This article was published in Atmospheric Environment, 117, 271-281, 2015

http://dx.doi.org/10.1016/j.atmosenv.2015.07.027

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

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.

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 behavior of the process (in-formation 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 PM10 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 (Cooperative 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 initialized according to the initial state probability distribution δ = (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 (A), which describes the probabilities for all the transitions. As elements of this matrix are probabilities, they are non-negatives, no greater than 1 and each of the rows sum to unity. Each of the hidden states addresses an associated statistical distribution from which the data are generated. These distributions are referred to in the literature as emission probabilities denoted by B. In this work, this distribution is represented by a weighted sum of Gaussian densities. In Fig. 1, they are represented by two Gaussian densities, N (15,3) and N (30,7), contributing equally ($\pi 1 = \pi 2 = 0.5$) to capturing the shape (histogram) of the TS. The example finishes at time t = 5 after having generated five observations of the TS. Lower concentration values in the TS are generated by the Gaussian distribution associated with state 1, while higher ones are generated by the Gaussian distribution associated with state 2. This observation leads to the formation of two groups of the TS observations depending on their concentration values. This idea helps to clarify the role of hidden states as elements that cluster the TS observations, modelling the temporal heterogeneity of any given TS.

Next, the elements defining an HMM may be given:

1. The number of states of the model, N. The individual states are denoted as S = {S1,…, S_N} and the state at time t is denoted by q_t.

2. The initial state probability distribution, determining in which state the Markov chain starts to transition at t = 1, defined as $\delta t = P(S_t = i)$, i = 1,…, N.

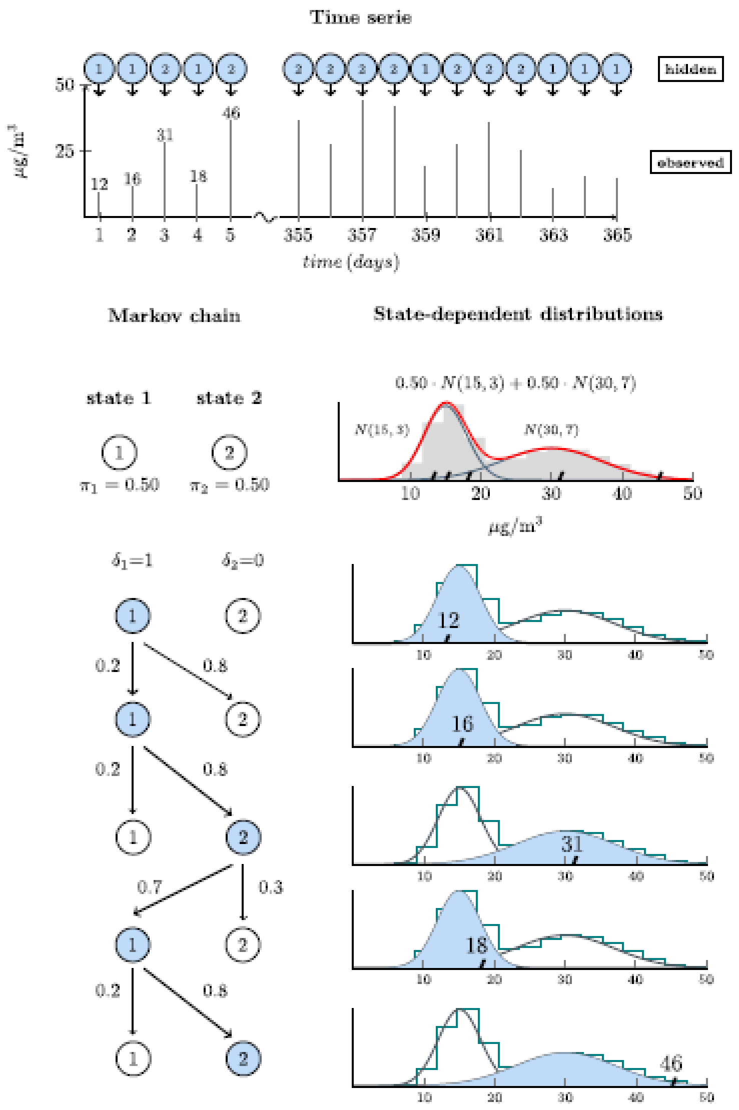
3. The state transition probability matrix A={ a_{ij}}, with elements:

$$a_{ij} = P(S_t = j|S_{t-1} = i) \quad i,j = 1,\ldots,N$$

indicating the probability the state at time t (j) given the state at time t - 1(i).

4. M, the number of distinct observations of the TS for each state. The individual observations are denoted by V = {v_1, v_2,…, v_M}.

5. The emission probability distribution in state Si, B = {bi(k)}, where bi(k) = P(v_k/q_t = S_i), is the probability that a particular observation of the TS is emitted in a state Si at time t, i = 1,…, N, k = 1,…, M. This element of the HMM includes the parameters of the weighted sum of N Gaussian distributions: the weighting coefficient (π_i), the mean (μ_i) and the standard deviation (σ_i) values, i = 1,…,N, of the Gaussian distributions. The weighting coefficients satisfy the constraint that their values sum to unity.



$$\mathbf{A} = \begin{array}{c} \mathcal{S} \\ \text{from 1} \\ \text{from 2} \end{array} \begin{array}{c} \begin{array}{cc} \text{to 1} & \text{to 2} \end{array} \\ \begin{pmatrix} 0.2 & 0.8 \\ 0.7 & 0.3 \end{pmatrix} \end{array}$$

$$\mu_m = \sum_{i=i}^{2} \pi_i \, \mu_i = (0.5 \times 15) + (0.5 \times 30) = 22.5 \ \mu g/\mathrm{m}^3.$$

Fig. 1. Example of a TS modelling with a two-state HMM with A representing the transition probability matrix of the unobserved Markov chain. Calculation of the mean value of the TS (µm).

2.3. Model estimation

The estimation problem lies in finding the parameters of the HMM that specify the model that is most likely to have generated a given TS. This is referred to as the maximum likelihood estimation (MLE) problem. Although there are many methods which can be used to estimate the parameters of an HMM, the Expectation-Maximisation (EM) algorithm (Dempster et al., 1977; Wu et al., 2008) is the most widely used. The EM algorithm can be applied for MLE when there are hidden data (e.g., hidden states) or when a problem can be reformulated in those terms. Briefly, the EM algorithm takes the observed data and an initial estimate of the parameters and uses them to estimate the hidden data (the expectation step); it then takes the observed data and the estimated hidden data and uses them to provide a new estimate of the parameters (the maximisation step) in an iterative fashion. The algorithm iterates until a convenient stopping criterion is met (Wilks, 2006). This criterion may be a permitted number of iterations of the algorithm, an acceptable minimum difference term (E) considering the parameter estimations at each iteration, or both. The standard error of the parameters of the HMMs was obtained by means of a bootstrap approximation as described in Basford et al. (1997) adapted to univariate TS data.

The computational implementation of the HMMs was accomplished using the "depmixS4" package (Visser and Speekenbrink, 2010) from the open-source software R (R Core Team, 2013). The appropriate model selection was performed by estimating models for different values of hidden states (S = 1,…,7) and later the optimum solution was selected based on the lowest BIC (Bayesian information criterion) value (Schwarz, 1978). The EM algorithm used to obtain the parameters of every HMM was setup with E = 10^{-8} and a maximum of 2000 iterations. The computational implementation and commented R code is available in the Supplementary Material of this work. To check the validity of the modelling results obtained with the "depmixS4" library, the "HiddenMarkov" (Harte, 2015) and "HMM" (Himmelmann, 2010) libraries were also used, and negligible differences were found in the parameter estimates.

2.4. Application of HMMs to the PM10 TS study

For many practical applications there is often some physical significance attached to the hidden states of the model. For instance, in economics, states of the HMM can be related to "state of the economy" (e.g. expansion and recession) and the interest is in studying the dynamics between them (Dias et al., 2010); in development psychology, the states of the HMM are used to quantify knowledge that subjects express in an implicit learning task (Visser et al., 2002); or in the study of sleep stages, the states of an HMM correspond to various stages such as REM (rapid eye movement) sleep, deep sleep and wakefulness (Flexer et al., 2002).

In this work, hidden states of the HMMs are assigned to represent different PM_{10} concentration ranges in each modelled annual TS (77 data sets; see Tables SM.1 to SM.6 in the Supplementary Material). For the sake of simplicity and a more intuitive approach, the hidden state concept and the term "concentration ranges" will be unified and the term regimes of concentration or simply regimes will be used. These regimes, as a result of applying a cluster technique, groups observations of the TS with a relatively similar PM10 concentration, which at the same time are dissimilar to other ones grouped in other regimes. However, values of these regimes may show some overlapping, which is required to precisely capture the data distribution in the TS.

HMM provides the mean and standard deviation values for every regime. These values define the different Gaussian distributions assigned to the regimes, providing very useful information as they fully characterise every analysed TS for a given time period and monitoring site. HMM enables this parameterisation to be summarised by using simple algebraic expressions, as indicated in Appendix A. These expressions calculate the mean and standard deviation values of every TS using these same values for every regime in a sum that is weighted by the representativeness of every regime (π) to the overall distribution of the TS data. These values defining the TS are denoted as μ_m (Fig. 1) and σ_m (where m stands for mixture). They are quantitatively similar to the arithmetic mean ($\bar{x}$) and standard deviation (s) values which could be obtained without considering a temporal dependence in the TS data. Therefore, μm serves as a quantitative indicator of the annual mean PM_{10} exposure level of populations and σm represents the level of variation of the pollutant distribution around the μm.

In addition, HMM results include the transition probability matrix, which indicates the dynamic transition among regimes as probability values. Thus, it is possible to determine how probable it is that after an observed PM_{10} concentration is assigned to a regime the next observation will be similar (stay at the same regime) or not (switch to other regime). This leads to the concept of stability of the TS, which is derived from the main diagonal elements of A. This diagonal is related to the probability of the process being in a given regime in the long run. Thus, the closer the elements of the diagonal are to 1, the greater the probability of expecting similar PM_{10} concentration values on a lasting basis for a given regime. For instance, in Fig. 1, these elements (A) show a remarkable instability of the hypothetical TS, as values of the main diagonal are closer to 0 than to 1 ($a_{11} = 0.2 \ll 1$; $a_{22} = 0.3 \ll 1$).

3. Results and discussion

3.1. HMM modelling and estimations

The numerical values of the HMM application to the daily average PM_{10} concentrations from the Temisas site during 2013 are shown in Table 1 (calculations of μm and σm are shown in Appendix A). A graphical depiction of this application is shown in Fig. 2 where the grouping of the daily average PM_{10} concentrations can be observed. This site and the others indicated in Table SM.1 from the Canarian archipelago have been used by other authors (Pérez et al., 2014; Pey et al., 2013; Querol et al., 2013a) to quantify the contribution of North African dust outbreaks in this area.

Fig. 2A and B are both equivalent and show the formation of four groups within the data as a TS and histogram, respectively. These groups correspond to the regimes (hidden states) of the HMM referred to in Section 2.2 (a four-state HMM). In Fig. 2A each observation is labelled with the number of the regime to which it has been assigned by the HMM, or is grouped below a coloured Gaussian line in Fig. 2B. The latter shows how the resulting density of the clustering (black line) captures the form of the data distribution (grey line) much better than any single density, whatever its family. In Table 1, it can be observed that the range of values for every regime shows the typical overlapping, but the similarity between the μm and σm values and the arithmetic mean ($\bar{x}$) and standard deviation (s) of the analysed data set can be appreciated.

After modelling the data, a meaning for each regime is assigned. The following definitions are proposed for the mean concentration value of every regime, applied both to the regional background case and period under study:

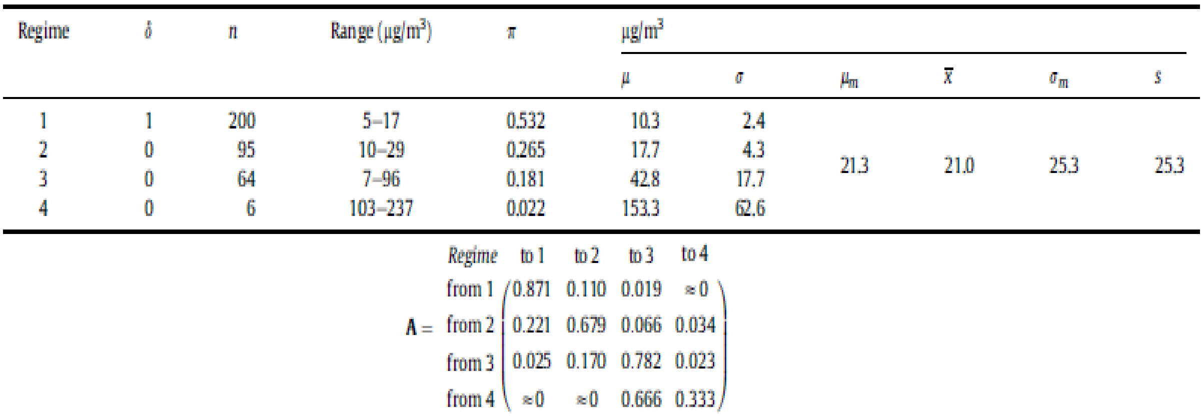
- μ1 (10.3 μg/m^3) represents the underlying or threshold concentration over which great changes in value are not expected over the years if atmospheric and pollution conditions remain relatively constant, being a characteristic of the studied area. Daily average PM_{10} concentrations assigned to this regime are supposedly not caused by any direct influence of natural or anthropogenic sources, or if they are, they are negligible.
- μ2 (17.7 μg/m^3) is the average PM_{10} concentration on the days affected by moderate contributions of anthropogenic sources due to activities that take place in the region. The value of μ2 is subject to slightly more variation than μ1 between years. The

referenced days may be affected by contributions from natural sources attributable to African dust transport episodes that have a minor impact on the observed PM_{10} concentrations.

- μ3 (42.8 μg/m^3) is the average PM_{10} concentration on days affected by characteristic, usual contributions from African outbreaks. These contributions are highly variable in concentration and are the main factor responsible for the exceedances of the 50 μg/m^3 limit value established by the Directive 2008/50/EC on ambient air quality and cleaner air for Europe (Directive, 2008) for this pollutant.
- μ4 (153.3 μg/m^3) is the average PM_{10} concentration on days with unusual but severe episodes of natural contributions from North African episodes.

Table 1

HMM results from the Temisas site and comparison of μm and σm with the arithmetic mean (x) and standard deviation (s) of the analysed data set. n indicates the number of observations grouped in each regime and the probability transition matrix is represented by A.



Regime	δ	n	Range (μg/m^3)	π	μg/m^3					
					μ	σ	μ_m	$\bar{x}$	σ_m	s
1	1	200	5–17	0.532	10.3	2.4	21.3	21.0	25.3	25.3
2	0	95	10–29	0.265	17.7	4.3				
3	0	64	7–96	0.181	42.8	17.7				
4	0	6	103–237	0.022	153.3	62.6				

$$\mathbf{A} = \begin{array}{c|cccc} \textit{Regime} & \text{to } 1 & \text{to } 2 & \text{to } 3 & \text{to } 4 \\ \hline \text{from } 1 & 0.871 & 0.110 & 0.019 & \approx 0 \\ \text{from } 2 & 0.221 & 0.679 & 0.066 & 0.034 \\ \text{from } 3 & 0.025 & 0.170 & 0.782 & 0.023 \\ \text{from } 4 & \approx 0 & \approx 0 & 0.666 & 0.333 \end{array}$$

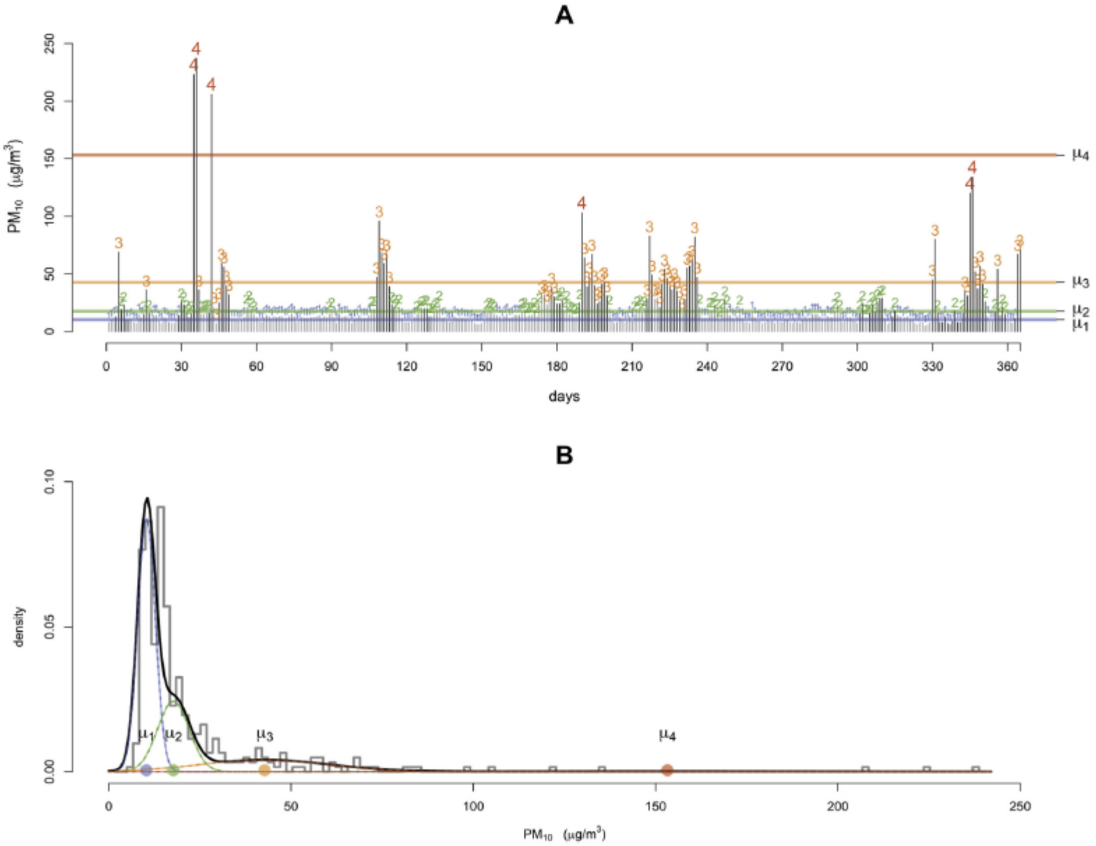


Fig. 2. A. Clustered PM10 TS at the Temisas site. Each observation is numbered after being assigned to a regime by the HMM (regime 1, 2, 3 and 4 in blue, green, orange and red, respectively). Mean values of every regimes (clusters) are indicated by μ1, μ2, μ3 and μ4 and are coloured accordingly. **B.** Mixture of Gaussian distributions capturing the shape of the data distribution. A and B show correspondence in colour. Black vertical lines in the TS indicate days in which African dust intrusions are detected. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Remarkably, the aforementioned definitions for regimes after applying HMM coincide from a conceptual point of view with the PM10 level separation made by Lenschow et al. (2001) at the Berlin (Germany) region, where local, urban and regional background fractions were made. The authors of this study gave these regimes the intuitive name of horizontal profiles of the ambient PM10 con-centration and particularly, what is here defined as the first regime was called natural background concentration. The later work by Escudero et al. (2007b) mentions the usefulness of interpreting the variability of the RB PM_{10} levels, since local contributions may be identified and thus plans and programmes for air quality improvement can be properly implemented.

It must be noted that the definitions given above have to be adapted to the geographical area under study. In general, these concentrations have an accumulative affect, as several pollution sources may contribute simultaneously to air quality (e.g. one observation belonging to regime 3 also includes contributions from the sources associated with regimes 1 and 2). Some useful quantities can be roughly estimated using these definitions and this assumption, namely:

- μ2 - μ1 (7.4 μg/m^3): average concentration due to anthropogenic contributions from the region.

- μ3 - μ2 (25.1 μg/m^3): average concentration associated with characteristic contributions from dry regions from North Africa when they occur.
- μ4 - μ2 (135.6 μg/m^3): average concentration from severe contributions from dry regions from North Africa when they occur.

More precise information from the modelling can be obtained as the contribution of every regime to the annual mean concentration of PM$_{10}$ at this site (21.3 μg/m^3). This can be easily calculated from the figures given in Table 1 by multiplying the representativeness of every regime (π) by its mean value (μ). Fig. 3 shows these contributions.

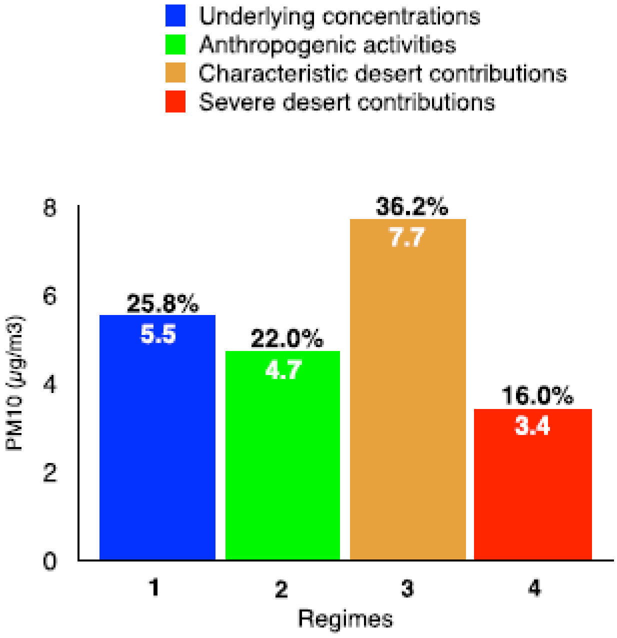


Fig. 3. Contributions of the different regimes to the annual PM10 average concentration (21.3 μg/m3) at the Temisas RB site. Each regime contribution to the annual PM10 average concentration is indicated within the bars (in μg/m3), and their representativeness in %. Contributions are calculated as πi · μi, i = 1,…, 4, from Table 1.

The transition probability matrix (A matrix -- Table 1) shows that regimes in this site are somewhat unstable, as the elements from the main diagonal of A are not close to 1 (each row sums to unity). Less probable transitions among regimes are close to 0, namely: a41, from regime 4 to regime 1, and a42, both due to atmospheric residence times after a severe African outbreak occurs; also, a14, for the impossibility of a sudden increase in PM$_{10}$ concentration from one to day to the next. In addition, the A matrix shows that the likelihood of two consecutive severe episodes is low (a44 = 0.333).

3.2. Behaviour of regimes in the Iberian Peninsula and archipelagos

The HMM described in Section 3.1 was applied to the data collected at 33 monitoring sites for 2013 indicated in Table SM.1. Fig. 4A displays these results, with each coloured dot representing the mean value of every PM10 regime at sites (the number of regimes detected and results of central tendency and dispersion measures are given in Supplementary Material 5), with a maximum of four being detected (μ1, μ2, μ3 and μ4), depending on the site. For the sake of simplicity, Fig. 4B sums up this information using boxplot diagrams, as the aim is not to describe every regime from monitoring sites but to gain a general overview of the behaviour of regimes by geographical area. Definition of the regimes in TS is a previous and necessary step for estimating the contribution of sources in a specific area using this modelling. Defining the regimes is always subjective and subject to several interpretations, and some consensus is needed among experts. Due to the number of studied sites, this task could not be carried out in this work. However, there are certain cases that are worth analysing. The remote Faial site in the Azores Archipelago (AZ) was significant. The fourth regime is missing and the concentrations of the existing ones are typically low (μm = 5.80 μg/m^3; μ1 = 2.30 μg/m^3, μ2 = 5.88 μg/m^3, μ3 = 11.13 μg/m^3). Whether this site is influenced by particulate matter from the desert or not cannot be determined, although the marine aerosol could be present, being represented partially by the third regime, considering the isolated location of this site in the Atlantic Ocean. Of three North sites, two lack the third regime. Sites located in the centre of the Iberian Peninsula (CE) present a fourth regime that could be represented both by dust resuspensions and by natural contributions from North Africa. The rest of the regime analyses are considered for further investigation.

The initial regime definitions given for the Temisas site may be considered again in this section, bearing in mind the background nature of all the studied site in this work. This is the case of regime 1, as it is assumable in theory that every site holds a minimum, underlying PM10 concentration that is characteristic of the site and that shows little variation over time if atmospheric conditions remain relatively constant. Consequently, the first regime (Fig. 4A, lower plot) may provide an indication of PM10 pollution at sites. Other regimes defined for the Temisas site are also applicable to the rest of monitoring sites from the Canarian Archipelago (CA in Fig. 4) due to the fact that all of them are affected by the same natural source of contributions. Therefore, Table 2 compares the estimated contribution from African episodes in the Canarian Archipelago given by Pérez et al. (2014) and the one using HMMs after considering the regime definitions given in Section 3.1 (regimes 3 and 4). As can be appreciated, the estimation results are similar. Those authors use a methodology based on the assessment of PM10 concentration TS at background sites by means of the calculation of the monthly mobile 40th percentile (Escudero et al., 2007a; Querol et al., 2013a, 2013b). This latter method is outlined in Section 3.4.

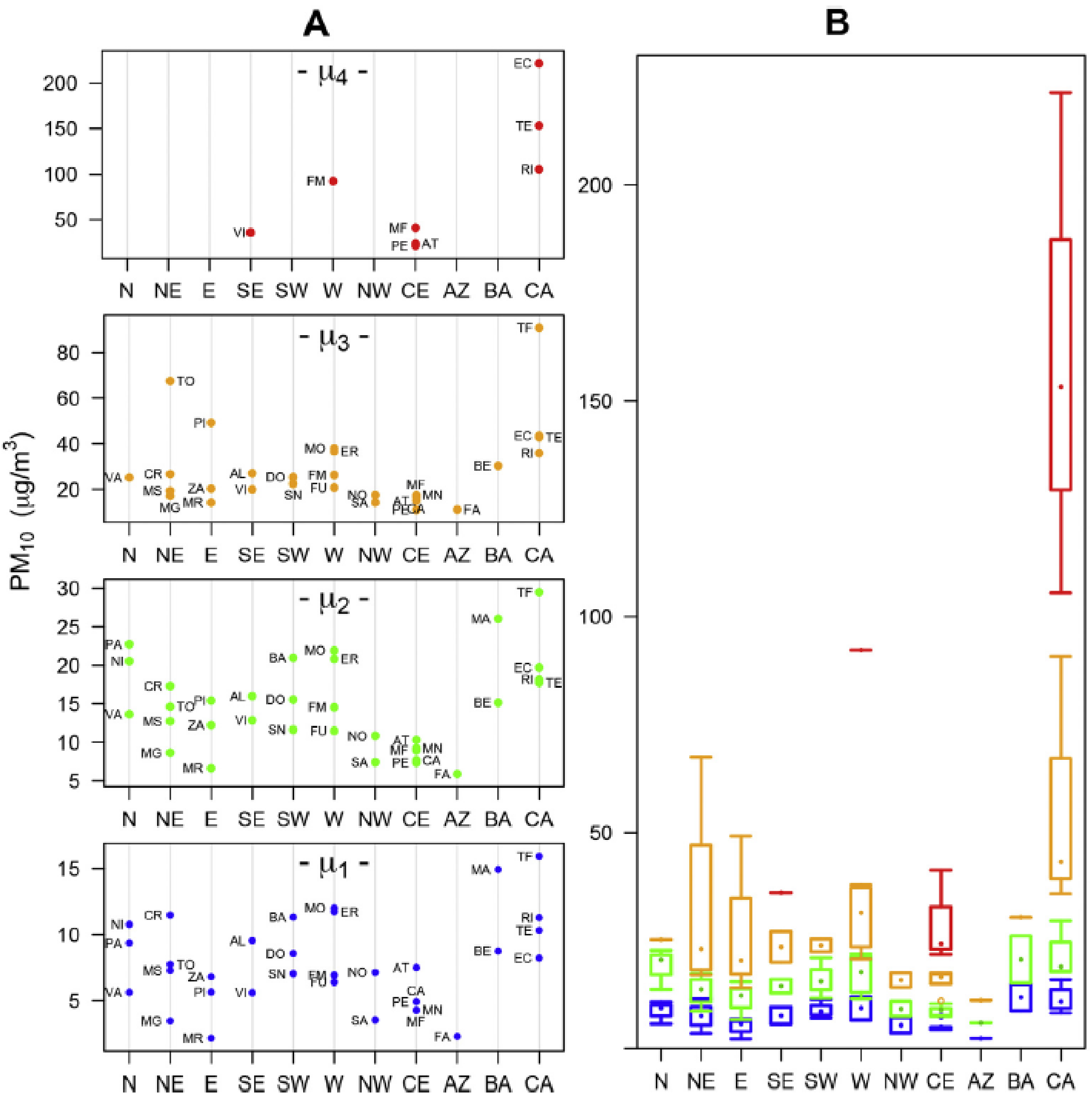


Fig. 4. A. Study of the regimes in the Iberian Peninsula and Canarian, Balearic and Azorean archipelagos during 2013. Each dot represents the mean values of PM10 regimes at sites, by geographical areas. **B.** Boxplot diagrams summing up the information from A. The colour codes for regimes are the same as in Fig. 2. Geographical areas are abbreviated as indicated in Table SM.1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2

Comparison between estimated contributions (in µg/m3) after African outbreaks in the Canarian Archipelago for 2013 and average values (x). The Temisas site included.

Site	HMMs	Pérez et al. (2014)
El Río	11.6	10.8
Temisas	11.1	9.0
Echedo	6.1	6.2
Tefia	9.9	10.4
	$\bar{x}$=9.7	$\bar{x}$=9.1

3.3. Study of the regimes over time

The same study of the behaviour of regimes is broadened from 2009 to 2013 for selected geographical areas (N, SE, CE and AZ) and the results are displayed in Fig. 5 (see Supplementary Material 6 for additional information). Fig. 5A shows the general trend of PM10 for that period using boxplot diagrams. Every boxplot describes all of the mm values at sites located in each area. The graph corresponding to the year 2013 in Fig. 5B repeats the results of Fig. 4B for these selected sites, and they are included again for comparative purposes with previous years. Sites from the Canarian Archipelago (CA) have not been included, as the air quality network of these islands was altered during the studied five years.

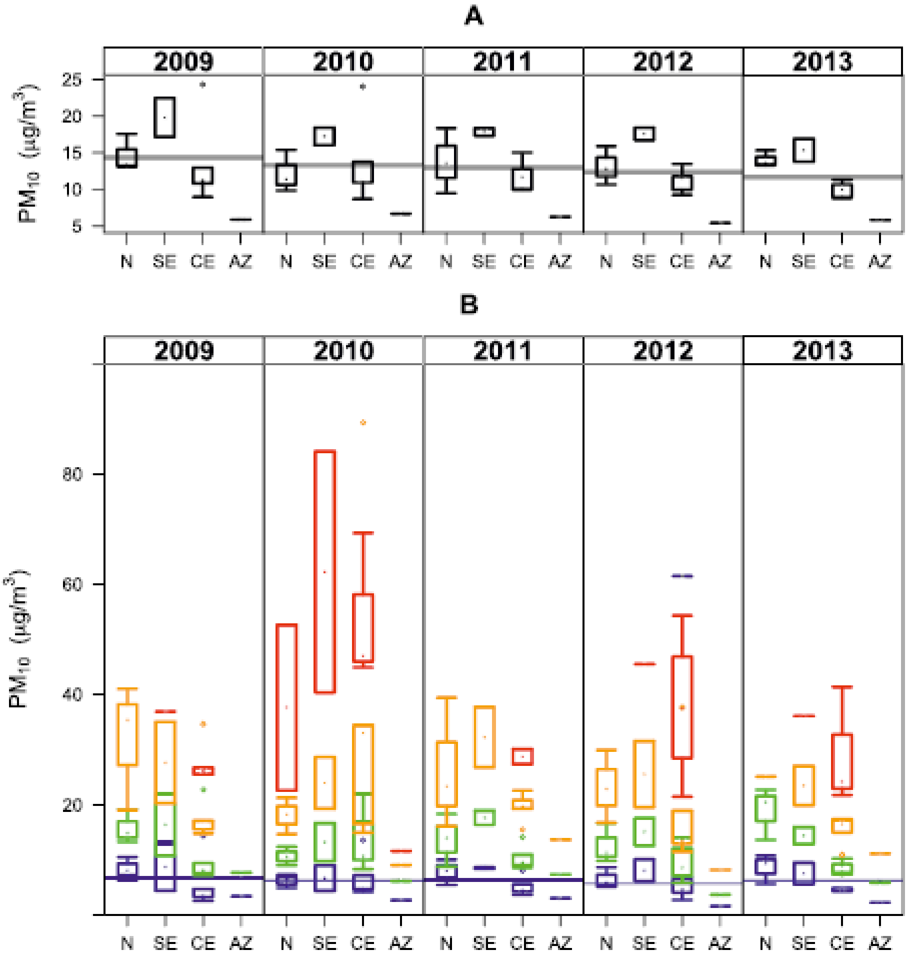


Fig. 5. A. Boxplot representation of mm values from 2009 to 2013 of sites in selected areas of the Iberian Peninsula and the Azores Archipelago. **B.** Boxplot study of the regimes of A. Abbreviation and colour codes are the same as in Fig. 4. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

As can be appreciated in Fig. 5A, a general, slightly downward trend in PM_{10} concentration is shown using a horizontal grey line, calculated using the mean value of all of the μm values for each year. This trend is also reflected in the disaggregation of the μm values by regimes shown in Fig. 5B, which was expected as every mm value is calculated considering the mean values of each regime. A close parallelism is detected between the first regime trend (boxplots and the horizontal line in blue) and the general trend (Fig. 5A), supporting the plausible use of this regime as a pollution indicator. This is due to the high representativeness of this regime in the overall distribution of the TS data (e.g. in the Temisas site, $\pi 1 = 0.532$).

The disaggregation shown in Fig. 5B may help to understand new/altered contribution phenomena, as the appearance of a fourth regime in 2010 for the N area (red boxplot) that is not pre-sent in the rest of years or a fifth regime in the CE area in 2012 (detected in S. Pablo). These events are by no means quantitatively important but could describe different apportions or changes from the expected ones. In the North case (N) for 2010, the fourth regime is described only at the Niembro and Pagoeta sites. These sites are located approximately 1 km and 6 km from the Cantabrian Coast respectively. In addition, a modification to the PM_{10} concentration levels is possible due to the marine aerosol apportion, although African contributions have also been described in studies carried out at the Niembro and Pagoeta sites (Pey et al., 2011) using the monthly mobile 40th percentile method. These studies determine that the African apportion in the Niembro and Pagoeta sites are 0.9 $\mu g/m^3$ and 0.7 $\mu g/m^3$, respectively. The HMM establishes that the contribution of the fourth regime ($\mu 4$) to the annual average PM_{10} concentration at these sites was 0.47 $\mu g/m^3$ and 2.33 $\mu g/m^3$ respectively.

In the case of the SE area for 2010, the fourth regime is present at the Viznar and Alcornocales sites. The fourth regime contributes to the annual average PM_{10} concentration for these sites by 4.31 $\mu g/m^3$ and 0.84 $\mu g/m^3$, respectively. Pey et al. (2011) calculate that the contribution from North African dust is 3.9 $\mu g/m^3$ and 2.0 $\mu g/m^3$ at these sites.

No direct conclusions from these HMMs results can be obtained, as regimes in these geographical areas have not been defined. However, similar results achieved by the currently used method-ology (Escudero et al., 2007a, 2007b) and HMMs suggest that further research is needed. It would be interesting to establish how both methodologies (and others) complement each other as HMMs may offer significant information once regimes are defined, namely: (i) net contributions due to different sources; (ii) contributions to the annual PM10 mean from every regime; and (iii) the probability of change among regimes. Because of this valuable in-formation provided, HMM represents a single statistical analysis with a strong theoretical background with which to characterise any TS. In addition, the ease of interpretation and reproducibility of the results should be considered, as well as the fact that this modelling can be performed making use of free software available to the research community.

3.4. New method proposed to quantify the apportionment of PM_{10} from deserts

Directive 2008/50/EC (Directive, 2008) allows for subtracting exceedances on the daily limit values of PM_{10} concentrations (50 $\mu g/m^3$) when they are attributable to natural events such as the transport of natural particles from dry regions (articles 2.15, 20.1 and 20.2). African dust outbreaks are responsible for a relevant percentage of the exceedances of the PM10 daily limit vale registered at rural and urban sites in the Mediterranean Basin (Salvador et al., 2014). The method that at present is more widely accepted (Viana et al., 2014) to estimate the daily African PM_{10} load was introduced by Escudero et al. (2007a) and it is based on the mobile 40th percentile (P40) method. This method is considered a reference method and is included in the text "Commission staff working paper establishing guidelines for demonstration and subtraction of exceedances attributable to natural sources under the Directive 2008/50/EC on ambient air quality and cleaner air for Europe" (EC, 2011).

Briefly, this method obtains the African PM10 load on a given day with the influence of African dust outbreaks, subtracting the RB level from the measured PM_{10} concentration. This RB level is obtained after applying the monthly mobile 40th percentile to the PM10 concentration TS at an RB site, after prior extraction from the TS of those days with African influence. The motivation to propose a new method is based on the suggestion of Viana et al. (2008) for further research on the current methodologies dealing with this specific contribution.

The aforementioned methodology was applied initially using the monthly mobile 30th percentile. This percentile was later shifted to the 40th percentile in order to correct the possible overestimation in the calculation of the North African contribution (Querol et al., 2009, 2013a). Although in the authors' opinion, this new percentile represents a suitable approximation for estimating the RB level on a given day, the method proposed in this work could improve the P40 method in two ways since it avoids: (i) the smoothed effect which is implicit in the P40 method after applying a mobile procedure in the TS treatments, and (ii) the empirical approach based on a correlation analysis applied in order to select this particular percentile (40th).

The example of Temisas (Section 3.1) is again taken into account and in particular, Fig. 1A shows a graphical and intuitive approach. The proposed method calculates the net contribution of African episodes as a result of subtracting the $\mu2$ value from the daily PM_{10} concentration when an African outbreak is detected. An example is given in Table 3 and compared to the P40 method. To gain a broader view of this comparison, Fig. 6 displays the different estimations of the dust loads in PM_{10} obtained through both methods for the Canarian archipelago sites (Temisas, Echedo and Tefia), and South (Viznar and Doñana), East (Zarra) and Centre (San Pablo, Peñausende and Campisábalos) sites on the Iberian Peninsula. As can be appreciated from this figure, the P40 method could overestimate the daily net load attributable to severe African episodes that frequently occur in the Canarian archipelago (Fig. 6AeC). This overestimation is less significant from a quantitative point of view when less severe contributions are observed (Fig. 6D, F-I; in Fig. 6E a slight underestimation is observed). Table 4 shows the difference between both methods when estimating average contributions on days affected by desert outbreaks for the analysed sites. Due to the empirical approach on which the P40 method is based, an analytical reasoning of these discrepancies cannot be directly derived. However, the smoothing effect referred to above is likely to be involved.

This proposed method is intuitive and simple, however a drawback is present. This is represented by the $\mu2$ concentration on which the method is based, and in particular, on its definition. As stated in Section 3.2, defining the regimes is always subjective and some consensus is needed among experts. This can discourage the end user from applying this method if knowledge on the main pollution sources of an area is missing. This difficulty is not pre-sent just in this application but in general when applying HMMs, as the regimes (hidden states) have to be given meaning. The given definition in this work for the Temisas site assumes that the main sources contributing to this regime have a regional anthropogenic origin and that the impact from natural sources to the range of concentrations on the regime is present in a lesser proportion. This regime definition is based on the study of the work by Viana et al. (2002, 2014) and Rodríguez et al. (2001).With respect to the rest of regimes, matrix A mentioned in Section 3.1 helps to clarify their behaviour and hence to give coherence to these definitions. The possible use of the quantity $\mu2$-$\mu1$ instead of $\mu2$ is not recommended for obtaining the net load as it is derived from the estimation of two mean values ($\mu2$ and $\mu1$) and therefore the difference is a rough estimation of the anthropogenic contributions.

Table 3

Example of five observations from the Temisas data set when using the proposed method to obtain the natural apportions from deserts. The last column shows the resulting quantity after applying the P40 method (in µg/m^3).

Date	Observed PM$_{10}$	Regime	μ_2	Desert contribution	
				Proposed method	P40 method
05 January 2013	69	3	17.7 ≈ 18	69–18 = 51	60
04 February 2013	223	4		223–18 = 205	213
05 February 2013	237	4		237–18 = 219	227
25 April 2013	21	2		21–18 = 3	10
12 December 2013	134	4		134–18 = 116	125

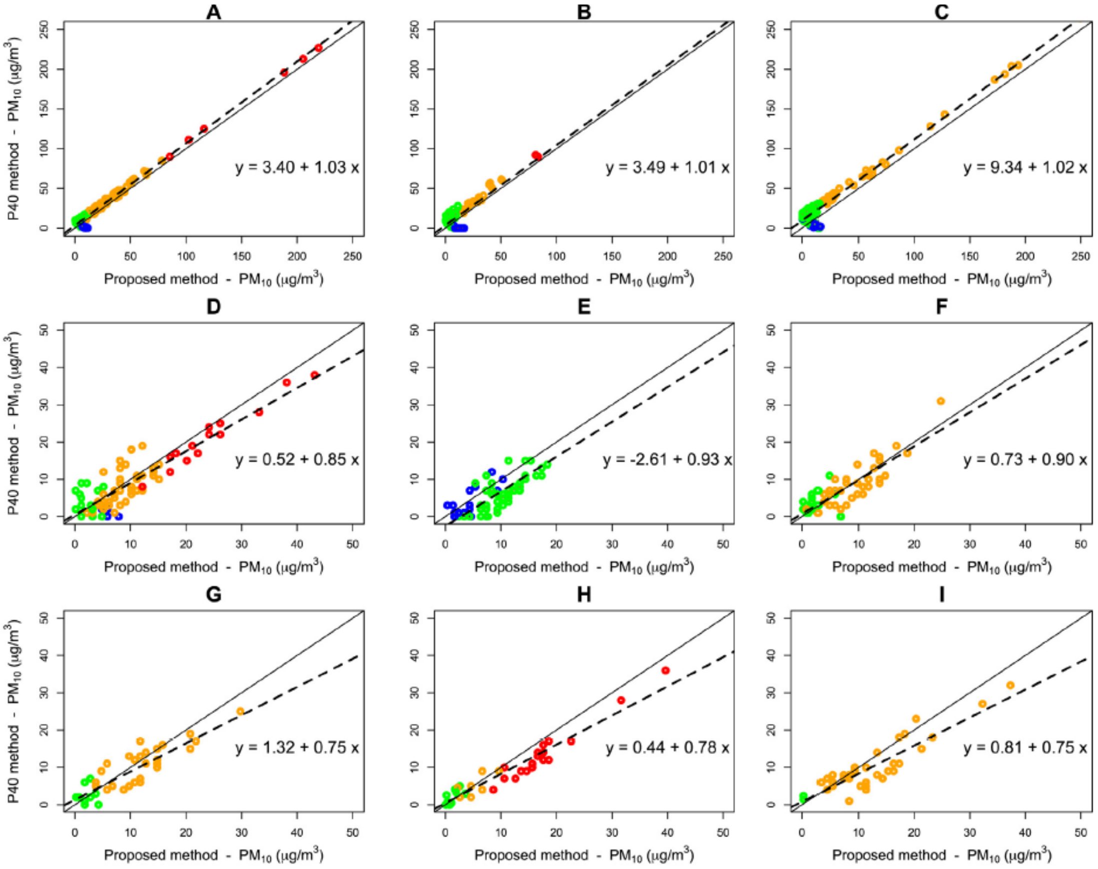
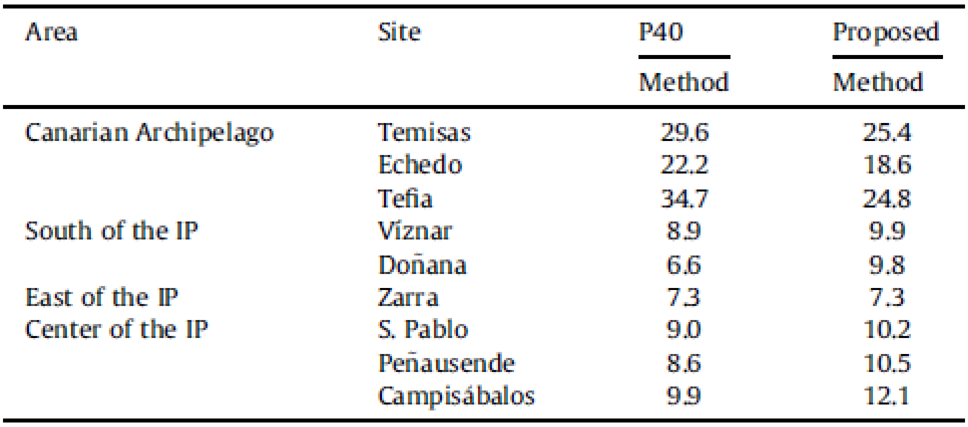


Fig. 6. Comparison between both methods for estimating the PM$_{10}$ contributions when desert outbreaks are detected. Dotted colours indicate the assignment of every quantity to a regime (the same colour code as in Fig. 2). The regression (dashed line) indicates the discrepancy between both methods and the black line indicates a hypothetically perfect correlation between them. Such discrepancy (in µg/m^3) is shown by a simple linear regression, where the equations regress the P40 method (y) on the proposed method (x). From **A** to **I**: Temisas, Echedo, Tefia, Viznar, Doñana, Zarra, San Pablo, Peñausende and Campisábalos sites, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Estimations of average PM_{10} contributions on days when African episodes are detected during 2013 using both methods (in µg/m^3) at Canarian archipelago and Iberian Peninsula (IP) sites.



Area	Site	P40	Proposed
		Method	Method
Canarian Archipelago	Temisas	29.6	25.4
	Echedo	22.2	18.6
	Tefia	34.7	24.8
South of the IP	Víznar	8.9	9.9
	Doñana	6.6	9.8
East of the IP	Zarra	7.3	7.3
Center of the IP	S. Pablo	9.0	10.2
	Peñausende	8.6	10.5
	Campisábalos	9.9	12.1

4. Conclusions

Properties and uses of a new SA methodology based on the grouping of TS are presented as well as a method for estimating the PM_{10} contributions from deserts. The results of the application of HMMs on daily average PM_{10} concentrations collected during different time periods at background sites from the Iberian Peninsula and some archipelagos were analysed. Net contributions due to different sources, contributions to the annual PM10 mean of every regime and probability of change among regimes at the Temisas site were estimated, after defining the regimes of its TS. This site is characterised by high PM_{10} contributions from the dry regions from North Africa. The first regime is proposed as an indicator of the background pollution in the analysed sites, taking into consideration the atmospheric variations in the time scale.

Regime defining for every TS is a previous and necessary step when this modelling is applied and the consensus of experts is necessary. These definitions provide a formal theoretical grounding to background pollutions fractions introduced by other authors, which must be considered when plans for the improvement of air quality are to be designed. The study of the regimes on a spatial scale helps to distinguish and quantify the different source contributions in geographical areas, although such studies must be complemented by other types of modelling to gain more robust SA deductions. The annual contribution of North African episodes to the PM10 mean value in the Canarian Archipelago coincides markedly with the same estimation made using the P40 method, applied by other authors. By adding a temporal scale to this analysis, the detection of new source contributions or the alteration of the expected ones is enabled in such areas of study.

The introduced method for estimating contributions from deserts seems to correct the net load of PM_{10} given by the P40 method and attributes less impact on areas suffering greater influence from African episodes on the daily PM_{10} concentrations.

The clustering of TS using HMMs provides an important methodological approach to exploratory methods used in SA but can also be used to complement other RM techniques that require time-consuming and expensive chemical speciation. The results of HMMs are easy to interpret and have a high degree of reproducibility. HMM implementation is also available through free software, which does not require advanced programming skills or advanced knowledge of statistics. The use of HMMs is therefore encouraged in the study of PM_{10} pollution.

Disclaimer

The authors declare that they have no actual or potential competing financial interest.

Acknowledgements

The data provided are based on the studies developed within the framework of a scientific collaboration between the Spanish Ministry of Agriculture, Food and the Environment, the Spanish National Research Council (CSIC) and the Spanish National Meteorology Service (AEMet). The authors are also grateful to CCDR do Centro, CCDR do Alentejo and Direcção Regional do Ambiente dos Açores for kindly providing their air quality data.

Appendix A

The aim of this appendix is to show how the values of μm and σm are calculated for the TS of the Temisas site (Table 1):

$$\begin{aligned}\mu_m &= \sum_{i=i}^{4} \pi_i \mu_i \\ &= (0.532 \cdot 10.325) + (0.265 \cdot 17.720) + (0.181 \cdot 42.752) \\ &\quad + (0.022 \cdot 153.256) \\ &= 21.298\end{aligned}$$

$$\begin{aligned}\sigma_m &= \left[\sum_{i=i}^{4} \left(\mu_i^2 + \sigma_i^2 \right) \pi_i - \mu_m^2 \right]^{1/2} \\ &= \left[\left(10.325^2 + 2.418^2 \right) \cdot 0.532 + \left(17.720^2 + 4.325^2 \right) \cdot 0.265 \right. \\ &\quad + \left(42.752^2 + 17.702^2 \right) \cdot 0.181 \\ &\quad \left. + \left(153.256^2 + 62.560^2 \right) \cdot 0.022 - 21.298^2 \right]^{1/2} \\ &= 25.315.\end{aligned}$$

Appendix B. Supplementary data

Supplementary material related to this article can be found at http://dx.doi.org/10.1016/j.atmosenv.2015.07.027.

References

Basford, K.E., Greenway, D.R., McLachlan, G.J., Peel, D., 1997. Standard errors of fitted component means of normal mixtures. Comput. Stat. 12, 1-17.

Belis, C.A., Karagulian, F., Larsen, B.R., Hopke, P.K., 2013. Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe. Atmos. Environ. 69, 94-108.

Belis, C.A., Larsen, B.R., Amato, F., Haddad, I.E., Favez, O., Harrison, R.M., Hopke, P.K., Nava, S., Paatero, P., Pre vô t, A., Quass, U., Vecchi, R., Viana, M., 2014. European Guide on Air Pollution Source Apportionment with Receptor Models. Joint Research Centre Reference Reports, Luxembourg. Report No.: EUR 26080 EN.

Cappe , O., Moulines, E., Ryde n, T., 2005. Inference in Hidden Markov Models. Springer, New York.

Dias, J.G., Vermunt, J.K., Ramos, S., 2010. Mixture hidden Markov models in finance research. In: Fink, A., et al. (Eds.), Advances in Data Analysis, Data Handling and Business Intelligence. Studies in Classification, Data Analysis and Knowledge Organization. Springer-Verlag, Heidelberg, pp. 451-459.

Directive, 2008. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on Ambient Air Quality and Cleaner Air for Eur¼ope [Internet] [cited 2014 Dic 23]. Available from: http://eur-lex.europa.eu/ LexUriServ/LexUriServ.do?uri OJ:L:2008:152:0001:0044:En:PDF.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood estimation from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. 39, 1e38.

Draxler, R.R., Rolph, G.D., 2003. HYSPLIT (Hybrid Single-particle Lagrangian Integrated Trajectory). NOAA Air Resources Laboratory, Silver Spring, MD. Model Access via NOAA ARL READY Website. http://ready.arl.noaa.gov/HYSPLIT.php.

Dong, M., Yang, D., Kuang, Y., He, D., Erdal, S., Kenski, D., 2009. PM2.5 concentration prediction using hidden semi-Markov model-based times series data mining. Expert Syst. Appl. 36, 9046-9055.

Escudero, M., Querol, X., Pey, A., Alastuey, A., Pe rez, N., Ferreira, F., Alonso, S., Rodríguez, S., Cuevas, E., 2007a. A methodology for the quantification of the net African dust load in air quality monitoring networks. Atmos. Environ. 41, 5516-5524.

Escudero, M., Querol, X., A vila, A., Cuevas, E., 2007b. Origin of the exceedances of the European daily PM limit value in regional background areas of Spain. Atmos. Environ. 41, 730e744.

EMEP, 2014. Transboundary Particulate Matter, Photo-oxidants, Acidifying and Eutrophying Components. Norwegian Meteorological Institute. EMEP Status Report 2014.

EN 12341, 1998. Air Quality e Determination of the PM10 Fraction of Suspended Particulate Matter - Reference Method and Field Test Procedure to Demonstrate Equivalence of Measurement Methods.

Flexer, A., Sykacek, P., Rezek, I., Dorffner, G., 2002. An automatic, continuous and probabilistic sleep stager based on a hidden Markov model. Appl. Artif. Intell. 16, 199-207.

Harte, D., 2015. HiddenMarkov: Hidden Markov Models, 2010. R Package Version 1.8-3.

Himmelmann, L., 2010. HMM: HMM-hidden Markov Models. Scientific Software Development, 2010. R Package Version 1.0.

Lenschow, P., Abraham, H.-J., Kutzner, K., Lutz, M., Preuß, J.-D., Reichenbächer, W., 2001. Some ideas about the sources of PM10. Atmos. Environ. 35 (1), S23-S33.

Pérez, N., Querol, X., Alastuey, A., Alonso-Pe rez, S., Cuevas, E., Orío, A., Reina, F., Pallare s, M., Salvador, P., Artín˜ano, B., de la Rosa, J., 2014. Episodios Naturales de Partículas 2013. CSIC, AEMet, Ministerio de Medio Agricultura, Alimentacio n y Medio Ambiente-Subdireccio n General de Calidad del Aire y Medio Ambiente Industrial.

Pey, N., Pe rez, N., Querol, X., Alastuey, A., Alonso-Pe rez, S., Cuevas, E., Moral, A., Jime nez, S., Pallare s, M., Salvador, P., Artín˜ano, B., de la Rosa, J., 2013. Episodios Naturales de Partículas 2012. CSIC, AEMet, Ministerio de Medio Ambiente, Medio Rural y Marino-Subdireccio n General de Calidad del Aire y Medio Ambiente Industrial.

Pey, N., Querol, X., Alastuey, A., Alonso-Pe rez, S., Cuevas, E., Gonza lez-Ortiz, A., Jime nez, S., Pallare s, M., Salvador, P., Artín˜ano, B., de la Rosa, J., Monjardino, J., Ferreira, F., 2011. Episodios Naturales de Partículas 2010. CSIC, AEMet, Ministerio de Medio Ambiente, Medio Rural y Marino-Subdireccio n General de Calidad del Aire y Medio Ambiente Industrial. http://bit.ly/1CWTHRW.

Querol, X., Alastuey, A., Rodríguez, S., Viana, M.M., Artín˜ano, B., Salvador, P., Mantilla, E., García do Santos, S., Fernandez Patier, R., de la Rosa, J., Sanchez de la Campa, A., Menendez, M., Gil, J.J., 2004. Levels of particulate matter in rural, urban and industrial sites in Spain. Sci. Total Environ. 334e335, 359-376.

Querol, X., Pey, J., Pandolfi, M., Alastuey, A., Cusack, M., Pe rez, N., Moreno, T., Viana, M., Mihalopoulos, N., Kallos, G., Kleanthous, S., 2009. African dust contributions to mean ambient PM10 mass-levels across the Mediterranean Basin. Atmos. Environ. 43, 4266-4277.

Querol, X., Alastuey, A., Pey, J., Escudero, M., Castillo, S., Orío, A., Pallare s, M., Jime nez, S., Ferreira, F., Marques, F., Monjardino, J., Cuevas, E., Alonso, Artíano, B., Salvador, P., de la Rosa, J., 2013a. Procedimiento para la identificacio n de episodios naturales de PM10 y PM2.5, y la demostracio n de causa en lo referente a las superaciones del valor límite diario de PM10. Instituto de Di agnostico Ambiental y Estudios del Agua (IDAEA), CSIC, Universidad Nova de Lisboa, AEMet-Izan˜a, CIEMAT, Universidad de Huelva. Ministerio de Medio Ambiente, Medio Rural y Marino, Ministe rio Do Ambiente, Ordenamiento Do Territorio e Desenvolvimiento Regional (Portugal), Age^ncia Portuguesa do Ambiente (Portugal).

Querol, X., Alastuey, A., Pey, J., Escudero, M., Castillo, S., Gonza lez-Ortiz, A., Pallare s, M., Jime nez, S., Cristo bal, A., Ferreira, F., Marques, F., Monjardino, J., Cuevas, E., Alonso, Artíano, B., Salvador, P., de la Rosa, J., 2013b. Methodology for the Identification of Natural Episodes in PM10 and PM2.5, and Justification with Regards to the Exceedances of the PM10 Daily Limit Value. Instituto de Diagnostico Ambiental y Estudios del Agua (IDAEA), CSIC, Universidad Nova de Lisboa, AEMet-Izan˜a, CIEMAT, Universidad de Huelva. Ministerio de Medio Ambiente, Medio Rural y Marino, Ministe rio Do Ambiente, Ordenamiento Do Territorio e Desenvolvimiento Regional (Portugal), Agência Portuguesa do Ambiente (Portugal).

R Core Team, 2013. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org/.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257-286.

Rodríguez, S., Querol, X., Alastuey, A., Kallos, G., Kakaliagou, O., 2001. Saharan dust contributions to PM10 and TSP levels in Southern and Easter Spain. Atmos. Environ. 35, 2433-2447.

Salvador, P., Alonso-Pe rez, S., Pey, J., Artín˜ano, B., de Bustos, J.J., Alastuey, A., Querol, X., 2014. African dust outbreaks over the western Mediterranean Basin: 11-year characterization of atmospheric circulation patterns and dust source areas. Atmos. Chem. Phys. 14, 6759-6775.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461e464. Viana, M., Querol, X., Alastuey, A., Cuevas, E., Rodríguez, S., 2002. Influence of African dust on the levels of atmospherics particulates in the Canary Islands air quality network. Atmos. Environ. 36, 5861e5875.

Viana, M., Kuhlbusch, T.A.J., Querol, X., Alastuey, A., Harrison, R.M., Hopke, P.K., Winiwarter, W., Vallius, M., Szidat, S., Pre voˆt, A.S.H., Hueglin, C., Bloemen, H., Wåhlin, P., Vechhi, R., Miranda, A.I., Kasper-Giebl, A., Maenhaut, W., Hitzenberger, R., 2008. Source apportionment of particulate matter in Europe: a review of method and results. J. Aerosol Sci. 39, 827-849.

Viana, M., Pey, J., Querol, X., Alastuey, A., de Leeuw, F., Lükewille, A., 2014. Natural sources of atmospheric aerosols influencing air quality across Europe. Sci. Total Environ. 472, 825-833.

Visser, I., Raijmakers, M.E.J., Molenaar, P.C.M., 2002. Fitting hidden Markov models to psychological data. Sci. Program. 10, 185-199.

Visser, I., Raijmakers, M.E.J., Maas, H.L.J., 2009. Hidden Markov models for individual time series. In: Valsiner, J., Molenaar, P.C.M., Lyra, M.C.D.P., Chaudhary, N. (Eds.), Dynamic Process Methodology in the Social and Developmental Sciences. Springer, Heidelberg, pp. 269-289.

Visser, I., Speekenbrink, M., 2010. depmixS4: an R package for hidden markov models. J. Stat. Softw. 36 (7), 1-21.

Visser, I., 2011. Seven things to remember about hidden Markov models: a tutorial on Markovian models for time series. J. Math. Psychol. 55, 403-415.

Wilks, D.S., 2006. Statistical Methods in the Atmospheric Sciences, second ed. Academic Press, Burlington.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J., Steinberg, D., 2008. Top 10 algorithms in data mining. Knowl. Inf. Syst. 14, 1-37.

Zucchini, W., MacDonald, I., 2009. Hidden Markov Models for Time Series. An Introduction Using R. CRC Press, New York.