Science of the Total Environment 485–486 (2014) 292–299

Contents lists available at ScienceDirect

Science of the Total Environment

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

Finite mixture models to characterize and refine air quality monitoring networks

Álvaro Gómez-Losada [a,b,*], Antonio Lozano-García [a,**], Rafael Pino-Mejías [b], Juan Contreras-González [c]

[a] Environmental and Water Agency of Andalusia, c/Johan G. Gutemberg s/n, Isla de la Cartuja 41092 Seville, Spain
[b] Department of Statistics and Operational Research, University of Seville, Avda. Reina Mercedes s/n, Seville, Spain
[c] Environmental Council of the Junta de Andalucía, Avda. Manuel Siurot 50, 41071 Seville, Spain

HIGHLIGHTS

- Mixture models used to fit observational data are summarized by means of μ_m and σ_m.
- Source attributions are easily detected through the use of these models.
- Imputation permits the estimation of pollutants unmonitored due to limited resources.
- New configurations of the monitoring network arise analyzing the overall information.
- PCA clarifies the network setup and determines site misclassifications.

ARTICLE INFO

Article history:
Received 27 December 2013
Received in revised form 22 February 2014
Accepted 19 March 2014
Available online xxxx

Editor: P. Kassomenos

Keywords:
Air quality monitoring networks
Finite mixture models
Random forests
Imputation
Missing data
Seville

ABSTRACT

Background: Existing air quality monitoring programs are, on occasion, not updated according to local, varying conditions and as such the monitoring programs become non-informative over time, under-detecting new sources of pollutants or duplicating information. Furthermore, inadequate maintenance may cause the monitoring equipment to be utterly deficient in providing information. To deal with these issues, a combination of formal statistical methods is used to optimize resources for monitoring and to characterize the monitoring networks, introducing new criteria for their refinement.
Methods: Monitoring data were obtained on key pollutants such as carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), particulate matter (PM_{10}) and sulfur dioxide (SO_2) from 12 air quality monitoring sites in Seville (Spain) during 2012. A total of 49 data sets were fit to mixture models of Gaussian distribution using the expectation–maximization (EM) algorithm. To summarize these 49 models, the mean and coefficient of variation were calculated for each mixture and carried out a hierarchical clustering analysis (HCA) to study the grouping of the sites according to these statistics. To handle the lack of observational data from the sites with unmonitored pollutants, the missing statistical values were imputed by applying the random forests technique and then later, a principal component analysis (PCA) was carried out to better understand the relationship between the level of pollution and the classification of monitoring sites. All of the techniques were applied using free, open-source, statistical software.
Results and conclusion: One example of source attribution and contribution is analyzed using mixture models and the potential for mixture models is posed in characterizing pollution trends. The mixture statistics have proven to be a fingerprint for every model and this work presents a novel use of them and represents a promising approach to characterizing mixture models in the air quality management discipline. The imputation technique used is allowed for estimating the missing information from key unmonitored pollutants to gather information about unknown pollution levels and to suggest new possible monitoring configurations for this network. Posterior PCA confirmed the misclassification of one site detected with HCA. The authors consider the stepwise approach used in this work to be promising and able to be applied to other air monitoring network studies.

Abbreviations: CO, carbon monoxide; NO_2, nitrogen dioxide; O_3, ozone; PM_{10}, particulate matter with diameter of 10 μm or less; SO_2, sulfur dioxide; EM, expectation–maximization; HCA, hierarchical clustering analysis; PCA, principal component analysis.

* Correspondence to: A. Gómez-Losada, Environmental and Water Agency of Andalusia, c/Johan G. Gutemberg s/n, Isla de la Cartuja, 41092 Seville, Spain. Tel.: +34 662974023; fax: +34 955044508.
** Corresponding author. Tel.: +34 697956143; fax: +34 955044508.
E-mail addresses: agomezlo@agenciamedioambienteyagua.es (Á. Gómez-Losada), alozano@agenciamedioambienteyagua.es (A. Lozano-García).

http://dx.doi.org/10.1016/j.scitotenv.2014.03.091
0048-9697/

1. Introduction

The finite mixture of distributions has been used extensively to model heterogeneous data in many fields and numerous examples can be found in McLachlan and Peel (2000). More recently, mixture models have been applied in bioinformatics and genetics (Delmar et al., 2005). However, applications in the environmental literature have been limited (Li et al., 2013) and the potential of finite mixture models has not been exploited in the area of air pollution research. In this study its potential is applied to this latter field.

In air pollution research, when a good fit is pursued, it is not always sufficient to use just one classical distribution to describe a data set. This is especially true if this data set is derived from two or more subpopulations or *generating processes*. In such cases it is necessary to fit a composition of distributions to the underlying data set. These distributions are called *mixture models* and are defined by the parameters specific to each component and the proportion in which the mixed components occur. The *clustering* of the set of observations comes about by deducing the component's parameters and classifying each observation by component. These mixture models may be well defined and summarized by their mean and variance values, and consequently, by their coefficients of variation constituting the fingerprint of these models (see Appendix A for a basic review of the main idea of these values and the notation). Mixture distributions can be fit using many techniques such as graphical methods, the method of moments, maximum likelihood estimation (MLE), Bayesian approaches, and others. The EM algorithm is a popular tool for the iterative MLE of mixture distributions (McLachlan and Basford, 1988). The basic idea is to introduce a multinomial indicator variable that identifies cluster membership for every observation. This represents a convenient technique for obtaining the parameters from the mixtures since analytical solutions are not available (see Wilks (2006), for an account of the EM algorithm in a general context).

The design of the air quality monitoring networks basically involves determining the number of stations and their location, class, and number of pollutants monitored with a view to the objectives, costs, and available resources. In most cases, air quality monitoring networks in metropolitan areas are designed to measure pollutants of concern such as CO, NO_2, O_3, PM_{10} and SO_2 (Chang and Tseng, 1999). Future air quality monitoring standards may not be met if the importance of these pollutants in identifying the sources in existing monitoring stations is not reassessed. On occasion, some of these pollutants are simultaneously monitored at neighboring stations, leading to a duplication of information by detecting similar levels of pollution (equipment redundancy in air monitoring networks was dealt with by Pires et al. (2008)). Proper maintenance and carrying out operational tasks are requirements if monitoring equipment is to function at satisfactory levels. In addition to these considerations, an important objective of any air quality monitoring network should be the effective and optimized statistical use of the information acquired therewith.

The aim of this work is to implement an integrated approach for fully exploiting the information provided by the air monitoring networks and to obtain consistent criteria for their refinement. Refinement essentially consists of refocusing the parameters monitored on sites, detecting duplicates, re-classifying the types of stations, and finally, implementing consequent and progressive modifications to the routine monitoring networks by network managers. Even without taking the possible refinement step to the network into account, this methodology is useful in that it offers valuable information about the intrinsic structure of the monitoring network.

This approach was applied using statistical techniques, all of which were available through the open-source software *R* (R Core Team, 2013). Every technique was applied in a stepwise fashion with the following objectives: 1) to obtain from every mixture model the level and variability of exposure to pollutants on sites, evaluate source attributions, and calculate the essential statistics μ_m and cv_m, model mean and model coefficient of variation, respectively; 2) to identify site groupings according to μ_m and cv_m values using HCA; 3) to impute missing μ_m and cv_m values from unmonitored pollutants by means of the random forests technique (Breiman, 2001); and 4) to study the reclassification of sites through PCA considering the new (imputed) information obtained.

The described procedure was applied to the air quality monitoring network of Seville. The province of Seville is located in Andalusia, in southern Spain, covering an area of 14,036 km^2 and the year the study was carried out had a total population of 1,935,364 (IECA, 2012). The city of Seville is the largest urban agglomeration in Andalusia and has a population of 1,217,811 (SG, 2012).

2. Material and methods

2.1. Observation sites and source of data

The air quality monitoring network of Andalusia includes 89 operating sites and the pollutants CO, NO_2, O_3, PM_{10} and SO_2 are simultaneously monitored at approximately 43 sites. This network is managed by the Environmental and Water Agency run by the Regional Ministry of Environment and Land Planning of Andalusia (Consejería de Medio Ambiente y Ordenación del Territorio de la Junta de Andalucía).

The focus of this study is on the analysis of the aforementioned pollutants performed on 2012 data from 10 sites within the metropolitan area of the city of Seville and 2 rural sites, permitting a wide range of pollution levels, contributions and locations to be considered.

The monitoring sites where the data were collected from were classified according to the type of area (R—Rural, S—Suburban, U—Urban) and the predominant emission sources on site (B—Background, I—Industrial, T—Traffic) (general site characteristics and analyzed pollutants are given in Table 1).

Since monitored data are available at a 10-minute temporal resolution, the observed pollutant concentrations were averaged to establish a single value for each day, ensuring independent statistical observations. The averaged values at every site and for each day were only calculated when 80% of the data were available (data on 19 out of 24 h). In this study, all concentration units are referred to in $\mu g/m^3$.

The reference monitoring methods established in European Directive, 2008/50/EC (Directive, 2008) were used for pollutants CO, NO_2, O3 and SO_2, and beta attenuation monitoring was applied for PM_{10}. This method can be related to the reference method by a single correlation factor (Hauck et al., 2004). In this study, a correction factor to the PM_{10} data were not applied nor were the natural contribution of particles taken into account; the values obtained by the monitoring stations were used directly.

Air quality monitoring networks are subject to intense maintenance programs, to ensure accurate values. The data obtained are validated by the Regional Ministry of Environment and Land Planning of Andalusia prior to undergoing analysis.

2.2. Model selection and computation

In this study, the EM algorithm is performed to fit mixture distributions (Dempster et al., 1977; McLachlan and Peel, 2000; McLachlan and Ng, 2009) to every data set obtained from pollutant monitoring on the sites cited in Table 1.

The EM algorithm seeks to maximize the log-likelihood in a missing data framework. Mixture models can be dealt with by EM by defining a set of unobserved variables z_{ij}, where $z_{ij} = 1$ if observation i comes from component j of the mixture, and where $z_{ij} = 0$ otherwise. After initializing the vector of parameter estimations, EM alternates between two steps: E for expectation and M for maximization. For mixtures of normal components, the E step computes the expected value of the z_{ij} variables using the last vector of parameter estimations. The M step performs a weighted maximum likelihood estimation using the values of the