentrations by month, day of the week, and every two hours. The estimates are shown in Fig. 6.

From Fig. 6, the variables for months, weeks and hours seem to follow the same pattern within each type of variable; that is, they increase and decrease at the same time. The summer months have a lower concentration of SO_2, and after September, there is a rebound that reaches its maximum in late November and then balances until summer. As with the descriptive analysis for weekdays (see Section II), the



(a) Location estimates (see Table 3).

(b) Scale estimates (see Table 4).

FIGURE 5. Location and scale estimates by year.

SO_2 concentration drops sharply on weekends and remains high on weekdays. In the first days of the week, the SO_2 concentration is lower because the drop over the weekend has to be compensated for. In addition, for the hours of the day, there is a rise in the concentration of SO_2 in the early hours of the morning, which drops abruptly for the rest of the hours of the day. Additionally, a small but relevant rise is seen at the end of the day. Again, for all the variables, the centralization measures are between the mean and the median.

Finally, from Fig. 6, all the scale estimates are bounded from below by the LMS point estimator. Additionally, there is a concordance between the location and scale estimates, according to which an increase in the SO_2 concentration leads to an increase in its variability.

B. CONFIDENCE INTERVALS: CLASSIC, BOOTSTRAP AND ROBUST

In this research, confidence intervals were built following the recommendations of [43], [44] and the suggestions of [21], [22]. In what follows, $t_{v,q}$ stands for the q-th quantile of the Student's t distribution with v degrees of freedom (DOFs) [38]. The confidence intervals are as follows:

1) $\left(\bar{X} \pm t_{n-1,\alpha/2}\ \dfrac{s_x}{\sqrt{n}}\right)$ for $(\overline{X},\ s_x)$, where $\overline{X}$ is the mean and s_x is given by (10).
2) $\left(M_e \pm t_{n-2,\alpha/2}\ \dfrac{MAD}{\sqrt{n}}\right)$ for $(M_e,\ MAD)$, where M_e is the median and MAD is given by (8).
3) $\left(M_e \pm \dfrac{t_{n-1,\alpha/2}}{1.075}\ \dfrac{IQR}{\sqrt{n}}\right)$ for $(M_e,\ IQR)$, where $IQR = Q_3 - Q_1$ is the interquartile range, with Q_1 being the first quartile and Q_3 being the third quartile. In the case under study, the IQR is considered to be similar to SRH (see (12)).
4) $\left(T(\alpha) \pm t_{n-2[n\alpha]-1,\alpha/2}\ \dfrac{s^W(\alpha)}{(1-2\alpha)\sqrt{n}}\right)$ for $(T(\alpha),\ s^W(\alpha))$, where $T(\alpha)$ is given by (5) and $s^W(\alpha)$ is given by (14).
5) $\left(T_{wa} \pm t_{[0.7\cdot(n-1)],\alpha/2}\ \dfrac{s_{wa}}{\sqrt{n}}\right)$ for $(T_{wa},\ s_{wa})$, where T_{wa} is given by (6) and s_{wa} is given by (15).
6) $\left(T_{bi} \pm t_{[0.7\cdot(n-1)],\alpha/2}\ \dfrac{s_{bi}}{\sqrt{n}}\right)$ for $(T_{bi},\ s_{bi})$, where T_{bi} is given by (9) and s_{bi} is given by (16).
7) Bootstrap confidence interval [23]:

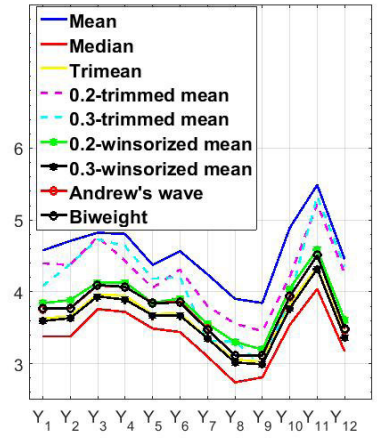
$$(M_e - t^*_{1-\alpha/2}\hat{s}^*,\ M_e + t^*_{\alpha/2}\hat{s}^*)$$

where $\hat{s}^*$ is the standard deviation unbiased estimator of the median of each bootstrap sample M_e^* and $t^*_{1-\alpha/2}$ and $t^*_{\alpha/2}$ are the percentiles of the statistic $\hat{M}_b^*$ given by

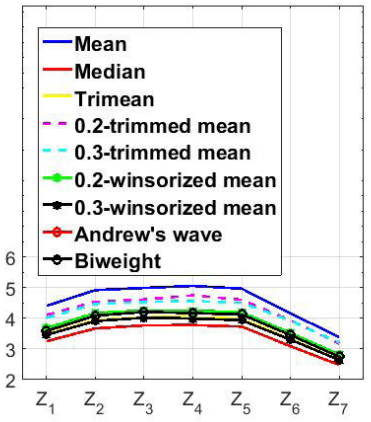
$$\hat{M}_b^* = \frac{M_e^* - M_e}{\hat{s}^*}.$$

In this paper, taking into account the previous information, eight confidence intervals were constructed for each of the twelve years under study. Specifically, these intervals were of the following types: five robust intervals, one nonparametric interval, one bootstrap interval, and one classic interval. In Fig. 7, these intervals are shown for the years 2008, 2014 and 2019. Showing more figures with confidence intervals would not provide relevant information here.

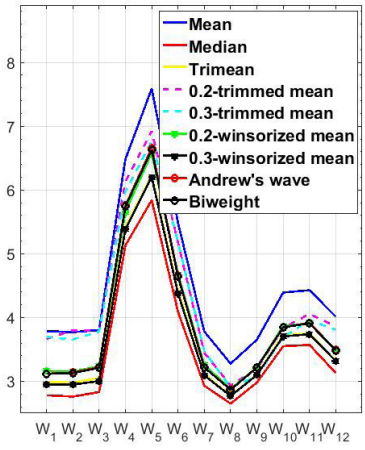
Despite the fact that only three of the twelve variables have been included in Fig. 7, it can be affirmed that they all have the same characteristics. First, the classic confidence intervals are the most displaced toward high values, while the median-based intervals are those with the lowest values. Furthermore, among these median-based intervals, the nonparametric and the bootstrap intervals are very similar.



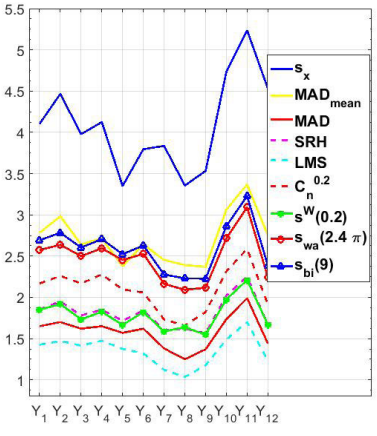
(a) Location estimates by month: Y_k, $k = 1, \ldots, 12$.



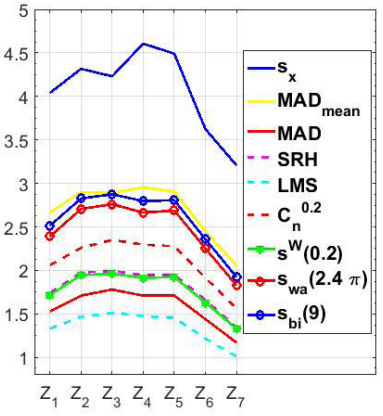
(b) Location estimates by day of the week: Z_k, $k = 1, \ldots, 7$.



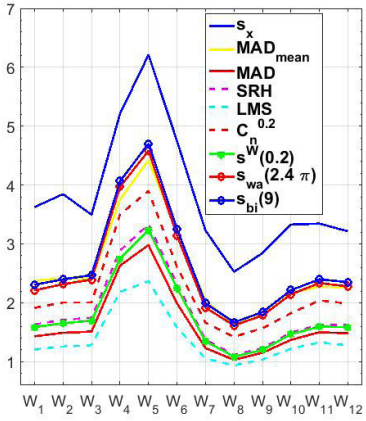
(c) Location estimates by hour in groups of two hours: W_k, $k = 1, \ldots, 12$.



(d) Scale estimates by month: Y_k, $k = 1, \ldots, 12$.



(e) Scale estimates by day of the week: Z_k, $k = 1, \ldots, 7$.



(f) Scale estimates by hour in groups of two hours: W_k, $k = 1, \ldots, 12$.

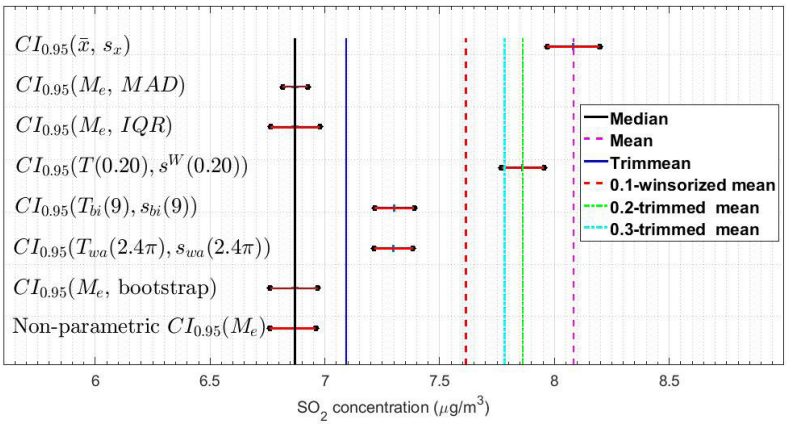
FIGURE 6. Location and scale estimates by month, day and every two hours.

Additionally, the narrowest intervals are those based on the pair (M_e, MAD) because the MAD scale estimator is the one with the lowest values among those chosen to construct the confidence intervals. Second, the intervals based on Andrew's wave and the biweight are analogous in all variables. Finally, the intervals based on the α-trimmed mean location estimators are those that are closest to the classic intervals.

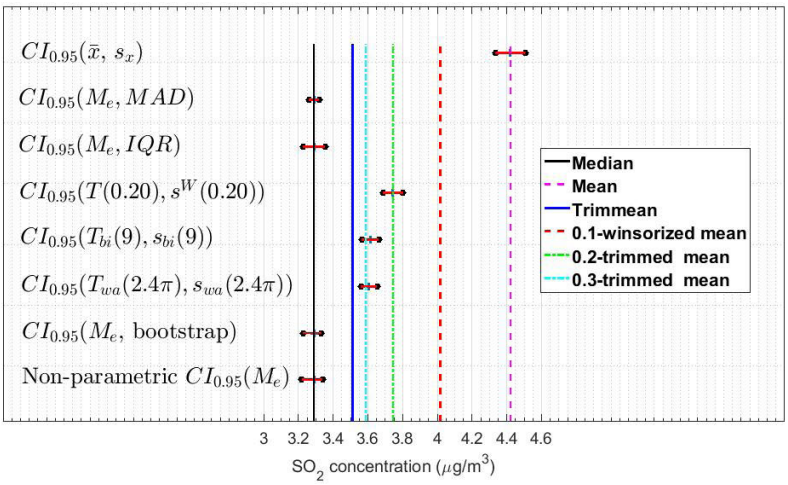
In accordance with what has been said in the previous paragraphs, the twelve variables under study are compared using the following pairs of estimators: ($T(\alpha)$, $s^W(\alpha)$) and (T_{bi}, s_{bi}). This decision was made because the classic intervals assumed that the underlying distribution is approximately normal, which is not true for the case under study. Additionally, the point estimators (M_e, MAD) and (M_e, IQR) and the bootstrap estimators yield results that are analogous to the results obtained in Section III by using the nonparametric estimators. However, the use of the pair (M_e, MAD) will produce more differences in the grouping of variables, eliminating the possibility of grouping similar behavior between the concentration of SO_2 by year. Moreover, the results for the estimators based on Andrew's wave and on the biweight are similar, so either of these two estimators can be chosen.

With a confidence level of 95%, both the confidence intervals and their lengths are shown in Table 5 for ($T(\alpha)$, $s^W(\alpha)$) and ($T_{bi}(c)$, $s_{bi}(c)$), with $\alpha = 0.2$ and $c = 9$.

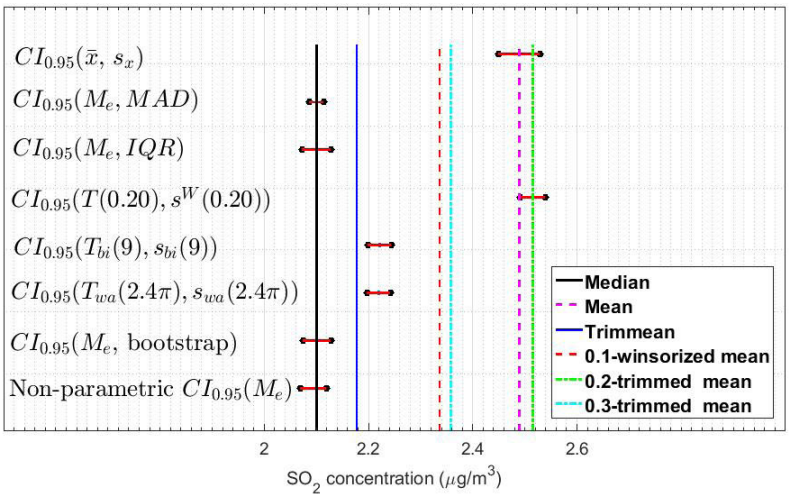
Figure 8 shows a graphical representation of the above-mentioned confidence intervals. Furthermore, to classify the variables, lines have been included. This classification is analogous to that carried out in Section III by using the Wilcoxon rank sum test for the medians. With the estimators ($T(0.2)$, $s^W(0.2)$) and ($T_{bi}(9)$, $s_{bi}(9)$), the classification of the variables is equivalent to that obtained by using nonpara-



(a) 95% confidence intervals for X_1 (2008).



(b) 95% confidence intervals for X_7 (2014).



(c) 95% confidence intervals for X_{12} (2019).

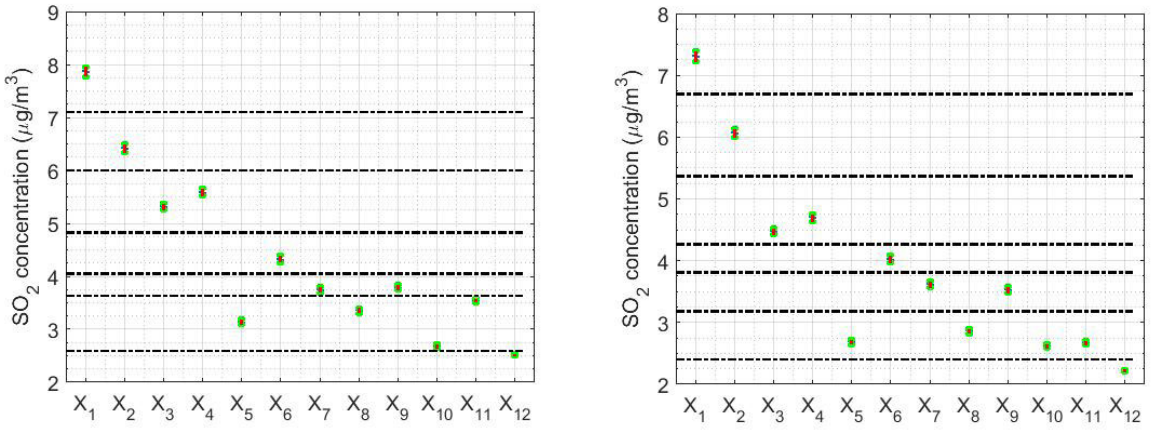
FIGURE 7. 95% confidence intervals (CI_{95}): classic, nonparametric, bootstrap, and robust confidence intervals.

metric estimators. However, the difference is that X_{10} can form a category by itself in the groupings given by the pair of estimators $(T(0.2),\ s^W(0.2))$.

Analogous to the nonparametric analysis, from Fig. 8, it can be observed that between 2008 and 2012, there is a trend of abrupt decrease in the SO_2 concentration

TABLE 5. Limits and lengths of the 95% confidence intervals for $(T(0.2),\ s^W(0.2))$ and $(T_{bi}(9),\ s_{bi}(9))$.

CI_{95}	Variable	Lower limit	Upper limit	Length	Variable	Lower limit	Upper limit	Length
$(T,\ s^W)$	X_1	7.2163	7.3891	0.1728	X_7	3.5645	3.6647	0.1001
$(T_{bi},\ s_{bi})$		7.7693	7.9546	0.1853		3.6873	3.8023	0.1149
$(T,\ s^W)$	X_2	5.9941	6.1408	0.1468	X_8	2.8206	2.9020	0.0814
$(T_{bi},\ s_{bi})$		6.3323	6.4916	0.1593		3.3037	3.4034	0.0997
$(T,\ s^W)$	X_3	4.4210	4.5300	0.1091	X_9	3.4817	3.5810	0.0992
$(T_{bi},\ s_{bi})$		5.2550	5.3719	0.1169		3.7382	3.8499	0.1118
$(T,\ s^W)$	X_4	4.6367	4.7578	0.1211	X_{10}	2.5875	2.6534	0.0659
$(T_{bi},\ s_{bi})$		5.5251	5.6565	0.1314		2.6504	2.7209	0.0705
$(T,\ s^W)$	X_5	2.6453	2.7261	0.0808	X_{11}	2.6339	2.7038	0.0699
$(T_{bi},\ s_{bi})$		3.0957	3.1841	0.0884		3.5100	3.5865	0.0765
$(T,\ s^W)$	X_6	3.9663	4.0859	0.1195	X_{12}	2.1990	2.2442	0.0453
$(T_{bi},\ s_{bi})$		4.2577	4.3916	0.1339		2.4910	2.5399	0.0489



(a) 95% confidence intervals based on $(T(0.2),\ s^W(0.2))$. (b) 95% confidence intervals based on $(T_{bi}(9),\ s_{bi}(9))$.

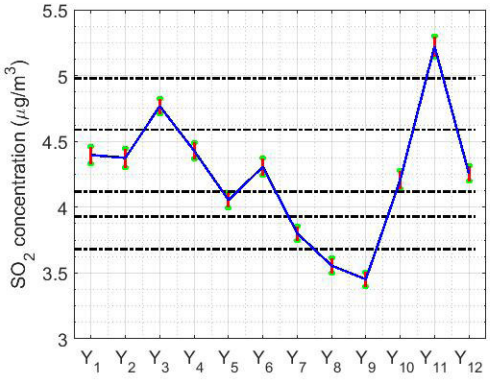
FIGURE 8. 95% confidence intervals based on $(T(0.2),\ s^W(0.2))$ and $(T_{bi}(9),\ s_{bi}(9))$.

and that, after a rise in 2013, there are fluctuations until 2019 with a slight downward trend. Regarding the amplitude, it can be concluded that the confidence intervals found with $(T_{bi}(9),\ s_{bi}(9))$ are less narrow than those found with $(T(0.2),\ s^W(0.2))$. The amplitudes of these intervals evolve in parallel with the values of the medians.

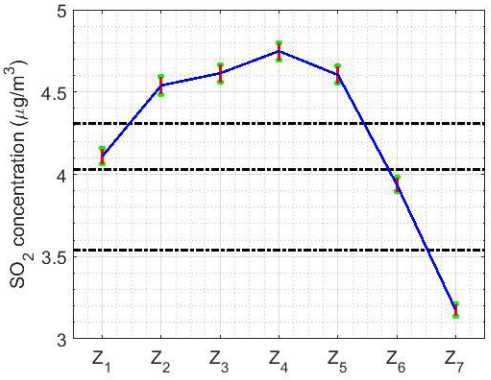
For the study of the confidence intervals with the variables grouped by month, day of the week and every two hours, the graphs shown in Fig. 9 are included. The graphs in Fig. 9(a) and 9(d) show that the lowest values per month are reached at the end of the summer, with SO_2 concentration values greater than $3\mu g/m^3$, after a pronounced decline that begins in early summer. After September, there is a growth that reaches its maximum in November, and in December, there is a decrease of approximately 20%, which tends to continue until February, although a small jump from December to January can be observed. Between winter and spring, there is a growth of 10%, and in June, the decline begins until the minimum of each year is reached. The results are similar to those shown in Fig. 4(a). Furthermore, with respect to the widths of the confidence intervals, these appear, in general, to be narrower than for the analysis of the years. However, the effect can also be oberved here that the greater the value of the median is, the greater the value of the width of the confidence intervals.

Regarding the analysis for the days of the week (see Fig. 9(b) and 9(e)), it is analogous to that obtained with the nonparametric estimators (see Section III and Fig. 4(b)). Specifically, on weekends, the minimum SO_2 concentration is obtained; that is, values 33% less than the values obtained on weekdays are obtained. In addition, there are four categories: one for each of the weekend days, one for Monday, and one that groups the other days of the week. The category of Monday, like that of Saturday, is between weekends and the rest of the working days because it is in transit between two states: working days and Sunday.

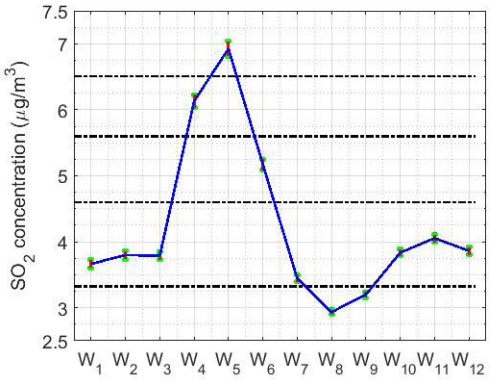
Finally, the results of the analysis for the hours (see Fig. 9(c) and 9(f)) are also very similar to those found with the nonparametric estimation (see Fig. 4(c)). Moreover, the results obtained with the estimators $(T_{bi}(9),\ s_{bi}(9))$ are the same as those obtained with the nonparametric estimation. However, differences appear in the estimates of $(T(0.2),\ s^W(0.2))$ because in the early hours of the day, the estimates appear to be somewhat the same as for late



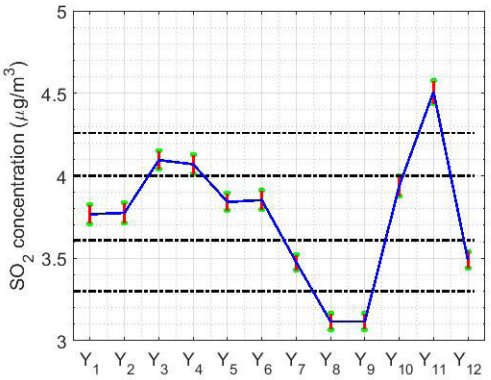
(a) 95% confidence intervals based on $(T(0.2),\ s^W(0.2))$: Y_k, $k = 1, \ldots, 12$.



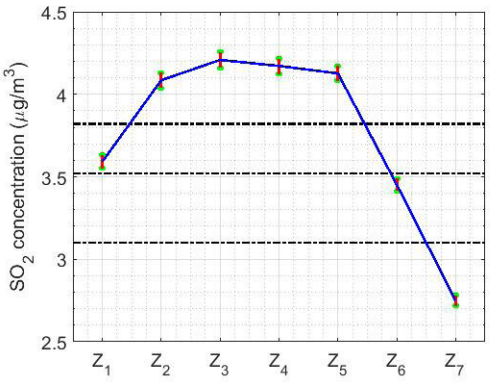
(b) 95% confidence intervals based on $(T(0.2),\ s^W(0.2))$: Z_k, $k = 1, \ldots, 7$.



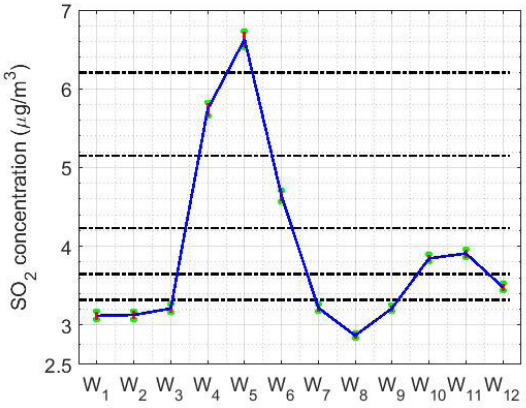
(c) 95% confidence intervals based on $(T(0.2),\ s^W(0.2))$: W_k, $k = 1, \ldots, 12$.



(d) 95% confidence intervals based on $(T_{bi}(9),\ s_{bi}(9))$: Y_k, $k = 1, \ldots, 12$.



(e) 95% confidence intervals based on $(T_{bi}(9),\ s_{bi}(9))$: Z_k, $k = 1, \ldots, 7$.



(f) 95% confidence intervals based on $(T_{bi}(9),\ s_{bi}(9))$: W_k, $k = 1, \ldots, 12$.

FIGURE 9. 95% confidence intervals based on $(T(0.2),\ s^W(0.2))$ and $(T_{bi}(9),\ s_{bi}(9))$ by month (Y_k), day of the week (Z_k), and every two hours (W_k).

afternoon hours. In any case, the general conclusions are the same: the maximum is reached at approximately 10:00 - 11: 00, and the SO_2 concentration decreases to more than 40% by 16:00 - 17:00. Then, the SO_2 concentration grows by 25% in the late afternoon. Later, the SO_2 concentration decreases again at night and in the early morning, although it does not decrease as much as in the afternoon.

V. CONCLUSION

The objective of this paper was to analyze the general behavior of the SO_2 concentration at the air quality monitoring station of Belisario, Quito, Ecuador over the last twelve years (i.e., from January 2008 to December 2019). To this end, four types of variables were considered: the year, month of the year, day of the week, and hour of the day.

After verifying that no separate sets of variables came from the same distribution, the aim was to determine the differences between the parameters that characterize these variables.

With an initial statistical summary, it was observed that all the SO_2 concentration values at the Belisario station were values that were found to be at the acceptable level of air quality pollution according to the Quiteño Air Quality Index. However, it was also observed that the SO_2 concentration values that exceed the value of desirable quality, according to the Quiteño Air Quality Index, were always extreme observations, although they were not the only ones. Furthermore, it should be noted that in the study carried out in this paper, the only extreme observations were from the right, never from the left. Additionally, all the variables of each time period presented characteristics that were compatible with the possibility that these variables followed heavy-tailed distributions.

Another important result of the research presented here was that after smoothing the data for all years, and for each of the years in particular, a tendency toward decreasing SO_2 concentration values was observed across the years. In addition, this feature also occurred across months, days of the week, and hours of the day.

Then, due to the impossibility of using classical inference, the variables under study were characterized by hypothesis testing and both nonparametric and robust confidence intervals. Additionally, different robust location and scale statistics were found, and some of them were used to determine the robust confidence intervals. During the analysis, it was observed that all the location estimates were between the mean and the median, that the amount of SO_2 concentration at the Belisario station decreased markedly between 2008 and 2012 and increased in 2013 and that from then on, oscillations occurred with a slight continued drop.

Regarding the analysis using the scale estimators, all these estimations were found to be in a band where the standard deviation was well above all of them, and the other estimations were bounded from below by the least median of squares estimator.

These observations highlight the fact that there is a parallel between the location estimates and scale estimates, in the sense that an increase or decrease in the value of the point estimates also produces an increase or decrease, respectively, in the value of the scale estimates. This result leads to the conclusion that the extreme observations, i.e., outliers, in quantity and value were the ones that determined the location and scale estimates in this research.

For reasons explained in the paper, the confidence intervals that were chosen to compare the variables were, on the one hand, the confidence intervals based on the α-trimmed mean and winsorized standard deviation and, on the other hand, the confidence intervals based on the biweight estimators. Here, the existence of a downward trend in the SO_2 concentration between 2008 and 2012 could be seen again; the concentration rises in 2013, and after that year, there are fluctuations that show a slight tendency to decrease.

In both the variables that represent the months and those that represent the days of the week and hours of the day, there was a certain periodicity. In the analysis by month, notable decreases were observed in summer, and there was an increase at the end of the year and some stability in the first quarter of each year. Additionally, in the analysis by day of the week, there was a clear difference in the SO_2 concentration between working days and weekends. Finally, the hourly analysis also showed minimum values in the early afternoon, maximum values in the early hours of the morning and night, and stable concentration values at night and in the early morning.

With the analysis carried out in this paper, it was possible to group the variables under study by comparing the results of the data by year, month, day of the week, and hour of the day. In addition, it was possible to find the differences between the categories that were established, and these differences were quantified by using confidence intervals. All of the above was exhaustively developed to robustly estimate the SO_2 concentration measurements at the Belisario station over the last twelve years.

The final conclusion of this paper is that the trend of the SO_2 concentration at Belisario station is downward, and this was measured with the precision provided by robust statistical methods. Therefore, it can be said that the measures that have been taken by the Quito city council over the last few years are yielding good results.

REFERENCES

[1] Autralian Department of Agriculture, Water and the Environment. *Sulfur Dioxide (SO_2)*. Environment Protection Publications and Resources. Accessed: Apr. 12, 2020. [Online]. Available: https://www.environment.gov.au/protection/publications/factsheet-sulfur-dioxide-so2

[2] Minnesota Pollution Control Agency. *Sulfur Dioxide (SO_2)*. Air Pollutants. Accessed: Apr. 12, 2020. [Online]. Available: https://www.pca.state.mn.us/air/sulfur-dioxide-so2

[3] WHO 2006. (2006). *WHO Air Quality Guidelines for Particulate Matter, Ozone, Nitrogen Dioxide and Sulfur Dioxide. Global Update 2005. Summary of Risk Assessment*. World Health Organization. Accessed: Apr. 12, 2020. [Online]. Available: https://www.pca.state.mn.us/air/sulfur-dioxide-so2

[4] United States Environmental Protection Agency. *Sulfur Dioxide Basics*. Sulfur Dioxide (SO_2) Pollution. Accessed: Apr. 13, 2020. [Online]. Available: https://www.epa.gov/so2-pollution/sulfur-dioxide-basics

[5] K. Nakano, R. Hirofuji, T. Ohnishi, M. Hauta-Kasari, I. Nishidate, and H. Haneishi, "RGB camera-based imaging of oxygen saturation and hemoglobin concentration in ocular fundus," *IEEE Access*, vol. 7, pp. 56469–56479, 2019.

[6] J. Niu, L. E. Flynn, T. Beck, Z. Zhang, and E. Beach, "Evaluation and improvement of the near-real-time linear fit SO_2 retrievals from suomi NPP ozone mapping and profiler suite," *IEEE Trans. Geosci. Remote Sens.*, early access, May 19, 2020, doi: 10.1109/TGRS.2020.2992429.

[7] L. Flynn, C. Long, X. Wu, R. Evans, C. T. Beck, I. Petropavlovskikh, G. McConville, W. Yu, Z. Zhang, J. Niu, E. Beach, Y. Hao, C. Pan, B. Sen, M. Novicki, S. Zhou, and C. Seftor, "Performance of the ozone mapping and profiler suite (OMPS) products," *J. Geophys. Res., Atmos.*, vol. 119, no. 10, pp. 6181–6195, May 2014.

[8] P. S. Kanaroglou, M. D. Adams, P. F. De Luca, D. Corr, and N. Sohel, "Estimation of sulfur dioxide air pollution concentrations with a spatial autoregressive model," *Atmos. Environ.*, vol. 79, pp. 421–427, Nov. 2013.

[9] J. F. Rogers, G. G. Killough, S. J. Thompson, C. L. Addy, R. E. Mckeown, and D. J. Cowen, "Estimating environmental exposures to sulfur dioxide from multiple industrial sources for a case–control study," *J. Exposure Sci. Environ. Epidemiology*, vol. 9, no. 6, pp. 535–545, Dec. 1999.

[10] J. Z. Holland, "A meterological survey of the Oak Ridge area: Final report covering the period 1948-1952," Weather Bureau, Oak Ridge, TN, USA, Tech. Rep. USAEC/Report ORO-99, 1953.
[11] P. J. Meade and F. Pasquill, "A study of the average distribution of pollution around Staythorpe," *Int. J. Air Pollut.*, vol. 1, pp. 60–70, Oct. 1958.
[12] R. Kh Turgumbayeva, M. N. Abdikarimov, R. Mussabekov, and D. Sartayev, "Application of statistical analysis method to determine distribution of sulfur dioxide and phosphorus oxides emissions from industrial enterprise 'KazFosfat,'" in *Proc. IOP Conf. Ser., Earth Environ. Sci.*, vol. 194. Bristol, U.K.: IOP Publishing, Nov. 2018, p. 022042, doi: 10.1088/1755-1315/194/2/022042.
[13] P. Bolzern, G. Fronza, E. Runca, and C. Überhuber, "Statistical analysis of winter sulphur dioxide concentration data in vienna," *Atmos. Environ.*, vol. 16, no. 8, pp. 1899–1906, Jan. 1982.
[14] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken, NJ, USA: Wiley, 2016.
[15] P. J. Brockwell and R. A. Davis, *Introduction to Time Series an Forecasting*, 2th ed. New York, NY, USA: Springer, 2002.
[16] Y. Hou, L. Wang, Y. Zhou, S. Wang, and F. Wang, "Analysis of the sulfur dioxide column concentration over Jing-Jin-Ji, China, based on satellite observations during the past decade," *Polish J. Environ. Stud.*, vol. 27, no. 4, pp. 1551–1557, 2018.
[17] J. Tomić, M. Kušljević, M. Vidaković, and V. Rajs, "Smart SCADA system for urban air pollution monitoring," *Measurement*, vol. 58, pp. 138–146, Dec. 2014.
[18] S. J. Smith, J. van Aardenne, Z. Klimont, R. J. Andres, A. Volke, and S. D. Arias, "Anthropogenic sulfur dioxide emissions: 1850–2005," *Atmos. Chem. Phys.*, vol. 11, no. 3, pp. 1101–1116, Feb. 2011.
[19] L. S. Gharib and M. Al Sarawi, "Sulfur dioxide (SO_2) and heavy metals accumulation in soils around oil refineries: Case study from three southern oil refineries in the state of kuwait," *Amer. J. Environ. Sci.*, vol. 14, no. 1, pp. 12–43, Jan. 2018.
[20] *Belisario*. Secretaría de Ambiente del Municipio del Distrito Metropolitano Quito. Accessed: Apr. 20, 2020. [Online]. Available: http://www.quitoambiente.gob.ec/ambiente/index.php/belisario
[21] D. C. Hoaglin, F. Mosteller, and J. W. Tukey, *Understanding Robust and Exploratory Data Analysis*. Hoboken, NJ, USA: Wiley, 2000.
[22] R. A. Maronna, R. D. Martin, and R. D. Yohai, *Robust Statistics: Theory and Methods*. Chichester, U.K.: Wiley, 2006.
[23] R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing*, 3rd ed. Waltham, MA, USA: Academic, 2012.
[24] *Red de Monitoreo Atmosférico*. Secretaría de Ambiente del Municipio del Distrito Metropolitano Quito. Accessed: Apr. 20, 2020. [Online]. Available: http://www.quitoambiente.gob.ec/ambiente/index.php/politicas-y-planeacion-ambiental/red-de-monitoreo
[25] M. C. Bryson, "Heavy-tailed distributions: Properties and tests," *Technometrics*, vol. 16, no. 1, pp. 61–68, Feb. 1974.
[26] M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric Statistical Methods*, 3rd ed., Hoboken, NJ, USA: Wiley 2014.
[27] J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference*, 5th ed. Boca Raton, FL, USA: Chapman & Hall, 2011.
[28] W. Hernandez, A. Mendez, A. M. Diaz-Marquez, and R. Zalakeviciute, "PM_2.5 concentration measurement analysis by using non-parametric statistical inference," *IEEE Sensors J.*, vol. 20, no. 2, pp. 1084–1094, Jan. 2020.
[29] W. Hernandez, A. Mendez, R. Zalakeviciute, and A. M. Diaz-Marquez, "Analysis of the information obtained from $PM_{2.5}$ concentration measurements in an urban park," *IEEE Trans. Instrum. Meas.*, early access, Jan. 13, 2020, doi: 10.1109/TIM.2020.2966360.
[30] W. Hernandez, A. Mendez, A. M. Diaz-Marquez, and R. Zalakeviciute, "Robust analysis of PM2.5 concentration measurements in the ecuadorian park la carolina," *Sensors*, vol. 19, no. 21, p. 4648, Oct. 2019.
[31] W. Hernandez, A. Mendez, R. Zalakeviciute, and A. M. Diaz-Marquez, "Robust confidence intervals for PM2.5 concentration measurements in the ecuadorian park la carolina," *Sensors*, vol. 20, no. 3, p. 654, Jan. 2020.
[32] Y. Zhang, Y. Wang, M. Gao, Q. Ma, J. Zhao, R. Zhang, Q. Wang, and L. Huang, "A predictive data feature exploration-based air quality prediction approach," *IEEE Access*, vol. 7, pp. 30732–30743, 2019.
[33] G. ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, 2017, pp. 3149–3157.
[34] J. Ma, Y. Ding, V. J. L. Gan, C. Lin, and Z. Wan, "Spatiotemporal prediction of PM2.5 concentrations at different time granularities using IDW-BLSTM," *IEEE Access*, vol. 7, pp. 107897–107907, 2019.
[35] X. Chen, H. Li, S. Zhang, Y. Chen, and Q. Fan, "High spatial resolution $PM_{2.5}$ retrieval using MODIS and ground observation station data based on ensemble random forest," *IEEE Access*, vol. 7, pp. 44416–44430, 2019.
[36] V. Díaz, "Informe Calidad del Aire 2017," *Secretaría de Ambiente del Distrito Metropolitano de Quito*, Accessed on: Apr. 20, 2020. [Online]. Available: http://www.quitoambiente.gob.ec/ambiente/index.php/informes#informe-calidad-del-aire-2017.
[37] T. Scientific. *Model 43i SO_2 Analyzer*. Thermo Fisher Scientific. Accessed: Mar. 21, 2020. [Online]. Available: https://www.thermofisher.com/order/catalog/product/43I#/43I
[38] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed. New York, NY, USA: McGraw-Hill, 2002.
[39] F. R. Hampel, "The influence curve and its role in robust estimation," *J. Amer. Stat. Assoc.*, vol. 69, no. 346, pp. 383–393, Jun. 1974.
[40] J. W. Tukey, *Exploratory Data Analysis*. Boston, MA, USA: Addison-Wesley, 1977.
[41] C. Croux and P. J. Rousseeuw, "A class of high-breakdown scale estimators based on subranges," *Commun. Statist. Theory Methods*, vol. 21, no. 7, pp. 1935–1951, Jan. 1992.
[42] N. M. S. Rock, "ROBUST: An interactive FORTRAN-77 package for exploratory data analysis using parametric, ROBUST and nonparametric location and scale estimates, data transformations, normality tests, and outlier assessment," *Comput. Geosci.*, vol. 13, no. 5, pp. 463–494, Jan. 1987.
[43] W. J. Dixon and J. W. Tukey, "Approximate behavior of the distribution of Winsorized t (Trimming/Winsorization 2)," *Technometrics*, vol. 10, no. 1, pp. 83–98, 1968.
[44] F. Mosteller and J. W. Tukey, *Analysis and Regression; A Second Course in Statistics*. Reading, MA, USA: Addison-Wesley, 1977.

WILMAR HERNANDEZ (Senior Member, IEEE) received the degree in electronics engineering and the specialist degree in microelectronics from the Instituto Superior Politecnico Jose Antonio Echeverria (ISPJAE), Havana, Cuba, in 1992 and 1994, respectively, and the M.S. degree in signal treatment and the Ph.D. degree in electronic engineering from Enginyeria La Salle, Universitat Ramon Llull, Barcelona, Spain, in 1997 and 1999, respectively. From 1992 to 1995, he was a Lecturer with the Electrical Engineering Faculty, ISPJAE, and a Researcher with the Microelectronics Research Center, ISPJAE. From 1999 to 2003, he was with the Department of Electronics and Instrumentation, University Institute for Automobile Research, Universidad Politecnica de Madrid (UPM), Spain, where he was the Technical Director of the department, from January 2003 to January 2004. From January 2004 to March 2013, he was an Associate Professor of circuits and systems with the Department of Circuits and Systems, EUIT de Telecomunicacion, UPM. From September 2014 to September 2015, he was a Researcher with SENESCYT, Ecuador, under the Prometeo Fellowship Program. From December 2015 to November 2017, he was a Professor with the Universidad Tecnica Particular de Loja, Ecuador. Since January 2018, he has been a Professor with the Universidad de Las Americas, Ecuador.

ALFREDO MENDEZ was born in Madrid, Spain, in June 1958. He received the degree in mathematical sciences (fundamental mathematics), the M.S. degree in mathematical sciences, and the Ph.D. degree in mathematical sciences (statistics and operational research) from the Universidad Complutense de Madrid (UCM), in 1981, 1987, and 1995, respectively. From 1983 to 1993, he was a Lecturer with the EUIT Agricola, Universidad Politecnica de Madrid (UPM), Spain. Since 1993, he has been an Associate Professor of mathematics with the Departamento de Matematica Aplicada a las Tecnologias de la Informacion y Comunicaciones (DMATIC), ETSIS de Telecomunicacion, UPM, where he was the Director of the DMATIC, from May 2004 to May 2012.

VICENTE GONZÁLEZ-POSADAS was born in Madrid, Spain, in 1968. He received the B.S. degree in radio communication engineering from the Universidad Politecnica de Madrid (UPM), Madrid, in 1992, the M.S. degree in physics from the Universidad Nacional de Educacion a Distancia, Madrid, in 1995, the M.S. degree in high strategic studies from the CESEDEN, Madrid, in 2009, and the Ph.D. degree in telecommunication engineering from the Universidad Carlos III de Madrid, Madrid, in 2001. He is currently a Full Professor with the ETSIS de Telecomunicacion, UPM. He has authored or coauthored over 60 technical conferences, letters, and journal articles. His current research interests include active antennas, microstrip antennas, CRLH lines and metamaterials, microwave technology, and RFID.

JOSÉ LUIS JIMÉNEZ-MARTÍN was born in Madrid, Spain, in 1967. He received the B.S. degree in electrical engineering with a minor in radio communication engineering, the M.S. degree in telecommunications engineering, and the Ph.D. degree from the Universidad Politecnica de Madrid (UPM), Madrid, in 1991, 2000, and 2005, respectively, and the M.S. degree in high strategic studies from CESEDEN, Madrid, in 2007. He is currently an Associate Professor with the ETSIS de Telecomunicacion, UPM. He has authored or coauthored over 60 technical conferences, letters, and journal articles. His current research interests include oscillators, amplifiers, and microwave technology.