Applied Mathematics and Computation 241 (2014) 1–10

Contents lists available at ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

Air quality parameters outliers detection using functional data analysis in the Langreo urban area (Northern Spain)

J. Martínez [a,*], Á. Saavedra [b], P.J. García-Nieto [c], J.I. Piñeiro [b], C. Iglesias [b], J. Taboada [b], J. Sancho [a], J. Pastor [a]

[a] Centro Universitario de la Defensa, Academia General Militar, Zaragoza, 50090 Zaragoza, Spain
[b] Department of Natural Resources and Environmental Engineering, University of Vigo, 36310 Vigo, Spain
[c] Department of Mathematics, Faculty of Sciences, University of Oviedo, 33007 Oviedo, Spain

ARTICLE INFO

Keywords:
Air quality control
Outliers detection
Gas pollution
Functional analysis

ABSTRACT

Polluted air of cities is a harmful factor to health that may eventually cause respiratory problems and cardiovascular disease. The monitoring and control of pollutants is an essential activity in order to protect the environment and the health by minimizing pollution levels through the detection of contaminants.

Contaminants are emissions of substances to the atmosphere (mainly gases and particulate matter) whose values are greater than the limits allowed by the environmental legislation (they are anomalous values). Thus they are considered as vector samples where each component represents the gas concentration value in the air.

In this sense, a model based on functional analysis has been implemented for the outliers detection in air quality samples in this research work. This model transforms the vectorial sample by creating a new functional sample in order to determine functional outliers by adjusting the concept of depth to the functional event. This method has been compared to classical outliers analysis from a vectorial point of view, emphasizing the power of use of such functional techniques over the traditional ones.

The main aim of this research work is to compare the results corresponding to the classical and the functional methods and to obtain the most appropriate methodology to analyze this type of dataset in order to reach a better solution for the air quality control.

1. Introduction

Air pollution is an important environmental problem in cities [1–4]. Air is never perfectly clean and polluted air is a continuing threat to human health and welfare [5,6]. An average adult male requires about 13.5 kg of air each day compared with about 1.2 kg of food and 2 kg of water. Therefore, clean air should certainly be as important to us as clean water and food.

There are a number of sources of air pollution that affect human health [1,7]. Information on meteorological pollution, such as that produced by carbon monoxide (CO), nitrogen oxides (NO and NO_2), sulfur dioxide (SO_2), ozone (O_3) and particulate matter (PM_{10}), is increasingly important due to the harmful effects on human health [4,8]. Automated measurement of

[*] Corresponding author.
E-mail addresses: jmtorres@unizar.es (J. Martínez), saavedra@uvigo.es (Á. Saavedra), pauli@constru.uniovi.es (P.J. García-Nieto), jpdiblasi@uvigo.es (J.I. Piñeiro), carlaiglesias@uvigo.es (C. Iglesias), jtaboada@uvigo.es (J. Taboada), jsanchov@unizar.es (J. Sancho), jjpastor@unizar.es (J. Pastor).

http://dx.doi.org/10.1016/j.amc.2014.05.004
0096-3003/© 2014 Elsevier Inc. All rights reserved.

the concentration of these pollutants provides instant records of harmful pollution that inform or alert local residents of a possible hazard. European Union and national environmental agencies have set standards and air quality guidelines for allowable levels of these pollutants in the air [5,6,9]. When the pollutant concentration levels exceed air quality guidelines, short–term and chronic human health problems may occur [10].

The source of pollutants such as historical industrial sites, mines, gas works, rubbish dumps, etc., may be known to local residents. These locations should be investigated to avoid or minimize potential risks. It is reasonable to assume that values for potentially polluted air samples behave as outliers in an urban environmental database. Outliers are observations that differ substantially from the rest of the data and can be detected by comparing the values in question with all the other values. They can be classified as local outliers [9,10] or global outliers. In comparison with global outliers, local outliers can be detected by comparing the values in question with neighboring values spatially located within a certain distance. For the purpose of polluted air investigation in urban areas, global high-value outliers exceeding the air quality guideline values indicate that a source should be further investigated. Observations which are not excessively high but still different from neighboring values may also contain information on unusual processes such as pollution.

A dataset may contain a small percentage of data objects (outliers) which are considerably dissimilar to the rest of the data based on some measurement. Outliers may merely be noisy observations. Alternatively, they may indicate abnormal behavior in the system. These abnormal values are very important and may lead to useful information or significant discoveries.

New techniques of functional data analysis (FDA) were developed for their application to vector problems. The appearance of this new methodology was motivated by the inefficiency of the classical data mining techniques to deal with the treatment of vector data [11]. FDA applications are so diverse that they were often used for environmental [12,13] and medical research [14,15], sensors [16], and industrial methods [17–19]. This functional model presents two important characteristics: firstly, it takes into account the time correlation structure of the data and secondly the comparisons are more general, leading to a global vision of the problem. The comparison of the curves is done by the application of the functional depth concept, a measure that represents the centrality of a given curve within a group of curves [20]. The functional depth has been used by the authors in different environmental problems [21,22].

The aim of this research was to construct a model to identify outliers in gas emissions in Langreo urban area and its surroundings (Principality of Asturias, Northern Spain). Many methods can be applied to identify outliers, but yet there is no universally agreed best method. This study was carried out in parallel with conventional methods and functional data analysis, and it is concluded with a comparative study between the two methodologies.

This innovative research work is structured as follows. Firstly, the necessary materials and methods to carry out this study are described; then, the obtained results are shown and discussed; finally, the main conclusions drawn from the results are exposed.

2. Materials and methods

2.1. Study area and dataset

Langreo is a municipality and town of Northern Spain, in the autonomous community of The Principality of Asturias. It is the 4th largest town in Asturias with 45,000 inhabitants and a density of 529.31 inhabitants per square kilometre. Land area is 82.46 km^2 being the maximum altitude of 1,021 m above sea level. Langreo is 22 km from Oviedo (Asturian capital city) and 35 km from Gijón. Fruit and cider are produced in this area, existing also important coal mines, foundries and factories for the manufacture of coarse cloth. It has been one of the most important mining and metallurgical points of Spain since the 18th century, being also well known because of workers struggles and its cultural life. What points out its importance in the past is the fact that the 3rd railway to be built in the Iberian Peninsula was the Langreo train. Due to the Spanish ‘‘Industrial Conversion’’, Langreo lost its industrial importance, although today it conserves important factories like Bayer, where 100% of the acetylsalicylic acid of the German enterprise is produced. Langreo is composed of the following most important districts: La Felguera, Sama, San Martín, Riaño, Ciaño, Lada and Barros.

The climate of Langreo, as with the rest of Northwest Spain, is more varied than in southern parts of Spain. Summers are generally humid and warm, with considerable sunshine but also some rain, whereas winters are cold and generally rainy, with some very cold spells, especially in the mountains surrounding the city, where snow is usually present from October to May.

The Lada Power Plant is a conventional cycle thermoelectric plant located next to the AS-117 road and to Nalón river, between the districts of Lada and La Felguera (both towns of Langreo). It consists of 2 groups of 350 and 175 thermal MW using coal as fuel, and it is owned by Iberdrola Ltd. Therefore, it provides most of the electrical energy used in Langreo urban area and is also a main source of its pollution. Nowadays, the only pollution caused by coal-fired power plants comes from gases (CO, NO, NO_2 and SO_2) released into the air. Acid rain is caused by emissions of nitrogen oxides and SO_2, which react in the atmosphere and create acidic compounds (such as sulfurous acid, nitric acid and sulfuric acid).

The Industry and Energy Department of the Government of Asturias has three automatic meteorological stations in the Langreo urban area, located in La Felguera, Sama and San Martín del Rey Aurelio, respectively (see Fig. 1), which measure the following primary and secondary pollutants every 15 min: CO, NO, NO_2, O_2, PM_{10} (particulate matter less than 10 μm) and

SO_2. This data are collected and processed daily and monthly on average. In this study, we used the data collected for the 72 months between January 2006 and December 2011. Therefore, the dataset used here was collected over 6 entire years (2006–2011).

2.2. Classical analysis

Effective monitoring strategy for early fault detection and diagnosis is very important from a safety and economic point of view [23]. The application of statistical process control methods can determine when significantly high measurements have been recorded. These methods can be applied to control individual measurements, x_j, using individuals charts, or to control means

$$\overline{x_j} = \frac{1}{n_2 - n_1 + 1}\sum_{j=n_1}^{n_2} x_j,$$

using average charts. Control charts are often implemented in statistical process control software packages and are used in conditioning monitoring for fault detection [24].

2.2.1. Classical analysis of outliers with control charts

Univariate and multivariate statistical processes control via control charting are a powerful technique for the interpretation of an out–of–control signal [25]. Control charts were initially developed by W. Shewhart in 1931 [26], with the aim of investigating if a process is under statistical control. The specific X–Bar and R charts he developed are also called Shewhart charts. The graphical interpretations of these graphs allow us to recognize when there is statistical evidence that excessively high measurements are being registered.

The classical analysis using control charts is developed in two stages: learning stage and control stage. In the learning stage the normality of the data should be tested in addition to checking the existence of discrepant measurements that might be eliminated of the database. The value of the center line is established with the control sample, setting then the warning limits at a distance ±2s from the central value n and control limits at a distance ±3s with s the standard deviation associated to the process. The coefficients for the construction of the warning and control limits are widely outlined in the literature [27].

In the control stage, measurements are plotted against time in the Shewhart charts in order to detect trends and situations out of control. Confirmation on whether the system is under statistical control is obtained by visual observation of the control charts. If the plotted points on the graph are distributed in a random pattern, the system is under statistical control. Control charts help to detect effects including the appearance of bias, presence of a progressive trend towards rising or falling values or periodic behaviors.

There are a number of well-known rules for assessing if a system is under statistical control. These rules can be found in Western Electric Corp [28].

2.2.2. Box–Cox transformation

The basis of control charts relies on the normality of the measurements. Under the assumption of normality the process is said to be out of control if there are points outside the control limits, or if trends are shown.

The power transformation is defined as a continuously varying function, with respect to a power parameter λ. This is a useful data transformation technique used to stabilize variance and make the data more normal distribution-like. The most widely used polynomial transformation is the Box–Cox transformation, defined as follows:

$$x_j^{(\lambda)} = \begin{cases} \frac{x_j^{\lambda}-1}{\lambda}, & \text{if} \quad \lambda \neq 0 \\ log(x_j), & \text{if} \quad \lambda = 0 \end{cases}$$

where λ is the value that maximize the profile likelihood function of the data x_j.

2.3. Functional analysis

2.3.1. Smoothing

In the first place, the functional model transforms the sample vector formed by experimental measurements into a final functional sample. This function is created from building the best-fitting curves from the initial discrete values. Thus, the model uses a set of continuous functions over time representing a set of observations instead of using a set of discrete points.

Functional data are discrete point observations of a random and continuous process in time [13,14]. A set of observations $x(t_j)$ in a set of n_p observation points, where $t_j \in \Re$ represents each time step, are considered as discrete observations of the function $x(t) \in \chi \subset F$, with F as a functional space. This function $x(t)$ is estimated taking into account that $F = span\left\{\phi_1, \ldots \phi_{n_b}\right\}$ is a functional space formed by a set of basis functions $\{\phi_k\}$, with $k = 1, 2, \ldots, n_b$ and n_b is the number of this basis functions necessary to form the functional space F. Normally, the basis functions used for this analysis are the splice or Fourier functions, but there are different types that can be used too. The expansion used is as follows [29–31]:

$$x(t)=\sum_{k=1}^{n_b}c_k\phi_k(t),$$

where $\{c_k\}_{k=1}^{n_b}$ are the coefficients of the function $x(t)$ with respect to the selected set of basis functions. Then, the smoothing problem is determined with the following regularization problem [30,31]:

$$\min_{x\in F}\sum_{j=1}^{n_p}\{z_j-x(t_j)\}^2+\lambda\Gamma(x),\tag{1}$$

where $z_j=x(t_j)+\epsilon_j$ is the observation value x in the point t_j with ϵ_j as the value of the zero–mean random noise, λ is a regularization parameter whose function adjusts the intensity of the regularization and the operator Γ is used to impose a penalty to the solution complexity. Accordingly Eq. (1) can be written into the following expression [31,32]:

$$\min_c\left\{(z-\Phi c)^T(z-\Phi c)+\lambda c^TRc\right\},$$

where $z=(z_1,\dots,z_{n_p})^T$ is the observation vector, $c=(c_1,\dots,c_{n_b})^T$ is the vector of coefficients of the functional expansion, Φ is the $n_p\times n_b$ matrix of elements $\Phi_{jk}=\phi_k(t_j)$ and R is the matrix formed by $n_b\times n_b$ elements [31]:

$$R_{kl}=\langle D^2\phi_k,D^2\phi_l\rangle_{L_S(T)}=\int_T D^2\phi_k(t)D^2\phi_l(t)dt,$$

where $D^n\phi_k(t)$ corresponds to the nth–order differential operator of the function ϕ_k. This problem can be expressed as follows:

$$c=\left(\Phi^t\Phi+\lambda R\right)^{-1}\Phi^Tz.$$

2.3.2. Functional depth concept

The concept of depth was established in multivariate analysis and is defined as a measure of the centrality of an observation in comparison with a set of observations, a cloud of points. Then, in a Euclidean space, where the observation presented by points can be distributed from the centre to the periphery, those points closer to the centre will have a greater depth. This definition has been extended to functional domain [30–32], where the depth concept is considered a measure of a curve x_i centrality with respect to a set of curves $x_1,\dots,x_n$.

Principal measures of functional depth are presented as follows:

- Fraiman–Muniz depth (FMD): where $F_{n,t}(x_i(t))$ is the cumulative empirical distribution function [20] for the values of the curves $\{x_i(t)\}_{i=1}^n$ in a time $t\in[a,b]$ ruled by the next expression [31]:

$$F_{n,t}(x_i(t))=\frac{1}{n}\sum_{k=1}^{n}I(x_k(t)\leqslant x_i(t))$$

where $I(\cdot)$ is the indicator function. Therefore, the FMD for a curve x_i with regard to set $x_1,\dots,x_n$ is estimated by:

$$FMD_n(x_i(t))=\int_a^b D_n(x_i(t))dt$$

where $D_n(x_i(t))$ is the depth of the point $x_i(t),\ \forall t\in[a,b]$ obtained by:

$$D_n(x_i(t))=1-\left|\frac{1}{2}-F_{n,t}(x_i(t))\right|$$

- H-modal depth (HMD): The functional mode definition is based on the mode concept, and it is defined as the curve most densely surrounded by other curves of the sample. HMD is exposed as follows [30–32]:

$$HMD_n(x_i,h)=\sum_{k=1}^{n}K\left(\frac{\|x_i-x_k\|}{h}\right)$$

where $K:R^+\to R^+$ is a kernel function, $\|\cdot\|$ is a norm in a functional space and h is the bandwidth parameter. L^2 is one of the most used norms for a functional space, given below:

$$\|x_i(t)-x_j(t)\|_2=\left(\int_a^b(x_i(t)-x_j(t))^2dt\right)^{1/2}.$$

Sometimes, the infinite norm can be used as follows:

$$\|x_i(t)-x_j(t)\|_\infty=\sup_{t\in(a,b)}|\,x_i(t)-x_j(t)\,|$$

For this research, the truncated Gaussian kernel was selected among all the different kernel functions $K(\cdot)$ due to its efficacy in functional outliers detection [30–32]:

$$K(t) = \frac{2}{\sqrt{2\pi}} exp\left(-\frac{t^2}{2}\right), \ t > 0$$

2.3.3. Functional outliers

A set of functional samples may contain elements which, although not erroneous in themselves, may have patterns different to the other elements in the set. Depth measurements used to identify outliers in functional samples enable sets of data observed over time and fitted to curves to be directly compared, rather than just the mean values for the measurement time interval.

An outlier is a sample with a lack of depth, meaning that depth and outliers are two opposed concepts. In order to identify the outliers, a search is conducted to find the curves with the greatest depths. In this study, HMD criterion was used, but the model was not substantially sensitive to the selection of h (we checked various values, such as 10 and 20, without the results varying). The bandwidth has been selected according to the standard methodology of the functional outliers [33] and C, the cut–off value, is selected according to [34], in such way that a maximum of 1% of correct observations were identified as outlier (type I error):

$$Pr(HMD_n(x_i(t) \leqslant C)) = 0.01, \ i = 1, \ldots, n$$

Among the different existing methods [30–32], the technique chosen to calculate the value of C is based on bootstrapping [30,32,34–36] because it is an unknown value a priori. The technique consists in bootstrapping original curves with a probability which is proportional to depth. This method is based on three main steps [31], starting with the extraction (with replacement) of a new sample from the original sample, with a resampling order of 10 selected. In this process, after the element extraction, each element is replaced and so it may be selected again. Secondly, based on the new sample, the population parameter of interest is derived on the basis of the construction of a statistic. Finally, the steps above are repeated using the results obtained, until a large number of estimates is collected.

3. Results and discussion

This study is carried out with a sample $\{x_{ij}\}_{j=1}^{72}$ which corresponds to the 72 months from January 2006 to December 2011, being x_{ij} the daily gas measurements taken the day i of the month j. In the case of functional analysis, each month has 30 days, that is to say, $i = 1, 2, \ldots, 30$.

Air quality in the Langreo urban area can seriously affect the health of its population. Indeed, air pollution is a significant risk factor for multiple health conditions including respiratory infections, heart disease, and lung cancer. In this sense, major primary pollutants produced due to the Lada coal–fired power plant (see Fig. 1) activity mainly include [5,6,10,37]:

- Sulfur dioxide (SO_2): produced by volcanoes and in various industrial processes as coal–fired power plants. Coal and petroleum often contain sulfur compounds, and their combustion generates sulfur dioxide. Further oxidation of SO_2, usually in the presence of a catalyst such as NO_2, forms H_2SO_4, and thus acid rain [5,6,10,37]. This is one of the causes for concern over the environmental impact of the use of these fuels as power sources. Its maximum allowable concentrations are 0.03 ppm (80 μg/m^3) in annual arithmetic mean and 0.14 ppm (365 μg/m^3) in 24–hour average according to the ambient air quality standards [5,6,10,37].
- Nitrogen dioxide (NO_2): expelled from high temperature combustion from the Lada coal–fired power plant (see Fig. 1). This gas is one of the most prominent air pollutants and also participates in the formation of acid rain as nitric acid (HNO_3) by reacting with water droplets from the atmosphere. Its maximum allowable concentration is 0.053 ppm (100 μg/m^3) in annual arithmetic mean according to the ambient air quality standards [5,6,10,37].
- Carbon monoxide (CO): a product by incomplete combustion of fuel such as natural gas, coal or wood. Vehicular exhaust is also a source of carbon monoxide. Its maximum allowable concentrations are 9 ppm (10 mg/m^3) in 8–hour average and 35 ppm (40 mg/m^3) in 1–hour average according to the ambient air quality standards [5,6,10,37].

Therefore, our analysis is focused in the study of the mean daily measurements of CO, NO_2 and SO_2 gases.

Next, the results obtained with both methodologies used for the air pollution analysis in the Langreo urban area are presented separately.

3.1. Results of classical analysis

Data collected by the monitoring stations are a sample vector $x(t_j) = x_j$ in a set of $n_p = 4$ points, where $t_j \in \Re$ stands for each time step. Experimental measurements, x_j, are discrete values of existing concentrations of certain pollutants in the air. These measurements, taken at regular intervals of time, conform the database through which, using statistical process control methods, can be determined at which moments significantly high pollutant measurements are recorded.

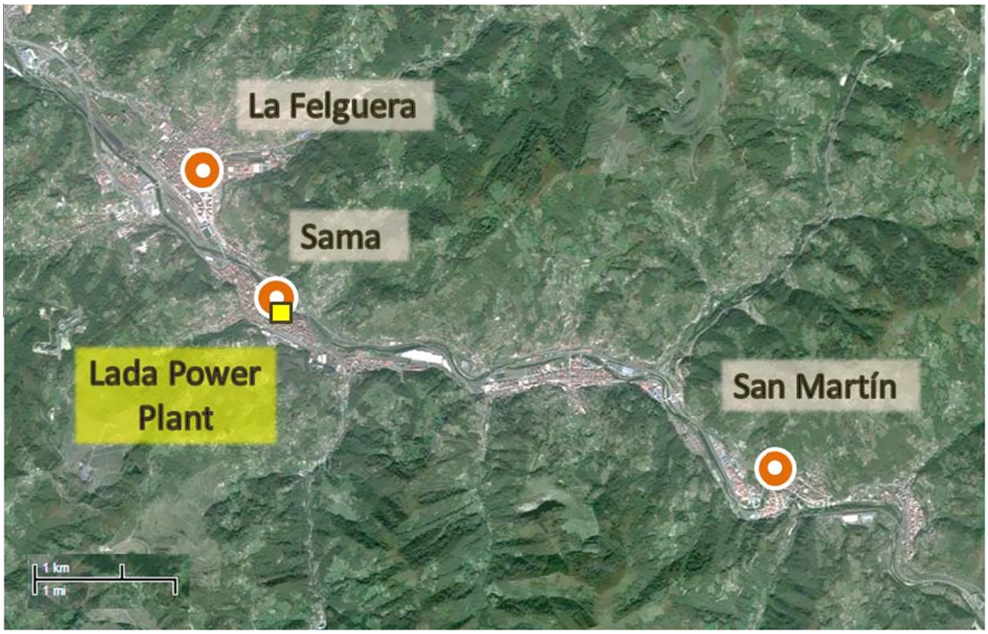

Fig. 1. Aerial photograph of the study area showing the location of the three automatic monitoring stations corresponding to the Langreo urban area and its surroundings.

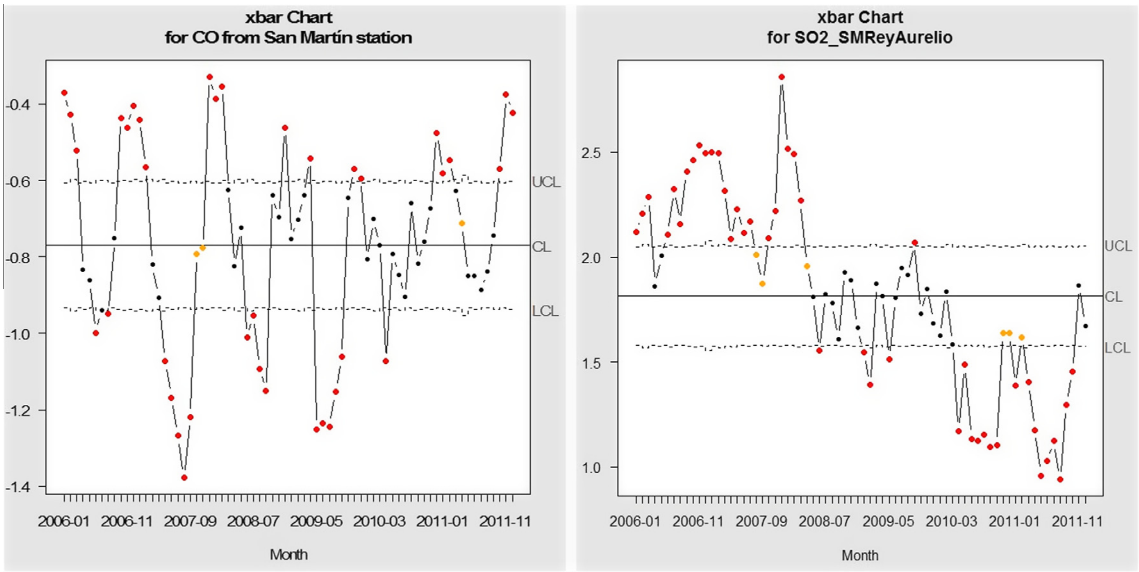

Fig. 2. X-bar charts for CO and SO_2 monthly averages recorded in San Martín del Rey Aurelio station.

In the learning stage it was observed the non-normality of the measurements, so we applied a Box–Cox transformation which ensured that the monthly averages of the concentrations verified the normality assumption. It should be noted that once the detection of outliers has been carried out, the study is considered to have reached its goal and then, no backtransformation is performed. If the aim of the study were the prediction of future measurements, then further post-processing of the data would be required: a backtransformation would be necessary in order to obtain predictions within the original space of sampling.

3.1.1. Checking the presence trends and periodic behaviors

Shewhart graphs corresponding to the control stage established the low stability of the averages of the pollutants. Fig. 2 shows the X–bar charts of the monthly averages of CO (left) and monthly averages of SO_2 (right) recorded in San Martín station (a total of 9 charts were obtained, Fig. 2 shows two of them as an example of the results).

In these figures, the colors of the dots are as follows:

- Black dots: the process is under statistical control, being the dots between the UCL (Upper Control Limit) and the LCL (Lower Control Limit).
- Yellow dots: the process is outside the warning limits.
- Red dots: these dots indicate that the process is not under statistical control, being outside the control limits UCL and LCL.

On the one hand, visual interpretation of these graphs indicates that, in both cases, there is a cyclical behavior which leads to values beyond the control limits. This periodic behavior was recorded in all stations and for all measurements of pollutants, although it is more acute in the case of CO. Thus higher values are recorded during the months between November and February while the lowest ones correspond to the months between May and August. On the other hand, as it can be seen in Fig. 2, there is a downward trend in the monthly averages of SO_2. Measurements of NO_2 also showed a downward trend, but not as marked as in the case of SO_2.

3.1.2. Study of outliers for the pollutants in the three stations using classical statistical analysis

Table 1 shows the months in which the average concentrations of the pollutants were above the control limits. These concentrations agree with the data displayed in Fig. 2, where the values outside the control limits are represented in red, taking into account that those concentrations below the control limits are not considered as pollution indicators and thus are not included in Table 1.

As set out in the previous section, many of these records are too high due to the trend and cyclical behavior recorded during the years under study.

3.2. Results of functional analysis

Taking the initial sample $\{x_{ij}\}_{j=1}^{72}$ and applying the smoothing process of the functional problem, a new sample $\{x_j(t)\}$ is obtained, where each x_j is now a function. A set of 100–element Fourier basis functions has been used in this analysis, obtaining a correlation coefficient of 0.99 between the discrete values of the initial sample and their corresponding values in the functions created.

The results obtained through the functional method are presented in Table 2, including every month where an abnormal content of a certain compound in any station has been detected.

3.3. Discussion

A wide number of outliers are detected with the classical analysis, even being identified as outliers more than half of the observations in many cases. This issue occurs independently of the station and the compound studied.

Regarding the outliers detected with the functional analysis, the number is considerably lower, as it can be stated in the previous Tables 1 and 2. It is also noted that the outliers detected with the functional analysis are among those identified with the classical method.

Table 1
Classical data analysis results.

Gas	Station	No. outliers	Months
CO	Sama	27	jan06, feb06, mar06, apr06, may06, jun06, jul06, aug06, sep06, oct06, nov06, dec06 jan07, feb07, dec07, mar08, dec09, jan10, feb10, mar10, apr10, nov10, dec10, jan11, feb11, mar11, dec11
	Felguera	26	jan06, feb06, mar06, apr06, may06, jun06, aug06, sep06, oct06, nov06, dec06, jan07, feb07, mar07, apr07, oct07, nov07, dec07, jan08, feb08, mar08, apr08, jun09, dec10, jan11, feb11
	San Martín	21	jan06, feb06, mar06, oct06, nov06, dec06, jan07, feb07, dec07, jan08, feb08, dec08, apr09, nov09, dec09, dec10, jan11, feb11, oct11, nov11, dec11
NO_2	Sama	24	jan06, feb06, mar06, apr07, may07, jun07, nov07, dec07, jan08, feb08, nov08, dec08, dec09, jan10, feb10, mar10, nov10, dec10, jan11, feb11, mar11, oct11, nov11, dec11
	Felguera	23	jan06, feb06, dec06, jan07, feb07, apr07, oct07, nov07, dec07, jan08, feb08, mar08, nov08, dec08, jan09, feb09, dec09, jan10, feb10, dec10, jan11, nov11, dec11
	San Martín	25	jan06, feb06, mar06, oct06, nov06, oct07, nov07, dec07, jan08, feb08, dec08, jan09, feb09, mar09, oct09, dec09, jan10, feb10, mar10, dec10, jan11, feb11, oct11, nov11, dec11
SO_2	Sama	18	sep06, oct06, jan07, feb07, apr07, jul07, feb08, sep08, oct08, nov08, dec08, jan09, feb09, dec09, nov10, dec10, nov11, dec11
	Felguera	25	jan06, feb06, mar06, may06, jun06, sep06, oct06, dec06, jan07, feb07, mar07, apr07, may07, jun07, oct07, nov07, dec07, jan08, feb08, mar08, nov08, dec08, jan09, dec09, dec10
	San Martín	24	jan06, feb06, mar06, jun06, jul06, aug06, sep06, oct06, nov06, dec06, jan07, feb07, mar07, apr07, may07, jun07, jul07, oct07, nov07, dec07, jan08, feb08, mar08, sep09

Table 2
Functional data analysis results of the dataset collected in the three stations situated in the Langreo urban area.

Gas	Station	No. outliers	Months
CO	Sama	5	feb06, jul06, aug06, sep06, feb10
	Felguera	3	feb08, aug08, jan09
	San Martín	9	dec06, jan07, mar07, nov07, dec07, jan08, feb08, apr08, jul08
NO_2	Sama	1	dec07
	Felguera	6	jan06, feb06, dec07, jan08, feb08, jan11
	San Martín	0	–
SO_2	Sama	4	sep06, aug08, dec08, jan09
	Felguera	1	feb06
	San Martín	3	jan07, dec07, jan08

This fact might be due to the non-normality of the data studied. The classical methods are very robust when the data are characterized by their normality, or when the sets studied have a known statistical distribution. However, when this method is applied to the study of sets which lack these characteristics, the efficacy decreases.

The detection of outliers in this case is complex since the classical methods study data as punctual observations. For the study of airborne pollution, the real pollution due to a certain compound is not dependant of a single measurement but of a series of measurements that exceed the limits in a certain time interval.

On the other hand, the functional analysis does not treat the data as punctual observations but as a temporal series, which is actually the proper approach to this problem. As it was explained, the aerial pollution due to a certain compound is given by a set of observations which, in a certain time interval, exceed the established limits instead of an individual measurement. This argumentation explains why the classical analysis identifies a high number of outliers compared to those obtained with the functional analysis. Furthermore, with this methodology the possible measurement errors, which are abnormal punctual measurements identified as outliers by the classical analysis, are obviated. Thus, the elimination of the instrumental error in a first stage makes the results more reliable.

Weather conditions explain why the curves created by the functional method were detected as outliers. Fig. 3 shows the average monthly temperature and overall monthly rainfall in the Langreo urban area between January 2006 and December 2011. Winter of 2006 was colder than that in previous years, what led to a greater consumption of electricity and heating, increasing pollutant emissions in this period. Specifically, the periods from November 2008 to March 2009, January 2010, and from December 2010 to January 2011, were characterized by cold winters with a lower level of precipitations. This combination is perfect to increase the concentration of pollutants in the air, since the rain and snow are able to collect gases and particulate matter during their gravitational fall in the atmosphere and therefore they can be considered as a natural clean-

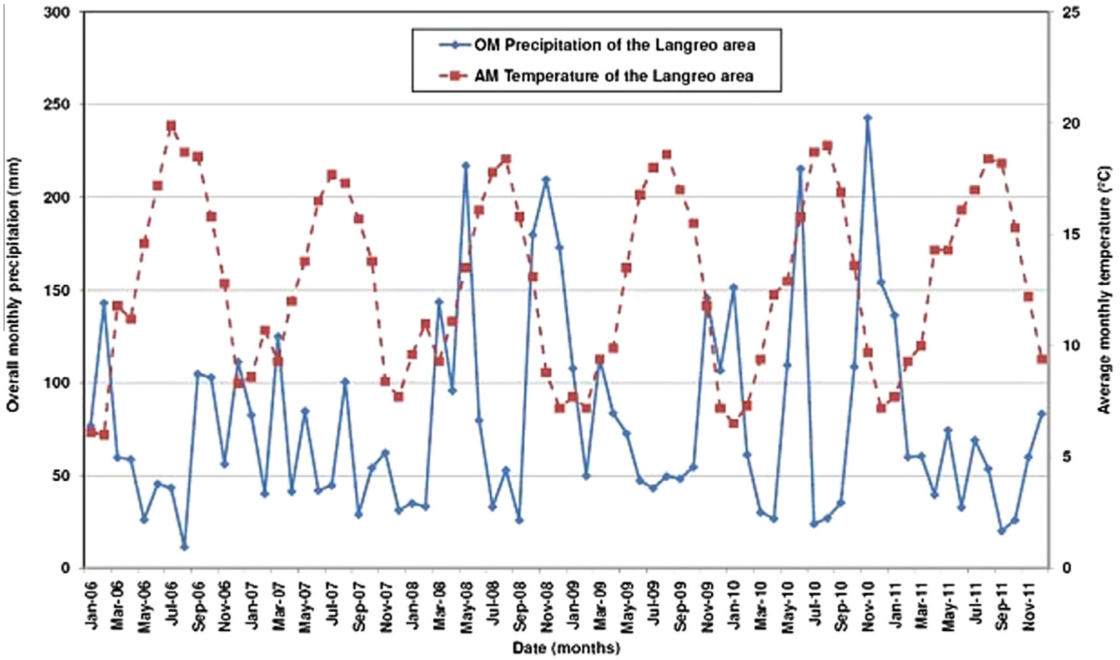

Fig. 3. Overall monthly rainfall and average monthly temperature for the Langreo urban area from January 2006 to December 2011.

ing agent. In similar way, these phenomena also cause that warmest periods between July 2006 and September 2006 or between July 2008 and August 2008 show a high level of pollutant gases.

4. Conclusions

Control charts are statistical tools simple to build and very useful for monitoring trends and stability of a process. However, the supervision of a statistician is needed both in the learning and control stages. When the aim is the identification of outliers, classical methods of quality control may be ineffective to study the presence of trends or the cyclical behavior of the measurements.

This research work presents a methodology for the detection of outliers from a functional point of view based on the concept of functional depth. Thus, the presence of outliers is determined among a sample of pollutants, allowing the establishment of conclusions regarding the air quality in the studied region.

Furthermore, this new methodology is compared with the classical methodologies used for the study of outliers, concluding that it is more powerful for the efficient determination of outliers, since the probability of detecting a measurement error as an outlier is lower.

Future research works are focused on the extension of the statistical study from a functional point of view with the introduction of scans rules to analyze trends and small shifts in functional samples. The possibility of applying this methodology to other type of problems is also under study.

Acknowledgments

The authors wish to acknowledge the computational support provided by the Departments of Mathematics at University of Oviedo and Natural Resources and Natural Resources and Environmental Engineering at the University of Vigo. They wish to acknowledge as well the Section of Industry and Energy from the government of Asturias (Spain) for supplying the pollutant data in the city of Oviedo. Furthermore, C. Iglesias is acknowledged to the Spanish Ministry of Education for her FPU 12-02283 grant. Additionally, the authors gratefully acknowledge the support from the Spanish Ministry of Science and Technology through research project ECO2011-22650, and Innovation projects MTM2008-03129 and MTM2011-23204 (FEDER support included).

References

[1] P.J. García Nieto, Parametric study of selective removal of atmospheric aerosol by coagulation, condensation and gravitational settling, Int. J. Environ. Health Res. 11 (2001) 151–162.
[2] A. Akkoyunku, F. Ertürk, Evaluation of air pollution trends in Istanbul, Int. J. Environ. Pollut. 18 (2003) 388–398.
[3] F. Karaca, O. Alagha, F. Ertürk, Statistical characterization of atmospheric PM10 and PM2.5 concentrations at a non-impacted suburban site of Istanbul, Turkey, Chemosphere 59 (8) (2005) 1 183–1 190.
[4] P.J. García Nieto, Study of the evolution of aerosol emissions from coal-fired power plants due to coagulation, condensation, and gravitational settling and health impact, J. Environ. Manage. 79 (4) (2006) 372–382.
[5] T. Godish, Air Quality, Lewis Publishers, Boca Raton, Florida, 2004.
[6] L.K. Wang, N.C. Pereira, Y.T. Hung, Air Pollution Control Engineering, Humana Press, New York, 2004.
[7] T. Elbir, A. Muezzinoglu, Evaluation of some air pollution indicators in Turkey, Environ. Int. 26 (1–2) (2000) 5–10.
[8] A.C. Comrie, J.E. Diem, Climatology and forecast modeling of ambient carbon monoxide in Phoenix, Atmos. Environ. 33 (1999) 5023–5036.
[9] C.D. Cooper, F.C. Alley, Air Pollution Control, Waveland Press, New York, 2002.
[10] F.K. Lutgens, E.J. Tarbuck, The Atmosphere: An Introduction to Meteorology, Prentice Hall, New York, 2001.
[11] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Springer, New York, 1997.
[12] J.M. Paruelo, F. Tomasel, Prediction of functional characteristics of ecosystems: a comparison of artificial neural networks and regression models, Ecol. Model. 98 (1997) 173–186.
[13] J.M. Matías, C. Ordóñez, J. Taboada, T. Rivas, Functional support vector machines and generalized linear models for glacier geomorphology analysis, Int. J. Comput. Math. 86 (2) (2009) 275–285.
[14] R. Viviani, G. Grön, M. Spitzer, Functional principal component analysis of FMRI data, Hum. Brain Mapping 24 (2) (2005) 109–129.
[15] D.A. Dombeck, M.S. Graziano, D.W. Tank, Functional clustering of neurons in motor cortex determined by cellular resolution imaging in awake behaving mice, J. Neurosci. 29 (44) (2009) 13751–13760.
[16] D. Wu, S. Huang, J. Xin, Dynamic compensation for an infrared thermometer sensor using least-squares support vector regression (LSSVR) based functional link artificial neural networks (FLANN), Meas. Sci. Technol. 19 (10) (2008) 105202.1–105202.6.
[17] M. López, J.M. Matías, J.A. Vilán, J. Taboada, Functional pattern recognition of 3D laser scanned images of wood–pulp chips, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 4477, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 298–305 (1).
[18] J.I. Park, S.H. Baek, M.K. Jeong, S.J. Bae, Dual features functional support vector machines for fault detection of rechargeable batteries, IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 39 (4) (2009) 480–485.
[19] J. Sancho, J.J. Pastor, J. Martínez, M.A. García, Evaluation of harmonic variability in electrical power systems through statistical control of quality and functional data analysis, Procedia Eng. 63 (2013) 295–302.
[20] R. Fraiman, R., G. Muniz, Trimmed means for functional data, Test 10 (2001) 419–440.
[21] J.I. Piñeiro, J. Martínez, P.J. García Nieto, J.R. Alonso, C. Díaz, J. Taboada, Analysis and detection of outliers in water quality parameters from different automated monitoring stations in the Miño river basin (NW Spain), Ecol. Eng. 60 (2013) 60–66.
[22] J. Sancho, J. Martínez, J.J. Pastor, J. Taboada, J.I. Piñeiro, P.J. García Nieto, New methodology to determine air quality in urban areas based on runs rules for functional data, Atmos. Environ. 83 (2014) 185–192.
[23] N.S. Che Din, N.A.F. Abdul Samad, S.Y. Chin, Fault detection and diagnosis for gas density monitoring using multivariate statistical process control, J. Appl. Sci. 11 (13) (2011) 2400–2405.
[24] T. Friebel, R. Haber, Detection of signal drifts by different control charts, in: IFAC Proceedings, (IFAC-Papers On line) 2 (PART 1), 2009.
[25] S. Bersimis, S. Psarakis, J. Panaretos, Multivariate statistical process control charts: an overview, Qual. Reliab. Eng. Int. 23 (5) (2007) 517–543.

[26] W. Shewhart, The Economic Control of Quality of Manufactured Products, D. Van Nostrand, New York, 1931.

[27] E.L. Grant, R.S. Leavenworth, Statistical Quality Control, McGraw-Hill, New York, 1998.

[28] Western Electric Corp., Statistical Quality Control Handbook, in: AT&T Technologics, Indianapolis, 1956.

[29] I. Alameddine, M.A. Kenney, R. Gosnell, K.H. Reckhow, Robust multivariate outlier detection methods for environmental data, J. Environ. Eng. 136 (11) (2010) 1 299–1 304.

[30] J. Martínez Torres, P.J. García Nieto, L. Alejano, A.N. Reyes, Detection of outliers in gas emissions from urban areas using functional data analysis, J. Hazard. Mater. 186 (2011) 144–149.

[31] C. Díaz Muñiz, P.J. García Nieto, J.R. Alonso Fernández, J. Martínez Torres, J. Taboada, Detection of outliers in water quality monitoring samples using functional data analysis in San Esteban estuary (Northern Spain), Sci. Total Environ. 439 (2012) 54–61.

[32] A. Cuevas, M. Febrero-Bande, R. Fraiman, On the use of the bootstrap for estimating functions with functional data, Comput. Stat. Data Anal. 51 (2006) 1063–1074.

[33] A. Cuevas, R. Fraiman, A plug-in approach to support estimation, Ann. Stat. 25 (6) (1997) 2 300–2 312.

[34] M. Febrero-Bande, P. Galeano, W. González-Manteiga, A functional analysis of NOx levels: location and scale estimation and outlier detection, Comput. Stat. 22 (3) (2007) 411–427.

[35] M. Febrero-Bande, P. Galeano, W. González-Manteiga, Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels, Environmetrics 19 (4) (2008) 331–345.

[36] L. Peng, Y. Qi, Bootstrap approximation of tail dependence function, J. Multivariate Anal. 99 (8) (2008) 1 807–1 824.

[37] G.R. Visgilio, D.M. Whitelaw, Acid in the Environment: Lessons Learned and Future Prospects, Springer, New York, 2007.