Contents lists available at ScienceDirect

Journal of Hazardous Materials

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

Detection of outliers in gas emissions from urban areas using functional data analysis

J. Martínez Torres[a], P.J. Garcia Nieto[b,*], L. Alejano[c], A.N. Reyes[c]

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

ARTICLE INFO

Article history:
Received 12 June 2010
Received in revised form
29 September 2010
Accepted 24 October 2010
Available online 2 November 2010

Keywords:
Functional data analysis
Outliers
Air pollution
Gas emissions
Functional depth

ABSTRACT

In this work a solution for the problem of the detection of outliers in gas emissions in urban areas that uses functional data analysis is described. Different methodologies for outlier identification have been applied in air pollution studies, with gas emissions considered as vectors whose components are gas concentration values for each observation made. In our methodology we consider gas emissions over time as curves, with outliers obtained by a comparison of curves instead of vectors. The methodology, which is based on the concept of functional depth, was applied to the detection of outliers in gas omissions in the city of Oviedo and results were compared with those obtained using a conventional method based on a comparison of vectors. Finally, the advantages of the functional method are reported.

1. Introduction

Air pollution is an important environmental problems in cities [1–4]. Air is never perfectly clean [5] and polluted air is a continuing threat to human health and welfare [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. Clean air should certainly be as important to us as clean water and food.

There are a number of sources of air pollution that affects human health [1,7]. Information on meteorological pollution, such as that produced by carbon monoxide (CO), nitrogen oxides (NO and NO_2), sulphur 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 concentrations of these pollutants provide 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 that can be detected by comparing the values in question with all the other values. They can be classified as local outliers [11] or global outliers. In comparison with global outliers, local outliers can be detected by comparing the values in question with neighbouring 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 neighbouring 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 behaviour in the system. These abnormal values are very important and may lead to useful information or significant discoveries.

The aim of this research was to construct a model to identify spatial outliers in gas emissions in Oviedo, a city located in northwest Spain. Many methods can be applied to identifying outliers,

[*] Corresponding author. Tel.: +34 985 103417; fax: +34 985 103354.
E-mail address: lato@orion.ciencias.uniovi.es (P.J.G. Nieto).

0304-3894/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhazmat.2010.10.091

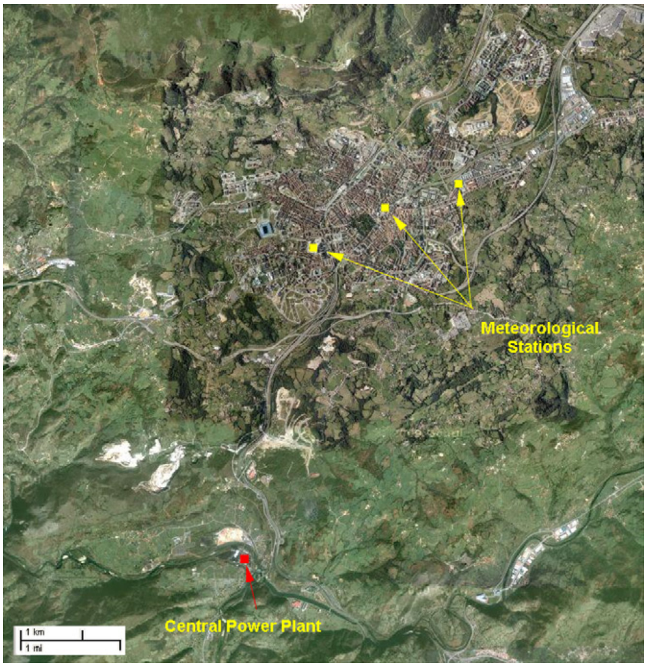

Fig. 1. Photograph of the study area showing the location of the metereological stations in the city of Oviedo and the coal-fired power plant.

but as yet there is no universally agreed best method. In this study, the method of the functional data analysis was applied.

This innovative research work is structured as follows. In the first place, the necessary materials and methods are described to carry our this study. Next the obtained results are shown and discussed. Finally, the main conclusions drawn from the results are exposed.

2. Materials and methods

2.1. Data

The data used for the functional data analysis used to detect outliers were collected over three years (2006–2008) from three metereological stations located in the city of Oviedo, capital of the Principality of Asturias in northern Spain and part of the municipality of the same name which is the administrative and commercial centre of the region.

The city of Oviedo has a population of 221,202 inhabitants, for a density of 1185.12 inhabitants per square kilometre. Land area is 186.65 km^2 and it is 232 m on average above sea level.

The climate of Oviedo, 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. 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 Soto de Ribera coal-fired power plant, lying 7 km south of the city (Fig. 1), provides most of the electrical energy used in Oviedo 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 sulphurous acid, nitric acid and sulphuric acid).

The industry and energy department of the government of Asturias has three meteorological stations in the city of Oviedo (Fig. 1), which measure the following primary and secondary pollutants every 15 min: CO, NO, NO_2, O_3, PM measuring less than 10 μm (PM_{10}) and SO_2. This data for the entire city is collected and processed once a month on average. In this study we used the data collected for the 36 months between January 2006 and December 2008.

Fig. 2 shows the concentrations of the different gases measured during the period of the study. It can be observed that the emission peaks occurred in late autumn and early winter (November to February) each year. Maximum emissions (51.20 g/m^3) occurred during the Christmas period of 2006, and minimum emissions (13.17 μg/m^3) in August 2007. This trend is general throughout the years studied, and reflects the higher electricity consumption of certain winter months and the lower consumption and reduced traffic of the summer holiday period. From the point of view of air quality standards, according to the US Environmental Protection Agency (EPA), the maximum allowable concentration of SO_2 expressed as an annual arithmetic mean is 80 μg/m^3. In our study, the annual arithmetic means for this gas in 2006, 2007 and 2008 were 24.0, 23.27 and 24.31 μg/m^3 respectively. Emissions were therefore below the maximum and complied with air quality standards during the three years, including at emission peaks.

2.2. Constructing curves from points: smoothing

Functional data are observations of a random continuous process observed at discrete points [12]. Given a set of observations $x(t_j)$ in a set of n_p points $t_j \in \mathbb{R}$, where t_j represents each instant of time, all the observations can be considered as discrete observations of the function $x(t) \in \chi \subset F$, where F is a functional space. In order to estimate the function $x(t)$ it is considered that $F = span\{\phi_1, \ldots, \phi_{n_b}\}$, where $\{\phi_k\}k = 1, \ldots, n$ is a set of basis functions. In view of this expansion:

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

where $\{c_k\}_{k=1}^{n_b}$ represent the coefficients of the function $x(t)$ with respect to the chosen set of the basis functions. The smoothing problem consists of solving the following regularization problem:

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

where $z_j = x(t_j) + \varepsilon_j$ (with ε_j as random noise with zero mean) is the result of observing x at the point t_j, Γ is an operator that penalizes the complexity of the solution and λ is a regularization parameter that regulates the intensity of the regularization.

Bearing in mind the expansion in Eq. (1), the above problem may be written as:

$$\min_{\mathbf{c}} \{(\mathbf{z} - \mathbf{\Phi c})^T (\mathbf{z} - \mathbf{\Phi c}) + \lambda \mathbf{c}^T \mathbf{Rc}\} \tag{3}$$

where $\mathbf{z} = (z_1, \ldots, z_{n_p})^T$ is the vector of observations, $\mathbf{c} = (c_1, \ldots, c_{n_b})^T$ is the vector of coefficients of the functional expansion, $\mathbf{\Phi}$ is the $n_p \times n_b$ matrix with elements $\mathbf{\Phi}_{jk} = \phi_k(t_j)$, and $\mathbf{R}$ is the $n_b \times n_b$ matrix with elements:

$$R_{kl} = \langle D^2\phi_k, D^2\phi_l \rangle_{L_2(\mathbf{T})} = \int_{\mathbf{T}} D^2\phi_k(t) D^2\phi_l(t) dt \tag{4}$$

The solution to this problem is given by:

$$\mathbf{c} = (\mathbf{\Phi}^t \mathbf{\Phi} + \lambda \mathbf{R})^{-1} \mathbf{\Phi}^t \mathbf{z} \tag{5}$$

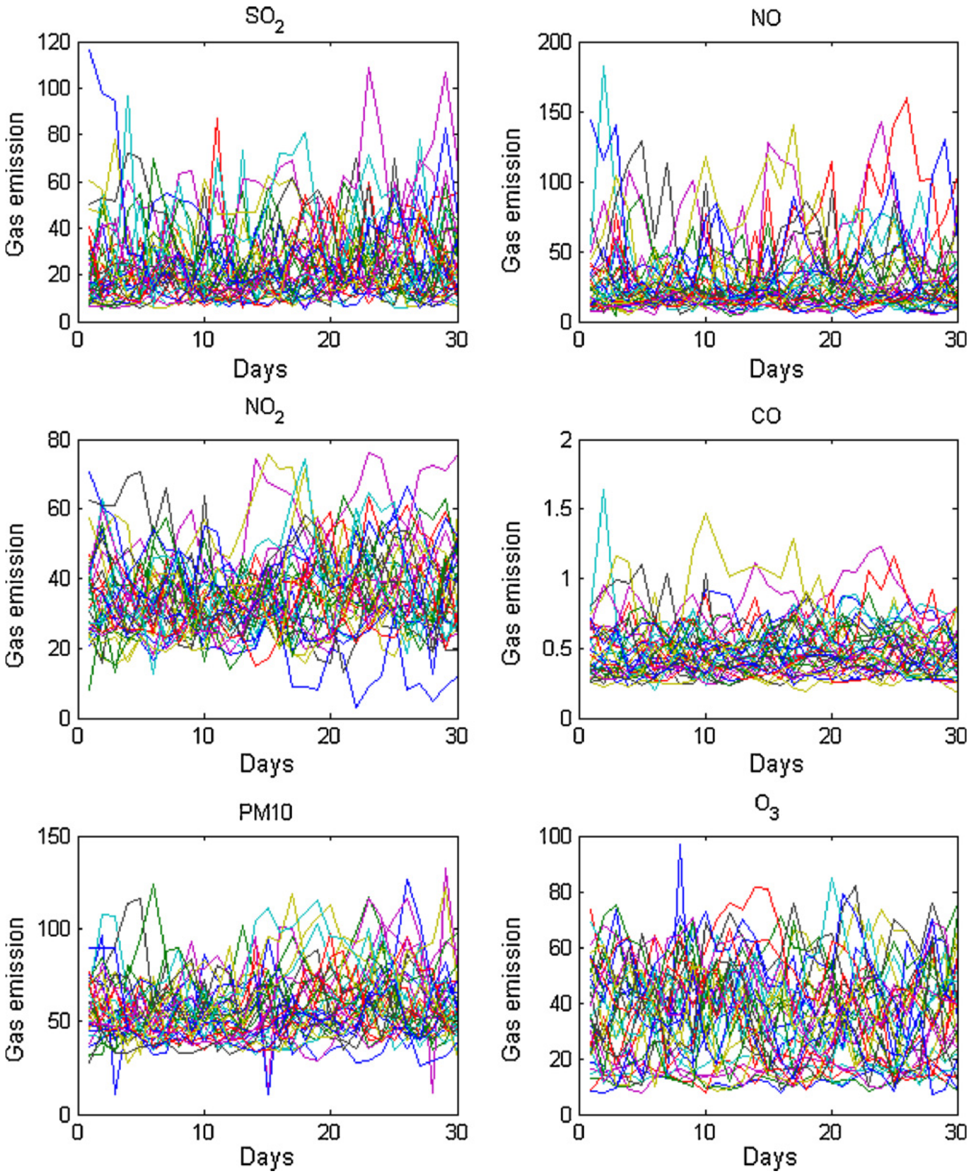

Fig. 2. Emissions of different gases during the 3-year study period.

2.3. The functional depth concept

Depth measurement was introduced originally in multivariate analysis in order to measure the centrality of a point with respect to a cloud of points. Depth provides a way of ordering points in a Euclidian space from the centre to the periphery, in such a way that the points closer to the centre will have greater depth. The notion of depth has recently been extended to functional data [13,14]. Functional depth measures the centrality of a curve x_i within a set of curves $x_1, \ldots, x_n$.

The most popular depth measurements are described below.

- Fraiman–Muniz depth (FMD): let $F_{n,t}(x_i(t))$ be the cumulative empirical distribution function [13] for the values of the curves $\{x_i(t)\}_{i=1}^{n}$ in an instant of time $t \in [a, b]$ given by:

$$F_{n,t}(x_i(t)) = \frac{1}{n}\sum_{k=1}^{n} I(x_k(t) \le x_i(t)) \tag{6}$$

where $I(\cdot)$ is the indicator function. The FMD for a curve x_i with respect to the set $x_1, \ldots, x_n$ is given by:

$$\mathrm{FMD}_n(x_i(t)) = \int_a^b D_n(x_i(t))dt \tag{7}$$

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

$$D_n(x_i(t)) = 1 - \left|\frac{1}{2} - F_{n,t}(x_i(t))\right| \tag{8}$$

- H-modal depth (HMD): the functional mode (which is based on the mode concept) is defined as the curve most densely surrounded by the other curves in the sample. HMD [14] is expressed as:

$$MD_n(x_i, h) = \sum_{k=1}^{n} K\left(\frac{||x_i - x_k||}{h}\right) \tag{9}$$

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

$$||x_i(t) - x_j(t)||_2 = \left(\int_a^b (x_i(t) - x_j(t))^2 dt \right)^{1/2} \tag{10}$$

The infinite norm L^∞ is sometimes used:

$$||x_i(t) - x_j(t)||_\infty = \sup_{t \in (a,b)} |x_i(t) - x_j(t)| \tag{11}$$

Different kernel functions $K(\cdot)$ can also be defined, among them the truncated Gaussian kernel [14]:

$$K(t) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right), \quad t > 0 \tag{12}$$

2.4. Functional outliers

A functional sample set may have elements that, although they do not incorporate error in themselves, may feature patterns different from the others. The depth measurements described above, used to identify outliers in functional samples, enables sets of observations over time fitted to curves to be compared, rather than just the mean values in the measurement time interval.

Depth and outlier are inverse concepts; hence, an outlier for a functional sample will have considerably less depth. The curves with the greatest depths are sought in order to identify functional outliers.

HMD was used to generate the outlier selection criterion selecting the value of bandwidth h as the 15th percentile of the empirical distribution $\{||x_i - x_j||_2\ i, j = 1, \ldots, n\}$ [15]. The cut-off C was selected in such a way that the percentage of correct observations poorly identified as outliers (type I error) was approximately 1% [16]:

$$\Pr(MD_n(x_i(t)) \le C) = 0.01, \quad i = 1, \ldots, n \tag{13}$$

Unfortunately, the distribution of the chosen functional depth is not known and so the value for C had to be estimated. Of the different approaches to estimating this value [14], we chose, for the purpose of this research, a method based on bootstrapping [14,16–18] the curves of the original set with a probability proportional to depth. The bootstrapping approach can be summarized as follows:

1. A new sample is extracted from the original sample by means of sampling with replacement (in other words, each extracted element is replaced after extraction and so it may be selected again). Furthermore, a resampling of order 10 has been selected.
2. Based on this new sample, the populational parameter of interest is estimated on the basis of the construction of a statistic.
3. The two steps above are repeated until a large number of estimates are obtained.
4. Finally, the empirical distribution of the statistic is determined.

2.5. Outlier detection by standard methods

A classical approach to comparing measures that is also used to detect outliers, is the Z-score, which is determined by:

$$z_i = \frac{x_i - x_r}{\sigma_r} \tag{14}$$

where x_i is the mean value of the observations by each point i, x_r is a reference value (the mean value for all the measures) and σ_r is the reference standard deviation (standard deviation for all the measures). Reference values are obtained by calculating the mean and standard deviation for all the observations regardless of their location.

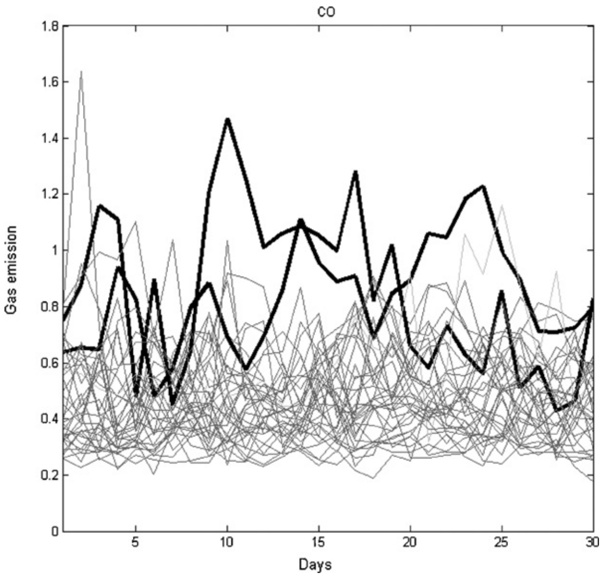

Fig. 3. CO gas emissions (in mg/m^3) with two functional outliers.

The interpretation criterion, according to the the International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) guidelines [19], is as follows:

$$\begin{aligned} &z_i \le 2 \rightarrow \text{Satisfactory} \\ &2 \le z_i \le 3 \rightarrow \text{Uncertain} \\ &z_i > 3 \rightarrow \text{Unsatisfactory} \end{aligned} \tag{15}$$

This means that possible outliers are those with a Z-score equal to or greater than two, with a Z-score of three indicating clear candidate outliers. The Z-score criterion assumes that the distrution of pollutant levels is Gaussian.

3. Results

For the studied gases, our sample $\{\boldsymbol{x}_{ij}\}_{j=1}^{36}$ corresponded to the 36 months between January 2006 and December 2008, where x_{ij} is the gas emission measurement for day i of month j, $i = 1, 2, \ldots, 30$. After the smoothing as described earlier, a sample $\{x_j\}_{j=1}^{36}$ is obtained where each x_j is now a function, considering a set of basis functions with 1000 elements. This calculation gives place to a correlation coefficient of 99% between the discrete values and the observations of each function in the corresponding points. In this sense, the generated functional sample is correlated with the discrete sample by 99%.

The functional analysis detected just two functional outliers, for the CO gas. Fig. 3 shows the sample for CO gas emissions, indicating the two outliers, corresponding to the two months of December 2006 and January 2007.

Weather conditions explain why these curves were detected as outliers. Fig. 4, which shows average monthly temperatures and rainfall in the city of Oviedo between January 2006 and December 2008, reveals that the temperatures for December 2006 and January 2007 were lower than in other years. This led to a greater consumption of electricity and heating, which, in turn, increased pollutant emissions in this period.

The histograms resulting from the application of the Z-score criterion to each of the gases. Unlike what happened with the functional analysis, it can be observed that the Z-score criterion indicated no outlier in the 36 study months for any gas, as most scores were one and none exceeded two. This is because this outlier

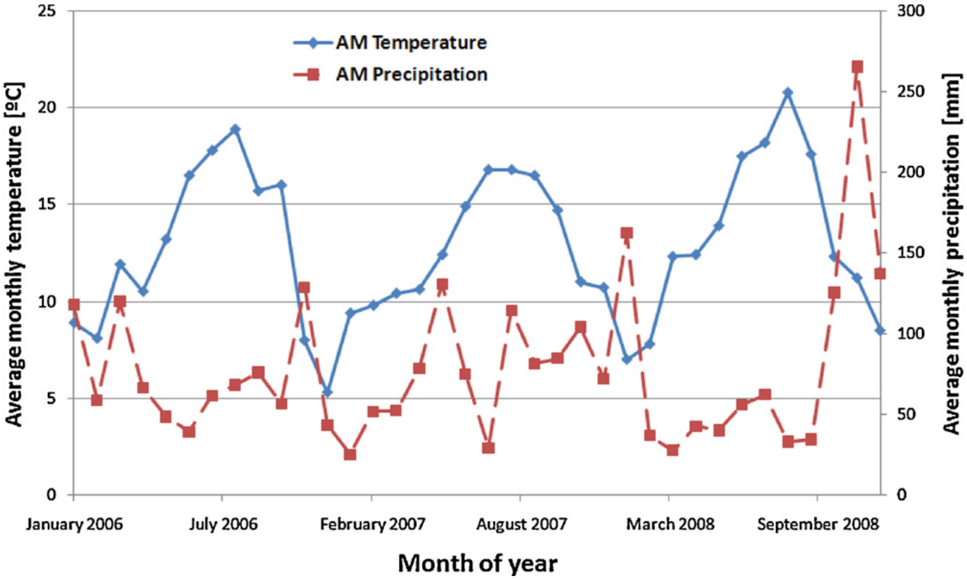

Fig. 4. Average monthly temperature and rainfall for the city of Oviedo from January 2006 to December 2008.

detection method is overly simple, based as it is on a comparison of means and failing to take into account data distribution over time. Furthermore, the statistics are affected by the outliers and the method is based on the mistaken assumption that concentrations for each month follow a Gaussian distribution.

4. Conclusions

To fix ideas, a sample of gas emissions for the city of Oviedo (Spain) has been analysed to detect outliers with success. In this research work, a classical vector technique (as used in European standards) and an innovative functional technique that treats the data as functions are used. The functional approach has the advantage of enabling more information to be recovered from the data than in the vector approach, which compares means and is unable to account for temporal variations. Furthermore, the vectorial technique assumes that the distribution of observations is normal, which is not always the case. The functional technique does not assume any kind of statistical distribution for the data and takes also into account the time correlation structure.

The results obtained revealed outliers in CO emissions for three months of the study period. A plausible explanation is that these corresponded to days when temperatures were low, leading to a greater consumption of energy. In fact, the proposed functional technique identifies two outliers while the standard vector method identifies no outliers. Fig. 4 above reveals the validity of the functional procedure in the evaluation and detection of outliers compared with the classical vector technique. The huge consumption of electricity and heating gives place to increasing pollutant emissions in this period and implies robust evidences that demonstrate that the result of the proposed innovative method is the correct one. The analysis of the outlier data using the functional technique provides an estimation of the common dispersion related to every concentration level in the studied range [15–17].

Outlier detection by functional data analysis can be used to evaluate polluted air in urban areas. As a general rule, emissions of pollutants increase in line with economic and demographic growth and decline during economic downturns. In this paper, a new methodology to find automatically outliers in the gas emissions set was developed. Our methodology can be applied to other cities with similar or different sources of pollutants, but it is always necessary to take into account the specificities of each location.

Acknowledgements

The authors wish to acknowledge the computational support provided by the Departments of Mathematics at University of Oviedo and Natural Resources and Environmental Engineering at Vigo University as well as pollutant data in the city of Oviedo supplied by the Section of Industry and Energy from the government of Asturias (Spain). We would like to thank Anthony Ashworth for his revision of English grammar and spelling of the manuscript.

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 PM_{10} and PM2.5 concentrations at a non-impacted suburban site of Istanbul, Turkey, Chemosphere 59 (8) (2005) 1183–1190.
[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] G. Lalor, C.S. Zhang, Multivariate outlier detection and remediation in geochemical databases, Sci. Total Environ. 281 (2001) 99–109.
[12] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Springer, New York, 1997.
[13] R. Fraiman, G. Muniz, Trimmed means for functional data, Test 10 (2001) 419–440.
[14] A. Cuevas, M. Febrero, R. Fraiman, On the use of the bootstrap for estimating functions with functional data, Comput. Stat. Data Anal. 51 (2006) 1063–1074.
[15] A. Cuevas, R. Fraiman, A plug-in approach to support estimation, The Annals of Statistics 25 (6) (1997) 2300–2312.

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

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

[18] L. Peng, Y. Qi, Bootstrap approximation of tail dependence function, J. Multivariate Anal. 99 (8) (2008) 1807–1824.

[19] ISO/IEC Guide 43-1, Proficiency Testing by Interlaboratory Comparisons. Part 1: Development and Operation of Proficiency Testing Schemes, 1997.