proceedings

Proceedings

Detection of Outliers in Pollutant Emissions from the Soto de Ribera Coal-Fired Plant Using Functional Data Analysis: A Case Study in Northern Spain [†]

Fernando Sánchez Lasheras [1,*], Celestino Ordóñez Galán [2], Paulino José García Nieto [1] and Esperanza García-Gonzalo [1]

[1] Department of Mathematics, University of Oviedo, 33007 Oviedo, Spain; pjgarcia@uniovi.es (P.J.G.N.), espe@uniovi.es (E.G.-G.)

[2] Department of Prospection and Mining Exploitation, University of Oviedo, 3004 Oviedo, Spain; ordonezcelestino@uniovi.es

[*] Correspondence: sanchezfernando@uniovi.es; Tel.: +34-985-10-3338

[†] Presented at the 2nd International Research Conference on Sustainable Energy, Engineering, Materials and Environment (IRCSEEME), Mieres, Spain, 25–27 July 2018.

Published: 5 November 2018

Abstract: The present research uses two different functional data analysis methods called functional high-density region (HDR) boxplot and functional bagplot. Both methodologies were applied for the outlier detection in the time pollutant emissions curves that were built using as inputs the discrete information available from an air quality monitoring data record station. Although the record of pollutant emissions is made in a discrete way, these methodologies consider pollutant emissions over time as curves, with outliers obtained by a comparison of curves instead of vectors. Then the concept of outlier passes from been a point to a curve that employed the functional depth as the indicator of curve distances. In this study, the referred methodologies are applied to the detection of outliers in pollutant emissions from the Soto de Ribera coal-fired plant which is in the nearby of the city of Oviedo, located in the Principality of Asturias, Spain. Finally, the advantages of the functional method are reported.

Keywords: functional data analysis; outlier detection; air pollution; gas emissions; functional bagplot; functional high-density region (HDR) boxplot

1. Introduction

Air pollution is an important environmental problem in cities [1,2]. There are a number of sources of air pollution that affects human health [1]. 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), is increasingly important due to the harmful effects on human health [3]. Automated measurement of concentrations of these pollutants provide instant records of harmful pollutants. When the pollutant concentration levels exceed air quality guidelines, short-term and chronic human health problems may occur [4].

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.

The aim of this research was to identify outliers in pollutant emissions from the Soto de Ribera coal-fired power plant affecting directly to Oviedo, a city located in northwest Spain with 221,870 inhabitants. Many methods can be applied to identifying outliers, but as yet there is no universally agreed best method. In this study, the method of the functional data analysis was applied.

2. Materials and Methods

2.1. Sources and Types of Air Pollutants

An air pollutant is a substance that can be unhealthy for humans and the environment. Pollutants can be found in the form of solid particles, liquid droplets, or gases. They may be man-made or natural and can be classified as primary or secondary. Mostly, primary pollutants come from a process, such as carbon monoxide from a motor vehicle exhaust, sulfur dioxide from factories, or ash from a volcanic eruption. Secondary pollutants form in the air when primary pollutants interact or react, and therefore, they are not emitted directly. For instance, an important secondary pollutant is ground-level ozone, which is one of the many secondary pollutants which make up photochemical smog [5]. Some pollutants can be both primary and secondary, that is, they have been both emitted directly and formed from other primary pollutants.

2.2. Study Area and Dataset

The dataset used for the functional data analysis to detect outliers were collected over five years (from 2013 to 2017) from Santa Marina air quality monitoring station located around the Soto de Ribera coal-fired power plant and close to the city of Oviedo. Every 60 minutes' measurements were taken of the following primary and secondary pollutants: SO_2, nitrogen oxides (NO and NO_2), CO, PM_{10}, and O_3.

2.3. Mathematical Model

2.3.1. Building Curves from Points: Smoothing

The first stage to solve the problem with the methodology proposed is the generation of the functional sample from the vector sample, that is to say, to build best-fitting curves with the points corresponding to the discrete values from the experimental measurements. In this way, we do not work with the set of observations as multivalued vectors, but with a set of observations considered as continuous functions over time.

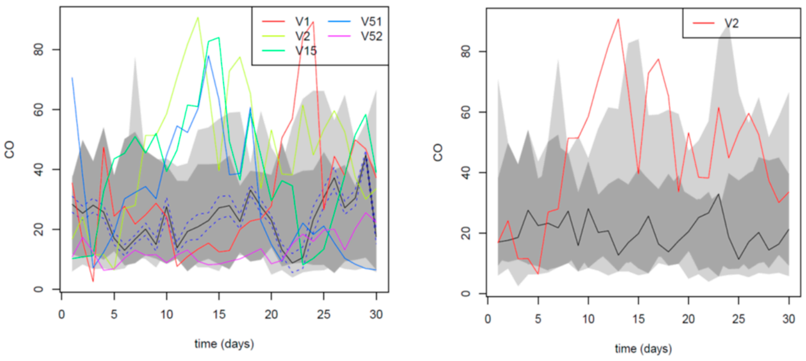
2.3.2. Functional Bagplot and Functional High-Density Region (HDR) Boxplot

Functional bagplots are an extension of the bivariate bagplots [6]. They are applied to the first two robust principal component scores. These bagplots have an advantage over other functional methods that detect outliers, such as those based on the concept of functional depth, in that it can identify both magnitude and shape outliers.

A bivariate bagplot is a generalization of the univariate boxplot. It is constructed on the basis of the halfspace location depth of a point relative to a bivariate dataset [7]. The bivariate bagplot is similar to a univariate boxplot in that it has a central point (the depth median), an inner region (the bag), and an outer region (the fence) beyond which outliers are shown as individual points. In a bagplot, the depth median (the point with highest halfspace depth) lies in the center and is surrounded by the bag, which contains 50% of the observations with the greatest depth (analogous to the inter-quartile range in a classical boxplot). The fence is obtained by magnifying the bag by a factor p. Observations outside the fence are flagged as outliers. It is also possible to visualize a confidence region for the depth median in the bagplot. In addition to the outliers, the bagplot allows observation of the characteristics of the data, such as location, spread, correlation, skewness, and tails. A variant of the functional boxplot, the functional HDR boxplot [8] orders the scores by means of a kernel bivariate density estimation.

3. Results and Discussion

For the studied pollutants, our sample corresponded to 60 months between January 2013 and December 2017. Figure 1 depicts the functional bagplot and HDR boxplot corresponding to the CO samples. The functional bagplot and HDR boxplots allows us to define the presence of outliers from a statistical point of view as curves that do not fall inside the 95% confidence bands (represented in light gray) indicated in Figure 1. Please note that CO has been selected as is the pollutant with the largest number of outliers.



(**a**) (**b**)

Figure 1. (**a**) Functional bagplot analysis of the CO curve; (**b**) HDR boxplot analysis of the CO curve.

The outlier detection results obtained from the two proposed functional methods are presented in Table 1, including every month where an abnormal concentration of a certain pollutant in Santa Marina air quality monitoring station has been detected during the 5-year study period (2013–2017). In this case, the functional HDR boxplot method performs the worst outlier detection between the two methods.

Table 1. Summary of the outliers detected by the two functional procedures from the dataset collected in the Santa Marina air quality monitoring station corresponding to Soto de Ribera coal-fired power plant pollutants emission (years from 2013 to 2017).

Pollutant/Functional Method	No. Outliers	Months
SO2/fbagplot	4	13 January, 13 February, 14 February, 14 March
SO2/HDR	1	13 February
NO/fbagplot	4	13 January, 13 February, 13 December, 14 February
NO/HDR	1	13 January
NO_2/fbagplot	3	13 February, 14 March, 17 February
NO_2/HDR	1	13 February
CO/fbagplot	5	13 January, 13 February, 14 March, 17 March, 17 April
CO/HDR	1	13 February
PM_{10}/fbagplot	3	16 February, 16 December, 17 January
PM_{10}/HDR	1	16 February
O_3/fbagplot	3	13 January, 13 October, 17 April
O_3/HDR	1	13 February

4. Conclusions

In this study, we have used a novel functional technique that treats the discrete data as continuous functions as a function of time. The functional approach has the advantage of enabling more information to be recovered from the data than in the vectorial 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 also takes the time correlation structure into account.

To fix ideas, functional analyses used in this study have several advantages over classical vector analyses [9] to detect outliers. In practice, the functional analysis does not require a Gaussian initial dataset or to subject the non-Gaussian initial dataset to statistical Box–Cox power transformations. In this sense, only the functional method has been performed here.

The results obtained revealed outliers in SO_2, NO, NO_2, CO, PM_{10} and O_3 emissions for sixty 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.

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. 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.

Author Contributions: P.J.G.N. and E.G.G. conceived and designed the experiments; P.J.G.N. and C.O.G. performed the experiments; F.S.L. and E.G.G. analyzed the data; E.G.G., F.S.L., C.O.G. and P.J.G.N. wrote the paper.

Acknowledgments: The authors wish to acknowledge the computational support provided by the Departments of Mathematics at University of Oviedo as well as pollutant data from the Santa Marina air quality automated monitoring station supplied by the Section of Industry and Energy from the government of Asturias (Spain).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. García-Nieto, P.J. Parametric study of selective removal of atmospheric aerosol by coagulation, condensation and gravitational settling. *Int. J. Environ. Health Res.* **2001**, *11*, 151–162.
2. García-Nieto, P.J. Study of the evolution of aerosol emissions from coal-fired power plants due to coagulation, condensation, and gravitational settling and health impact. *J. Environ. Manag.* **2006**, *79*, 372–382.
3. García Nieto, P.J.; Sánchez Lasheras, F.; García-Gonzalo, E.; de Cos Juez, F.J. PM 10 concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: A case study. *Sci. Total Environ.* **2018**, *621*, 753–761, doi:10.1016/j.scitotenv.2017.11.291.
4. García Nieto, P.J.; Sánchez Lasheras, F.; García-Gonzalo, E.; de Cos Juez, F.J. Estimation of PM 10 concentration from air quality data in the vicinity of a major steelworks site in the metropolitan area of Avilés (Northern Spain) using machine learning. *Stoch. Environ. Res. Risk Assess.* **2018**, 1–12, doi:10.1007/s00477-018-1565-6.
5. Crespo Turrado, C.; Meizoso López, M.C.; Sánchez Lasheras, F.; Rodríguez Gómez, B.A.; Calvo Rollé, J.L. Missing data imputation of solar radiation data under different atmospheric conditions. *Sensors* **2014**, *14*, 20382–20399, doi:10.3390/s141120382.
6. Rousseuw, P.J.; Ruts, I.; Tukey, J.W. The bagplot: A bivariate boxplot. *Am. Stat.* **1999**, *53*, 382–387.
7. Tukey, J.W. Mathematics and the picturing of data. In Proceedings of the International Congress of Mathematicians, Vancouver, BC, Canada, 21–29 August 1974; James, R.D., Ed.; Canadian Mathematical Congress: Vancouver, BC, Canada, 1975; pp. 523–531.
8. Shang, H.L.; Hyndman, R.J. Boxplot and Outlier Detection for Functional Data. In Proceedings of the First International Workshop on Functional and Operational Statistics, Toulouse, France, 19–21 June 2008.
9. Martínez Torres, J.; García Nieto, P.J.; Alejano, L.; Reyes, A.N. Detection of outliers in gas emissions from urban areas using functional data analysis. *J Hazard. Mater.* **2011**, *186*, 144–149, doi:10.1016/j.jhazmat.2010.10.091.