Atmospheric Environment 117 (2015) 271–281

Contents lists available at ScienceDirect

Atmospheric Environment

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

Time series clustering for estimating particulate matter contributions and its use in quantifying impacts from deserts

Álvaro Gómez-Losada [a,*], José Carlos M. Pires [b], Rafael Pino-Mejías [a]

[a] Departmento de Estadística e Investigación Operativa, Facultad de Matemáticas, Universidad de Sevilla, Avda. Reina Mercedes s/n, 41001 Sevilla, Spain
[b] LEPABE, Departamento de Engenharia Química, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

HIGHLIGHTS

- The impact of wind-blown desert to annual average PM_{10} concentrations was estimated.
- Hidden Markov models used to define and estimate regimes of PM_{10} concentrations.
- New methodology for calculating daily net PM_{10} loads from deserts.
- The modelling used complements other source apportionments techniques.

ARTICLE INFO

Article history:
Received 20 February 2015
Received in revised form
15 July 2015
Accepted 17 July 2015
Available online 23 July 2015

Keywords:
Apportionments
Hidden Markov Model
PM_{10}
Sahara

ABSTRACT

Source apportionment studies use prior exploratory methods that are not purpose-oriented and receptor modelling is based on chemical speciation, requiring costly, time-consuming analyses. Hidden Markov Models (HMMs) are proposed as a routine, exploratory tool to estimate PM_{10} source contributions. These models were used on annual time series (TS) data from 33 background sites in Spain and Portugal. HMMs enable the creation of groups of PM_{10} TS observations with similar concentration values, defining the pollutant's regimes of concentration. The results include estimations of source contributions from these regimes, the probability of change among them and their contribution to annual average PM_{10} concentrations. The annual average Saharan PM_{10} contribution in the Canary Islands was estimated and compared to other studies. A new procedure for quantifying the wind-blown desert contributions to daily average PM_{10} concentrations from monitoring sites is proposed. This new procedure seems to correct the net load estimation from deserts achieved with the most frequently used method.

1. Introduction

The main objective of many monitoring studies related to atmospheric aerosols is the identification and apportionment of pollutants to their sources. This information is crucial for the development and implementation of policies protecting human health and the environment as well as the design of effective mitigation strategies on a local or broader scale where the legislation thresholds are exceeded. Source apportionment (SA) is the practice of obtaining information about pollution sources and their contribution to ambient air pollution levels. There are three main groups of SA techniques (Viana et al., 2008): (i) methods that involve the assessment of monitoring data, (ii) methods that rely on emissions inventories and/or atmospheric dispersion modelling, and (iii) methods based on the statistical evaluation of the chemical data on particulate matter gathered from receptor sites (receptor models or RMs). The first group is considered to be based on basic numerical data treatment (Belis et al., 2013). It also includes simple time series (TS) modelling of data that may be used, for instance, to estimate natural PM_{10} contributions from deserts (Escudero et al., 2007a). The second one includes models to simulate aerosol emission formation, transport and deposition, although they are limited by the accuracy of emission inventories, when available. The third group is especially used for airborne particulate matter. The foundational principle of RMs is based on a mass balance between the emitter and the receptor, which assumes that the mass and species remain constant from one to the other or experience minimal change.

Abbreviations: HMM, Hidden Markov Model; PM_{10}, particulate matter with aerodynamic diameter of 10 μm or less; RB, regional background; RM, receptor model; SA, source apportionment; TS, time series.

* Corresponding author.
E-mail addresses: alvgomlos@alum.us.es, alvaro.gomez.losada@gmail.com (Á. Gómez-Losada).

http://dx.doi.org/10.1016/j.atmosenv.2015.07.027
1352-2310/

In addition to this classification, a basic statistical analysis is recommended before undergoing any SA study, which should include time-trend analyses or statistical distribution fitting that may describe the data sets under study (Belis et al., 2014). Simple statistical methods such as correlations or time-trend modelling are then used as an initial approach for suggesting SA or as a task prior to applying the time-consuming and more expensive RMs in which chemical speciation is required. Exploratory methods are varied and are not really SA oriented. Moreover, a strong statistical theory to back them is missing. More robust SA results can be obtained if the advantages of different types of modelling are combined, since no single model is completely adequate due to the theoretical assumptions. This represents the motivation behind this work.

Hidden Markov Models (HMMs) are scarcely used in predicting air quality due to their limited ability to accurately forecast pollutant concentrations (Dong et al., 2009). This limited ability is caused by the Markov property, by which only the present state provides any insight into the future behaviour of the process (information regarding the history of the process does not reveal anything new about the process). If no predictive statistics are desired with respect to pollutant concentration, HMMs show promise as flexible general purpose models for univariate (Cappé et al., 2005) and multivariate TS analyses (Zucchini and MacDonald, 2009), while at the same time allowing for relatively easy and straightforward interpretation (Visser et al., 2009; Visser, 2011).

HMMs constitute a starting point for SA based on the study and characterisation of PM_{10} TS, clustering their observations over time in homogeneous groups or *regimes* of concentrations. In this study, Gaussian HMMs are applied to univariate PM_{10} TS obtained from permanent background monitoring sites in the Iberian Peninsula and the Canarian, Balearic and Azorean Archipelagos. Interesting properties of HMMs are also applied to determine the probability of change between regimes or to obtain the average concentrations of the TS. The modelling was applied to the data relying on the authors' prior knowledge of SA as a prerequisite. To that end, the case of the Temisas site in Las Palmas de Gran Canaria Island (Canary Islands, Spain) is analysed. The SA on this archipelago has been previously studied by other authors (Rodríguez et al., 2001; Viana et al., 2002; Querol et al., 2004) and high contributions of particulate matter due to the transport of air masses from the Sahel and Sahara deserts (North Africa) has been confirmed.

This study aims: (i) to propose the use of homogenous HMMs as a routine exploratory tool to complement other SA techniques to estimate PM_{10} contributions from different sources; and (ii) to introduce a new method for deriving the dust net load from deserts using HMMs.

This study is outlined as follows. In Section 2 the data used in this study and the structure of the HMMs are explained. Section 3.1 deals with the application of HMMs to the Temisas site TS during 2013, defining their regimes, estimating different apportionments and how these regimes contribute to the annual mean PM_{10} concentration in this area. Sections 3.2 and 3.3 extrapolate this application to rest of the analysed sites, on a geographical and temporal scale, respectively. In Section 3.4 a new method for estimating contributions from deserts is proposed and finally concluding remarks are given in Section 4.

2. Material and methods

2.1. Monitoring sites and data

In this work, data sets of daily averages of PM_{10} concentrations collected at 33 background sites on the Iberian Peninsula and the Azorean, Balearic and Canarian archipelagos (Table SM.1 in Supplementary Material) have been studied at different years. Of these sites, 28 belong to the Spanish Ministry of Agriculture, Food and the Environment (MAFE) and are included in the Iberian background network for the detection of African episodes (Querol et al., 2013a), with 13 of them also being included in the EMEP (Co-operative Programme for Monitoring and Evaluation of the Long-Range Transmission of Air Pollutants in Europe) network (EMEP, 2014). The *Comissão de Coordenação da Direcção Regional* (CCDR) *do Centro*, CCDR *do Alentejo* and *Direcção Regional do Ambiente dos Açores* from Portugal manage 5 of these monitoring sites. The used data were provided by these Portuguese institutions and MAFE after validation.

The PM_{10} concentrations from the monitoring sites were determined using the gravimetric and automatic (beta-radiation attenuation and TEOM) methods. Therefore, in order to harmonise the TS data, the measurements were corrected by applying the correction factors obtained by a comparison with the gravimetric method (EN-12341, 1998). Occurrences of daily episodes of intrusions of particulate matter during 2013 due to North African transport of air masses applied in this work were established by Pérez et al. (2014) using a combination of methods (Querol et al., 2009), including HYSPLIT modelling (Draxler and Rolph, 2003).

2.2. Model definition

HMM is a time-dependent process generated by two interrelated probabilistic mechanisms, in which one is an underlying and hidden process, and a series of hidden states, while the other is the TS observation sequence determined by the current hidden state of a given Markov chain (Rabiner, 1989). HMM represents a flexible method of modelling TS that exhibits dependence over time as well as average PM_{10} concentrations collected in air quality monitoring networks. In most HMM applications, the hidden state outputs are represented by Gaussian distributions. Modelling daily average PM_{10} concentrations sampled during a year represents a problem because of the impossibility of capturing the asymmetrical distribution of this pollutant in a single distribution (e.g. log-normal distribution). One way to address this problem is to use multiple (a mixture) Gaussians to approximate the real distribution.

The model consists of two parts: firstly, the daily average PM_{10} concentrations (observations) which describe a TS of length T, and secondly, unobserved states, satisfying the Markov property, which are responsible for generating the observations. The states are hidden to the observer who just perceives the TS observations. The Markov property ensures that the highly temporal-dependent nature of PM_{10} concentrations on consecutive days is taken into account, a property which may be assumed when one day's concentration shows dependency on that of the previous day. States are distinct elements of the HMM, N being the number of states of the model. This number is also used to name the HMM (e.g. an N-state HMM).

In Fig. 1, how one hidden state transitions to another state generating the observations of an annual TS ($T = 365$) is first depicted and then the elements of an HMM are defined. For the sake of simplicity, this example uses a two-state HMM and the first five observations (from the first day -$t = 1$- to the fifth -$t = 5$-) of the TS are explained. Hidden states are denoted by circles and possible transitions among hidden states by arrows, with their probabilities given. The path generating the observation is indicated by highlighted arrows and blue circles. In the beginning ($t = 1$), the Markov chain is initialised according to the initial state probability distribution $\boldsymbol{\delta} = (1,0)$ and starts at state 1. Then the hidden state transfers from the initial state to the next state according to a transition probability matrix ($\mathbf{A}$), which describes the probabilities for all the