Journal of Hazardous Materials 186 (2011) 144–149

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

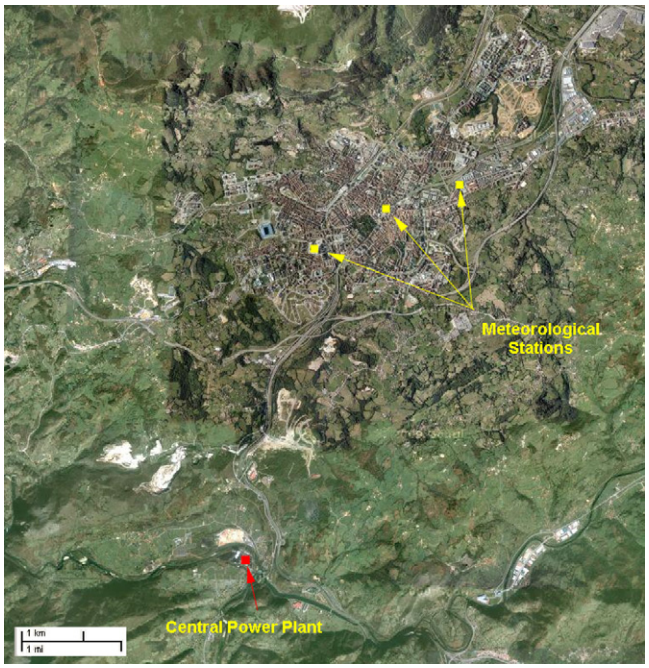

Fig. 1. Photograph of the study area showing the location of the metereological stations in the city of Oviedo and the coal-fired power plant.

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 $\mu g/m^3$) 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 $\mu g/m^3$. In our study, the annual arithmetic means for this gas in 2006, 2007 and 2008 were 24.0, 23.27 and 24.31 $\mu g/m^3$ 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:

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

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:

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