Supplementary material

# Estimation of particulate matter contributions from desert outbreaks in European Mediterranean countries (2015-2018) using the time series clustering method

Álvaro Gómez-Losada[*, a] and José C. M. Pires [b]

[a] Departamento de Estadística e Investigación Operativa, Facultad de Matemáticas, Universidad de Sevilla, Sevilla, España.

[b] Laboratório de Engenharia de Processos, Ambiente e Energia (LEPABE), Departamento de Engenharia Química, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal.

---

## Content



[*] Corresponding author.
E-mail addresses: aglosada@us.es, alvaro.gomez.losada@gmail.com (Á. Gómez-Losada); jcpires@fe.up.pt (José Carlos M. Pires).

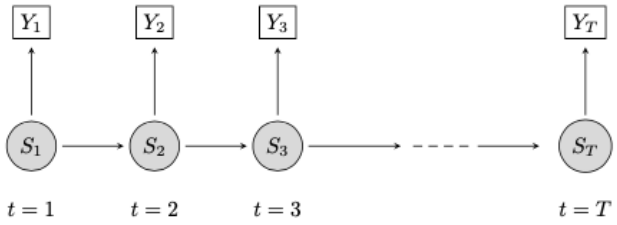
**SM.1** Definition and elements of Hidden Markov Models.

In this section, a some more technical Hidden Markov Models (HMM) definition is given than in the manuscript. HMM is a doubly embedded stochastic process in which one is an underlying Markov chain, a series of hidden states (state variables); and the other one is the observation sequence (the temporal series) determined by the current hidden state of a given Markov chain, the outcome of a certain hidden state. Only the observations are unhidden.

Let $Y_{1:T} \coloneqq (Y_1, Y_2, \dots, Y_T)$ be a time series of length $T$ and $S_{1:T} \coloneqq (S_1, S_2, \dots, S_T)$ *states variables*, these latter hidden to the observer. These variables $S_t$ are elements from a finite set $\mathcal{S} = \{1, \dots, n\}$ such that it can be written $S_t = i, i \in \mathcal{S}$. The set $\mathcal{S}$ is called the *state-space* of the HMM, and *n* is the number of states of the model. The observations $Y_t$ are dependent on the state variables $S_t$ such that the distribution of $Y_t$ can be written as: $f_i(Y_t) \coloneqq f(Y_t|S_t = i)$. Because the set $\mathcal{S}$ is finite, this means that the marginal distribution of the data (the temporal serie) is a mixture distribution with *n* components:

$$f\,(Y_t) = \sum_{i=1}^{n} \pi_i \; f_i(Y_t)$$

where $\pi_i$ are the mixing proportions in which every component (cluster) occurs with the constraints: $\pi_i \geq 0$, for $i \in \{1, \dots, n\}$ which sum to 1. Each $f_i(\cdot)$ is the conditional distribution of the data in component *i* and is called as *state-dependent* distribution of the model. The below figure shows the dependence graph in a HMM:




The Markov property in HMMs is determined by the dependence between the states, that can be expressed:

$$P(S_t|S_1, \dots, S_{t-1}) = P(S_t|S_{t-1})$$

which are referred as *transition probabilities*. These latter probabilities are denoted by the matrix $A(t)$ with entries:

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

and constraints:

$$\sum_{j=1}^{n} a_{ij}(t) = 1, \text{for each } \; i; a_{ij} \geq 0$$

The fundamental assumption of a dependent mixture model is that at any time point, the observations are distributed as a mixture with *n* components (clusters or states), and that time-dependencies between the observations are due to time-dependencies between the mixture

components (i.e., transition probabilities between the components) (Visser and Speekenbrik, 2010). A comprehensive account of the HMMs can be found at Visser et al. (2011) and Zucchini and MacDonald (2009).

**SM.2** Computational implementation of Hidden Markov Models.

The code below was used to fit the TS from the different monitoring sites reported in the paper (Section 2.1) using the *depmixS4* package in *R*. Those readers not familiar with this R computational environment are referred to http://www.r-project.org and http://cran.r-project.org/manuals.html where introductory material is available. This code is presented in a readable and comprehensive form although more sophisticated procedures have been used to manage all the data used in this work. These latter procedures are available to those interested readers under request.

```
library(depmixS4)

# The object named "data" contains the TS as a vector.
# It is necessary to create a data frame:

sample<-data.frame(y=data)

# Now, using the depmix function, we create 4 different models to fit the TS,
# from one hidden state (ns=1) to seven (ns=4).

m1<-depmix(y~1, data=sample, ns=1, ntimes=nrow(sample))
m2<-depmix(y~1, data=sample, ns=2, ntimes=nrow(sample))
m3<-depmix(y~1, data=sample, ns=3, ntimes=nrow(sample))
m4<-depmix(y~1, data=sample, ns=4, ntimes=nrow(sample))

# Every model is fit to obtain their parameters by means of the "fit" function:

fm1<-fit(m1, em=em.control(maxit=2000, tol=1e-08, crit="relative"))
fm2<-fit(m2, em=em.control(maxit=2000, tol=1e-08, crit="relative"))
fm3<-fit(m3, em=em.control(maxit=2000, tol=1e-08, crit="relative"))
fm4<-fit(m4, em=em.control(maxit=2000, tol=1e-08, crit="relative"))


# The BIC statistic is calculated for every model using the "BIC" function.
# Values are stored in the object named "bic" as a vector.

bic<-c(BIC(fm1),BIC(fm2),BIC(fm3),BIC(fm4))

# Now, the best model providing the best fit to the data
# must be selected. We look for the model with lowest BIC value:

> which.min(bic)
[1] 4

# The model with 4 states, "fm4", describes the TS data best.
# Now, we can obtain the probability transition matrix and the
# parameters of the 4 Gaussian distributions.
# The "summary" function is used:

> summary(fm4)
Initial state probabilties model
pr1 pr2 pr3 pr4
  1   0   0   0

Transition matrix
                toS1         toS2       toS3         toS4
fromS1  8.706526e-01 0.1099947038 0.01935269 6.945374e-56
fromS2  2.213438e-01 0.6793470074 0.06635697 3.295224e-02
fromS3  2.534007e-02 0.1697684571 0.78155134 2.334013e-02
fromS4 1.312388e-110 0.0001419835 0.66654146 3.333166e-01

Response parameters
Resp 1 : gaussian
    Re1.(Intercept)    Re1.sd
St1        10.32563  2.418200
St2        17.72293  4.346021
St3        42.75554 17.701547
St4       153.25253 62.565625

# Calculate the "pi" values of every Gaussian:

probs<-posterior(fm4)
```

```
colMeans(probs[,2:5])

> colMeans(probs[,2:5])
        S1         S2         S3         S4
0.53226225 0.26543964 0.18137000 0.01992811

# To solve the significant digits problem, the "S4" value can be obtained
# as 1-(S1+S2+S3).
# Missing some little precision must be assumed.
# The same can be applied to last column in the transition probability matrix.
```

## **SM.3** Analysed monitoring sites in Section 3.4 of the manuscript.

**Table SM.3a**. Monitoring station selected from each country for $PM_{10}$ analysis (Figure 5). m.a.s.l: meters above sea level.

| Country | EoI Code | Year | m.a.s.l |
|---|---|---|---|
| TR | R010213 | 2015 | 107 |
| ES | ES2002A | 2015 | 160 |
| PT | PT02019 | 2017 | 60 |
| BA | BA0001G | 2016 | 970 |
| FR | FR18039 | 2016 | 33 |
| HR | HR0011A | 2015 | 0 |

**Table SM.3b**. Monitoring station selected from each country for $PM_{2.5}$ analysis (Figure 6). m.a.s.l: meters above sea level.

| Country | EoI Code | Year | m.a.s.l |
|---|---|---|---|
| TR | R160613 | 2018 | 87 |
| ES | ES1802A | 2016 | 995 |
| PT | PT04006 | 2017 | 187 |
| FR | FR34038 | 2015 | 182 |
| HR | HR0011A | 2016 | 0 |

**SM.4** Supplementary material references.

Visser, I., Speekenbrink, M. depmixS4: An R package for Hidden Markov Models. Journal of Statistical Software 2010, 36(7):1-21. Available from: http://cran.r-project.org/web/packages/depmixS4/vignettes/depmixS4.pdf.

Visser, I. Seven things to remember about hidden Markov models: A tutorial on Markovian models for time series. Journal of Mathematical Psychology 2011, 55: 403-415. http://doi.org/10.1016/j.jmp.2011.08.002.

Zucchini, W., MacDonald, I. Hidden Markov Models for Time Series: An Introduction Using R. Monographs on Statistics and Applied Probability. CRC Press, Boca Raton. 2009.