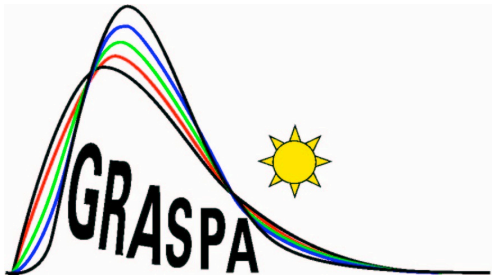
Web Working Papers
by
The Italian Group of Environmental Statistics



Gruppo di Ricerca per le Applicazioni della Statistica
ai Problemi Ambientali

www.graspa.org

A hidden Markov model for pollutants exceedances counts

Francesco Lagona and Antonello Maruotti

GRASPA Working paper n.33, January 2009

A hidden Markov model for pollutants exceedances counts

Francesco Lagona and Antonello Maruotti *

Abstract

Pollutant exceedances at the sites of a monitoring network are modeled via a hidden Markov model (HMM), where state transition probabilities depend on meteorological covariates and the observations are modelled in a generalized linear model framework, whose parameters depend on the hidden states. The estimated hidden states summarize the shape of the multinomial distribution of the pollutants at each time; define a model-driven air quality index and provide some useful insights into the processes driving pollutant exceedances. Model estimation is carried out by using a recursive forward-backward procedure, by taking a maximum likelihood approach. Parametric bootstrap is exploited to compute variances of parameter estimates. We model sequences of several air pollutants from monitoring stations in Rome.

1 Introduction

Air quality standards are referred to thresholds above which pollutants concentrations are considered to have serious effects on human health and the environment (WHO, 2006). In large urban areas, exceedances of these standards are usually recorded through a monitoring network, where concentrations of a number of pollutants are measured at different sites. Exceedances are typically counted to determine compliance with air quality regulations and to study short/long-term effects of air pollution exposure. In particular, daily number of stations reporting a violation of standards (i.e., the count of exceedances) is often used as a simple air quality indicator to detect air pollution episodes that occurred during a period of interest.

Whereas exceedances counts are helpful to communicate air conditions to the general public, they suffer of a number of obvious drawbacks.

First, exceedances data are often unbalanced, because urban monitoring networks are established for a variety of purposes and different stations typically measure non-homogeneous subsets of pollutants. Some pollutants are often monitored by fewer stations, while concentrations of other pollutants

*DIPES - University of Roma Tre and GRASPA Research Unit of Rome, Italy

are available at a larger number of sites. Moreover, some of these stations are often not in operation during part of the observation period and, hence, exceedances counts are daily computed on the basis of a time-varying number of stations. Exceedances data are therefore difficult to interpret if they are not examined conditionally on the number of stations that are available each day, for each pollutant.

Second, exceedances data are of little help in the evaluation of environmental risk, if the analysis is not adjusted for weather conditions. Although at present the formation and evolution of air pollution episodes in urban areas is only understood in general terms, it is well known that meteorological covariates have a significant influence on air quality. The severity of an episode should be therefore assessed by comparing exceedances data between days that share similar weather conditions. In other words, exceedances should be examined conditionally on meteorological covariates, if an analysis aims at addressing issues of environmental justice.

These drawbacks can be overcame by a conditional analysis of multivariate exceedances counts, that focuses on the estimation of weather-specific, joint exceedances probabilities. This requires the statistical modelling of multivariate exceedances counts.

In a regression context, a simple approach could be pursued, by modelling each pollutant separately and fitting a binomial regression to each time series of exceedances counts, where the probability of a pollutant-specific exceedance depends on the available meteorological covariates through a suitable link function (Kütchenhoff and Thamerus, 1996). This approach relies on an independence assumption between pollutant exceedances and can be unrealistic if significant interactions between pollutant concentrations occur, as it is often the case. As an alternative to independent models, finite mixtures of generalized linear models (Wang and Putermann, 1998) provide a parsimonious approach to capture interactions between exceedances of different pollutants. In fact, in a finite mixture framework, the dependence structure between pollutants exceedances is modelled by assuming that exceedances are conditionally independent, given a latent class. In other words, exceedances probabilities are represented as convex combinations of reference probabilities that can be interpreted as reference *states* of the air.

When a multivariate time series of exceedances counts is fitted by a mixture of binomial regressions, the temporal structure of the data is ignored and exceedances observed in different days are treated as independent samples. Temporal independence is often a strong assumption in environmental time series such as pollutant exceedances counts, especially when a conditional analysis is carried out by using a number of meteorological covariates that capture a small portion of the data variability. Hidden Markov models (HMM; Cappé et al. 2005) are parsimonious mixture models that account for the temporal dependence structure of the data, by assuming that temporal transitions from a state to another are driven by a Markov chain.

In this paper, we focus on the statistical analysis of the multivariate time series of exceedances counts, obtained when exceedances of a number of pollutants are recorded by a urban monitoring network. Conditionally on the available meteorological covariates and the number of stations operating each day, we aim at detecting typical patterns of exceedances probabilities that can be interpreted as reference air quality states. We specifically exploit a non-homogeneous, multivariate hidden Markov model, where exceedances counts are sampled from conditionally independent binomial distributions, given the covariates and a latent air quality state. Temporal transitions between different states are modelled by a non-homogeneous Markov chain, driven by covariate-specific transition probabilities.

The proposed model extends the specification of finite mixtures of generalized linear models (Wang and Putermann, 1998), to account for serially dependent data, and belongs to the family of the hidden Markov models discussed by MacDonald and Zucchini (1997) to analyze categorical time series. It can also be viewed as a multi-pollutants generalization of the hidden Markov model developed by Hughes, Guttorp and Charles (1999) in a study of precipitation occurrences.

After describing the environmental data used in this study (Section 2), the specification of a multivariate HMM is outlined in Section 3. Section 4 is devoted to discuss relevant computational details for likelihood-based parameter estimation, while Section 5 illustrates an application to air quality measurements from the monitoring network in Rome over the period January - November 2000. Section 6 provides some concluding remarks.

2 Data

Our analysis is based on daily, multivariate counts of exceedances of air quality standards, as computed from hourly pollutants concentrations that are typically available from the monitoring network in a large urban area.

In the application discussed in the present paper, we considered the concentrations data of particulate matter (PM_{10}), nitrogen dioxide (NO_2) and ozone (O_3), reported by the monitoring network of Rome, during a period of 328 days (1/3/2000 - 11/25/2000). The network includes 4 stations that measure particulates, 9 stations that measure nitrogen dioxide concentrations and 4 stations that measure ozone. The resulting $17 = 4 + 9 + 4$ time series include several missing values, because some of the stations were not in operation during parts of the considered period. Each pollutant is however daily recorded by at least one station, during the study period.

Daily exceedances counts were computed by counting the number of stations where (i) the 24-hour average concentration of particulate matter was above the threshold of $50\mu g/m^3$, (ii) the maximum hourly concentration of nitrogen dioxide was above the level of 200 $\mu g/m^3$ and (iii) the maximum

8-hour moving average of ozone concentrations exceeded the level of 120 $\mu g/m^3$. According to these cut-offs, we obtained three time series (one for each considered pollutant) of the daily number of stations where a violation occurred.

The count data are displayed in Figure 1, together with the number of stations operating each day.

Figure 1 about here

According to the above definitions of standards, we notice that ozone episodes occur during summer, while exceedances of particulate matter and nitrogen dioxide are scattered along the whole year. The severity of pollution episodes should however be assessed by taking in account the time-varying number of functioning stations. We also notice that the zeros in these time series are not structural, because each pollutant was daily recorded by at least one station, during the study period.

Figure 2 about here

In this paper, we aim at modelling the three time series in Figure 1, conditionally on a multivariate time series of covariates that summarize weather conditions. Figure 2 shows the standardized daily averages of temperature, humidity, pressure and global radiation recorded by the network during the study period, and used in this study to adjust our analysis for weather conditions.

3 Modeling exceedances counts

In this study, hourly concentrations of a generic pollutant i, $i = 1, \ldots, I$, recorded by a monitoring station h, $h = 1, \ldots, H$ during day t, $t = 1, \ldots, T$, are summarized by a daily binary variable $y_{iht} = 1$ in case of an exceedance and 0 otherwise. We aim at modelling the multivariate time series of exceedances counts of I pollutants, say $\mathbf{y}_t = \{y_{it}, i = 1 \ldots I, t = 1 \ldots T\}$, where $y_{it} = \sum_{h=1}^{n_{it}} y_{iht}$ indicates the number of exceedances of pollutant i, observed at the n_{it} stations operating in day t.

A HMM specification of the distribution of $\mathbf{y}_t$ allows for a parsimonious specification of both the temporal dependence between exceedances and the interactions between pollutants. In the following, we describe a two-state, discrete time, non-homogeneous hidden Markov model (Wang and Puterman, 2001; Cappé et al., 2005) that we propose to model multivariate exceedances counts.

Specifically, we consider a vector of latent states $s_{0:T} = (s_0, s_1, \ldots, s_T)$ and write the distribution of the observed data $\mathbf{y}_{0:T} = (\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_T)$ as a marginal distribution

$$Pr(\mathbf{y}_{0:T}) = \sum_{s_0} \sum_{s_1} \cdots \sum_{s_T} Pr(\mathbf{y}_{0:T}, s_{0:T}).$$

We write the joint probability of the observed and the hidden processes as

$$Pr(\mathbf{y}_{0:T}, s_{0:T}) = Pr(\mathbf{y}_{0:T} \mid s_{0:T}) Pr(s_{0:T}) \tag{1}$$

and assume that

$$Pr(\mathbf{y}_{0:T} \mid s_{0:T}) = \prod_{t=0}^{T} Pr(\mathbf{y}_t \mid s_{0:T}), \tag{2}$$

and

$$Pr(\mathbf{y}_t \mid s_{0:T}) = Pr(\mathbf{y}_t \mid s_t). \tag{3}$$

In particular, exceedances counts of different pollutant are assumed conditionally independent, given the latent state, and modelled by the product of I binomial distributions, as follows:

$$Pr(\mathbf{y}_{0:T} \mid s_{0:T}) = \prod_{t=0}^{T} Pr(\mathbf{y}_t \mid s_t) = \prod_{t=0}^{T} \prod_{i=1}^{I} \binom{n_{it}}{y_{it}} \pi_{it}^{y_{it}} (1 - \pi_{it})^{n_{it} - y_{it}}. \tag{4}$$

The canonical parameter $\pi_{it} = E(y_{it} \mid s_t, \mathbf{x}_t)$, i.e. the exceedance probability for pollutant i in day t, is assumed to depend on the weather conditions of that day, through a logit link function

$$\text{logit}(E(y_{it} \mid s_t, \mathbf{x}_t)) = \beta_{i0}(s_t) + \sum_{l=1}^{p} x_{tl} \beta_{il}(s_t)$$

where $\mathbf{x}_t^T = (x_{t1}, x_{t2}, \ldots, x_{tp})$ is a set of p atmospheric covariates and $\boldsymbol{\beta}_i(s_t) = (\beta_{i0}(s_t), \beta_{i1}(s_t), \ldots, \beta_{ip}(s_t))$ is an outcome-specific vector of regression parameters, depending on the latent state. This conditional independence model for $Pr(\mathbf{y}_t \mid s_t)$ assumes that the y_{it} are independent conditional on the latent state; unconditionally, counts y_{it} will be correlated owing the influence of the common latent state.

To complete the HMM specification, we assume that the joint distribution of the states sequence is driven by a two-states Markov chain with state space $\mathcal{S} = (0, 1)$, as follows:

$$Pr(s_{0:T}) = Pr(s_0) \prod_{t=1}^{T} Pr(s_t \mid s_{t-1}) = \delta_{s_0} \prod_{t=1}^{T} q_{s_{t-1} s_t} \tag{5}$$

where $\delta_{s_0} = Pr(s_0)$ and $q_{s_{t-1} s_t} = Pr(s_t \mid s_{t-1})$. In particular, the transitions probabilities at day t, $q_{s_{t-1} s_t}$, are assumed to be non-homogeneous and modelled as functions of linear predictors, through a logit transformation. We specifically assume that

$$\text{logit}(q_{00}) = \log\left(\frac{Pr(S_t = 0 \mid S_{t-1} = 0, \mathbf{x}_t)}{Pr(S_t = 1 \mid S_{t-1} = 0, \mathbf{x}_t)}\right) = \gamma_{00} + \sum_{l=1}^{p} x_{tl}\gamma_{0l} \quad (6)$$

$$\text{logit}(q_{11}) = \log\left(\frac{Pr(S_t = 1 \mid S_{t-1} = 1, \mathbf{x}_t)}{Pr(S_t = 0 \mid S_{t-1} = 1, \mathbf{x}_t)}\right) = \gamma_{10} + \sum_{l=1}^{p} x_{tl}\gamma_{1l} \quad (7)$$

where $\boldsymbol{\gamma}_{s_t} = (\gamma_{s_t 0}, \gamma_{s_t 1}, \ldots, \gamma_{s_t p})$ is a vector of state-specific regression parameters.

The two key assumptions of our model are hence the conditional independence between contemporaneous pollutants events, given the state (equation 4), and the Markovian dependence structure of the state sequence (equation 5).

Figure 3 about here

These two assumptions specifies the dependence structure of the HMM considered in this paper, as depicted by Figure 3, which displays the association graph of the model.

4 Estimation

Taking into account the assumptions defined in Section 3, we will define $L(\cdot)$ as the likelihood function. We can derive an expression for the likelihood in terms of multiple sums:

$$\begin{aligned} L(\cdot) &= \sum_{s_0}\sum_{s_1}\cdots\sum_{s_T} \Pr(\mathbf{Y}_{0:T} = \mathbf{y}_{0:T}, S_{0:T} = s_{0:T}) \\ &= \sum_{s_0}\sum_{s_1}\cdots\sum_{s_T} \delta_{s_0}\prod_{t=1}^{T} q_{s_{t-1}s_t}\prod_{i=1}^{I}\prod_{t=0}^{T} Pr(y_{it} \mid s_t) \end{aligned} \quad (8)$$

As it stands, this expression is of little or no computational use, because it has 2^T terms and cannot be evaluated except for very small T. Clearly, a more efficient procedure is needed to perform the calculation of the likelihood. The problem of computing these factors may be addressed through the Forward-Backward procedure (Baum et al., 1970; for a brief review see Welch, 2003). Let us start considering the forward variable

$$\alpha_t(j) = Pr(\mathbf{y}_{0:t}, S_t = j), \quad (9)$$

which represents the joint probability of the partial observed sequence until time t and state j at time t. Now, recursive factorization of $\alpha_t(j)$ is given inductively:

$$\alpha_0(j) = Pr(\mathbf{y}_0, S_0 = j) = \delta_j Pr(\mathbf{y}_0 \mid S_0 = j), \quad j = 0, 1. \quad (10)$$

$$\alpha_{t+1}(k) = Pr(\mathbf{y}_{0:t+1}, S_{t+1} = k) =$$
$$\left[\sum_{j=0}^{1} \alpha_t(j) q_{jk}\right] Pr(\mathbf{y}_{t+1} \mid S_{t+1} = k), \quad j,k = 0,1; 0 \leq t \leq T-1. \tag{11}$$

where $Pr(\mathbf{y}_t \mid S_t = j)$ is defined as in (4). As a by-product of the forward recursion, we obtain that the likelihood can be written as

$$L(\cdot) = \sum_{j=0}^{1} \alpha_T(j). \tag{12}$$

A reverse time recursion exists for the backward variable which is defined as

$$\tau_t(j) = Pr(\mathbf{y}_{t+1:T} \mid S_t = j), \tag{13}$$

i.e. the probability of the partial observation sequence from $t+1$ to the end, given state j at time t. Again we can solve for $\tau_t(j)$ inductively, as follows:

$$\tau_T(j) = 1, \quad j = 0,1. \tag{14}$$

$$\tau_t(j) = \sum_{k=0}^{1} q_{jk} Pr(\mathbf{y}_{t+1} \mid S_t = k) \tau_{t+1}(k), \quad j,k = 0,1; t = T-1, T-2, \ldots, 0. \tag{15}$$

The log-likelihood can be evaluated recursively, even for very long observed sequences; hence it is feasible to perform parameter estimation for HMMs by direct numerical maximization of the log-likelihood function. The maximization can be accomplished by solving m separate maximization problems defined by starting from a fixed initial state (Leroux and Puterman, 1992). An EM algorithm to find model parameter estimates can be used (e.g. Leroux and Puterman, 1992; Hughes, 1997; Bilmes, 1998). In the EM framework, $\mathbf{y}_{0:T}$ is referred to as the incomplete data, $s_{0:T}$ is called the "missing" data, while $(\mathbf{y}_{0:T}, s_{0:T})$ is the complete-data. Given a particular sequence of states, the complete-data log-likelihood can be easily computed as

$$\begin{aligned} \ell^c(\cdot) &= \sum_{j \in \mathcal{S}} \eta^*_{j0} \log \delta_j + \sum_{t=1}^{T} \sum_{j \in \mathcal{S}} \sum_{k \in \mathcal{S}} \xi^*_{jkt} \log q_{jk} \\ &+ \sum_{t=0}^{T} \sum_{j \in \mathcal{S}} \eta^*_{jt} \log Pr(\mathbf{y}_t \mid S_t = j). \end{aligned} \tag{16}$$

where η^*_{jt} is 1 if $S_t = j$ and 0 otherwise, ξ^*_{jkt} is 1 if a transition from j to k occurred are t and 0 otherwise, and $Pr(\mathbf{y}_t \mid S_t = j)$ and q_{jk} are defined as in (4) and (6), respectively. Let us define the conditional expectation of

the complete log-likelihood function, $\mathcal{Q}(\cdot,\cdot)$, which is obtained replacing the components of the missing data by their conditional means:

$$\mathcal{Q}(\cdot,\cdot) = \sum_{j\in\mathcal{S}} \eta_{j0} \log \delta_j + \sum_{t=1}^{T} \sum_{j\in\mathcal{S}} \sum_{k\in\mathcal{S}} \xi_{jkt} \log q_{jk} + \sum_{i=1}^{I} \sum_{t=0}^{T} \sum_{j\in\mathcal{S}} \eta_{jt} \log f(\mathbf{y}_t \mid S_t = j). \tag{17}$$

It can be seen that it is easy to differentiate with respect to model parameters, add the Lagrange multipliers and solve, where

$$\eta_{jt} = Pr(S_t = j \mid \mathbf{y}_{0:T}), \tag{18}$$

the posterior probability, given the observed data, of being in state j at time t and with

$$\xi_{jkt} \quad = \quad Pr(S_{t+1} = k, S_t = j \mid \mathbf{y}_{0:T}) \tag{19}$$

the posterior probability that the unobserved sequence visited state j at time t and made a transition to state k at time $t+1$, given the observed individual sequence.

We can compute (17) using the forward and the backward variables defined in (9) and (13) considering that the first and the third parts of the (17) can be seen as smoothing probabilities, while the second one is a bivariate smoothing probability. In fact,

$$\eta_{jt} = \frac{\alpha_t(j)\tau_t(j)}{\sum_{j=0}^{1} \alpha_t(j)\tau_t(j)} \tag{20}$$

$$\xi_{jkt} = \frac{\alpha_t(j) q_{jk} f_k(y_{t+1}) \tau_t(k)}{\sum_{j=0}^{1} \sum_{k=0}^{1} \alpha_t(j) q_{jk} f_k(y_{t+1}) \tau_t(k)} \tag{21}$$

In the M-step, we update all model parameter estimates. The estimates of the initial probability corresponds to the smoothing probability:

$$\hat{\delta}_j = Pr(S_0 = j \mid y_{0:T}) = \eta_{j0}. \tag{22}$$

All the other estimated parameters are the roots of the following M-step equations:

$$\frac{\partial \mathcal{Q}}{\partial \gamma} = \sum_{t=1}^{T} \sum_{j=0}^{1} \sum_{k=0}^{1} \xi_{jkt} \frac{\partial \log q_{jk}}{\partial \gamma}; \tag{23}$$

$$\frac{\partial \mathcal{Q}}{\partial \beta} = \sum_{i=1}^{I} \sum_{t=1}^{T} \sum_{j=0}^{1} \eta_{jt} \frac{\partial \log f(y_{it} \mid S_t = j)}{\partial \beta}. \tag{24}$$

The resulting equations are thus weighted sums of score equations for generalized linear models with common weights ξ_{jkt} and η_{jt} respectively. The E- and M-steps are repeatedly alternated until the log-likelihood (relative) difference changes by an arbitrarily small amount.

However, while the EM algorithm is useful for obtaining maximum likelihood estimates in such situations, it does not provide readily produce standard errors for parameters estimates.

We computed standard errors of parameter estimates using parametric bootstrap, as standard errors based on the observed information matrix are often unstable (see e.g. McLachlan and Peel 2000). Specifically, we re-fitted the model to the bootstrap data that were simulated from the estimated model. This process was repeated R times, and the approximate standard error of each model parameter κ was computed by

$$\hat{se}_R = \left\{ \frac{1}{R-1} \sum_{r=1}^{R} [\hat{\kappa}(r) - \overline{\kappa}(R)]^2 \right\}^{1/2}, \tag{25}$$

where $\hat{\kappa}(r)$ is the estimate from the r-th bootstrap sample and $\overline{\kappa}(R)$ is the sample mean of all $\hat{\kappa}(r)$.

5 Results

Three time series of daily exceedances counts were jointly fitted by the hidden Markov model outlined in Section 3. Figure 4 displays the estimated pollutant-specific exceedance probabilities, plotted against the observed exceedances proportions, i.e. the number of violations recorded each day by the network, divided by the number of stations in operation.

Figure 4 about here

Maximum likelihood estimates of the parameters are reported in Table 1, which displays the estimated influence of weather conditions on the conditional (logit-transformed) exceedances probabilities, given the latent state of the air, and the influence of these covariates on transition probabilities. Estimates should be interpreted by recalling that covariates were standardized. In particular we notice that the two states feature different subsets of significant covariates.

Table 1 about here

Estimates displayed in Table 1 can be conveniently interpreted by computing log-odds of pollutant-specific exceedances and plotting the conditional linear predictors, given the state (Figure 5).

Figure 5 about here

In particular, Figure 5 displays the log-odds of an exceedance at baseline and at two extreme values of each standardized predictor (± 2). Under state 0, the baseline log-odds of an exceedance of particulate matter is greater than the log-odds of an exceedance of nitrogen dioxide, which is greater than the log-odds of an ozone exceedance. Under state 1, this ordering is reversed. While a graphical examination of the data would suggest to cluster days into days of acceptable air conditions and days of air pollution episodes, the model exploited in this study proposes a different classification, based on two reference patterns of exceedances. State 0 represents a reference patterns, where likely episodes of particulate matter and nitrogen dioxide are compensated by unlikely exceedances of ozone. On the other side, the reference pattern detected by state 1 is featured by likely episodes of ozone, compensated by small exceedances probabilities of the other two pollutants. Days of acceptable air conditions (unlikely exceedances) and severely polluted days (likely exceedances) are represented as mixtures of these two reference patterns.

As well as the estimated effects of covariates reported in Table 1 capture departures from the two reference exceedances patterns, due to changes in weather conditions; differences between pollutants exceedances probabilities can be either compensated or enhanced, in the presence of good or adverse weather conditions. Differences enhancements and compensations are however different under the two states. For example (Figure 5), under state 0 log-odds differences increase as pressure and humidity increase. Under state 1, these differences decrease as pressure and humidity increase. A specular situation occurs in the case of global radiation. The impact of temperature on exceedances probabilities is instead of a different type: under state 0, temperature has a negative influence on both ozone and nitrogen dioxide, while positively influences particulate exceedances; a reversed situation occurs under state 1.

Figure 6 about here

Figure 6 represents the conditional linear predictors of exceedance probabilities, as estimated in the study period. Air quality is represented as a composition of a background state of pollution episodes (state 0: red solid line) and a air quality state where episodes occur together with adverse weather conditions (state 1: black solid line). Under state 0 exceedances of particulate matter and nitrogen dioxide are likely to occur regardless of changes in weather conditions, but are compensated by very small probabilities of ozone exceedances. Under state 1, air pollution episodes are likely

to occur in the presence of specific weather conditions, such as particulate matter exceedances in winter (due to houses heating) or ozone exceedances in summer (due to the adverse combination of high levels of temperature and global radiation).

6 Concluding Remarks

Since 1970 air-quality measures have been started in several of the world's cities. The index design depends on both the desired objectives of communication and the research goals and, as a result, general purpose indices simply do not exist. From a methodological viewpoint, however, design strategies can be clustered into data-driven and model-driven strategies.

The data-driven approach (Bruno and Cocchi, 2002) is the most popular strategy and is based on a deterministic aggregation of the hourly measurements on each pollutant at every site in the monitoring network. Since this approach does not use probabilistic assumptions on the data generating process, there are no obvious methods either to construct these indices in the presence of missing data or to forecast their values. As can be easily seen, pollutants time series we recorded are characterized by a relevant number of missing values leading to a possible bias in the estimate of an air-quality index due to the daily and pollutant varying number of monitoring sites.

We consider a model-based approach, based on regression model, which is used in various domains as environment for statistical data analysis when one need to model the relationship between a response variable and covariates, and which helps to build an air quality index even when missing values are present. Air pollution data often show temporal dependence when measurements are made hourly or at a shorter time intervals, hence a major issue is the specification of a model for temporally correlated data.

This paper presents a non-homogeneous HMM, i.e. a doubly stochastic process with an underlying stochastic process that is not directly observable (hidden) but can be observed only through another process that produces the sequence of observations. A parsimonious model has been considered and model parameters have a physical meaning, especially when the parameter estimates aim at defining an air quality index. The aim of this paper is to advocate the use of the HMM in analyzing environmental data when the usual regression failed in modeling the relationship between a response variable and covariates and the presence of latent states is suspected. HMM may help to overcome the lack-of-fit if the sample is made of several unobserved states of statistical units.

Furthermore, we adopt a multivariate model to identify factors associated with particulate matter, nitrogen dioxide and ozone. When we face multivariate variables, and the primary focus of the analysis is not only to build a regression model, but even to describe association among variables,

the univariate approach is no longer sufficient and needs to be extended. In this context, we are likely to face complex phenomena which can be characterized by having a non-trivial correlation structure (e.g. omitted covariates may affect more than one variable), which can be captured by introducing a latent structure. Furthermore, it is well known that, when responses are correlated, the univariate approach is less efficient than the multivariate one (see e.g. Davidson and MacKinnon, 1993).

References

Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occuring in the statistical analysis of probabilistic functions in Markov chains. The Annals of Mathematical Statistics, 41:164-171.

Bilmes, J.A. (1998). A gentle tutorial of the EM algorithm and its application to paramter estimation for gaussian mixture and hidden Markov models. International Computer Science Institute TR-97-021.

Bruno, F. and Cocchi, D. (2002). A unified strategy for building simple air quality indeces. Envirometrics, 13:243–261.

Cappé, O., Moulines, E. and Rydén, T. (2005). Inference in hidden Markov models. Springer Series in Statistcs.

Hughes, J.P. (1997). Computing the observed information in the hidden Markov model using the EM algorithm. Statistics and Probability Letters 32:107-114.

Hughes, J.P., Guttorp, P. and Charles, S.P. (1999). A Non-Homogeneous Hidden Markov Model for Precipitation Occurrence Applied Statistics, 48:15–30.

Kütchenhoff, H. and Thamerus, M. (1996). Extreme value analysis of Munich air pollution data. Environmental and Ecological Statistics, 3(2), 127-141.

Leroux, B.G. and Puterman, M.L. (1992). Maximum-penalized-likelihood estimation for independent and markov-dependent mixture models, Biometrics, 48, 545-558.

MacDonald, I.L. and Zucchini, W. (1997). Hidden Markov and other models for discrete valued time series, Chapman & Hall, London.

McLachlan, G.J and Peel, D. (2000). Finite Mixture Models. Wiley. Nwe York.

Wang, P. and Puterman, M.L. (1998). Mixed Logistic Regression Models. Journal of Agricultural, Biological, and Environmental Statistics, 3:175–200.

Wang, P. and Puterman, M.L. (2001). Analysis of longitudinal data of epileptic seizure counts - a two state hidden markov regression approach. Biometrical Journal, 43:941–962.

Welch, L.R. (2003). Hidden Markov models and the Baum-Welch algorithm. IEEE Information Theory Society Newsletter, 53(4):10–15

WHO (2006). Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide. Global update 2005. WHO Press: Geneva.

		PM_{10}	NO_2	O_3	State Variable
STATE 0	Constant	-1.37	-2.72	-6.33	-0.29
		(0.61)	(0.66)	(1.25)	(0.63)
	Temperature	1.47	-1.35	-0.53	5.40
		(0.66)	(0.38)	(0.56)	(1.36)
	Pressure	0.10	0.08	-0.60	2.53
		(0.41)	(0.29)	(0.58)	(0.61)
	Humidity	0.33	-0.16	-0.43	0.64
		(0.41)	(0.78)	(0.43)	(0.67)
	Radiation	-0.42	0.26	2.85	-1.26
		(0.52)	(0.36)	(1.21)	(0.47)
STATE 1	Constant	-8.89	-5.12	-4.79	2.13
		(2.41)	(1.06)	(0.97)	(0.25)
	Temperature	-1.18	-0.09	2.65	-0.35
		(0.46)	(0.42)	(0.75)	(0.42)
	Pressure	1.56	1.22	0.45	0.76
		(0.65)	(0.82)	(0.49)	(0.27)
	Humidity	0.98	0.12	-0.18	-0.47
		(0.45)	(0.22)	(0.39)	(0.35)
	Radiation	-2.14	0.94	0.77	-0.58
		(1.20)	(0.41)	(0.57)	(0.36)

Table 1: Parameter estimates (standard errors obtained using parametric bootstrap are displayed in parentheses)

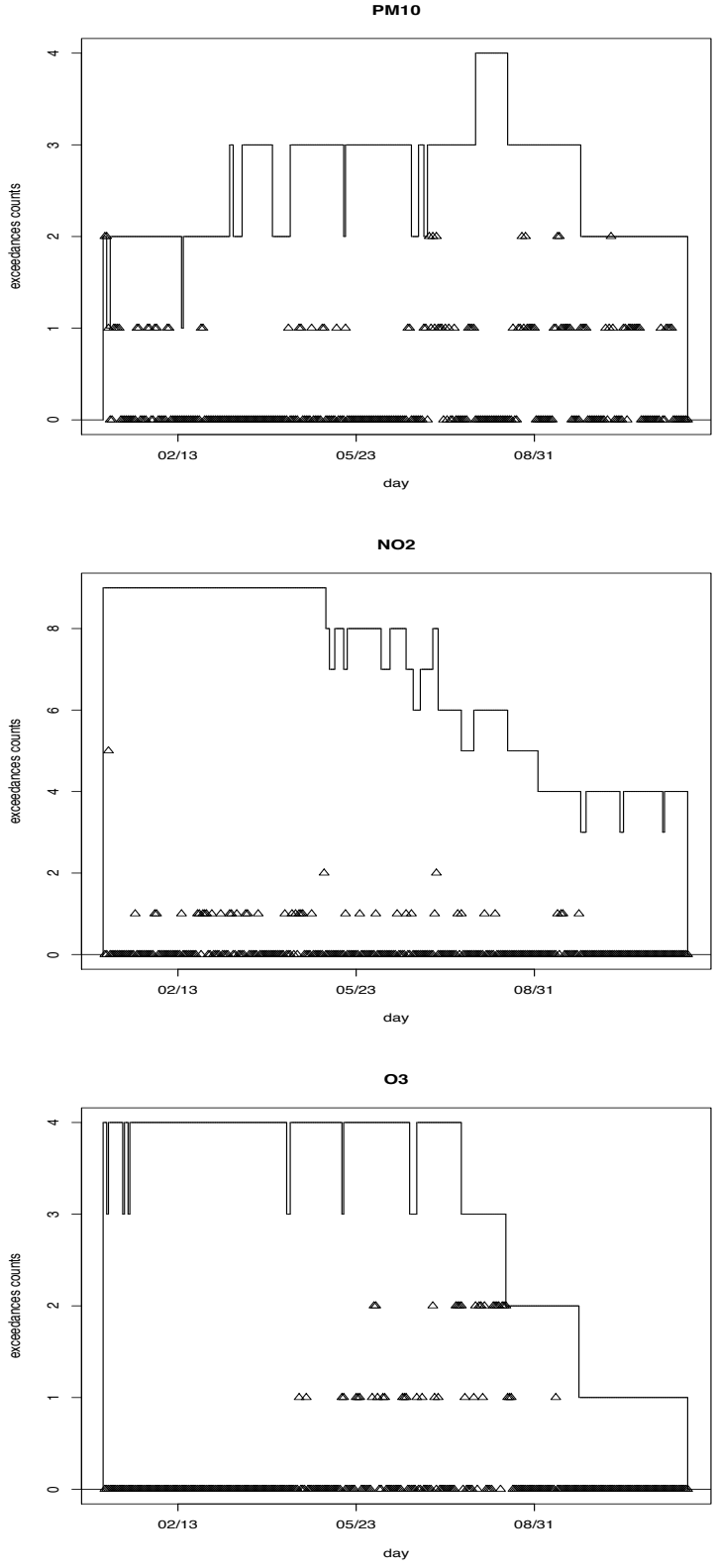


Figure 1: exceedances counts (triangles) recorded by the monitoring network of Rome in the period 1/3/2000 - 11/25/2000 relating to three pollutants - particulate matter (top), nitrogen dioxide (middle) and ozone (bottom) - and the number of stations in operation during the study period (solid line).

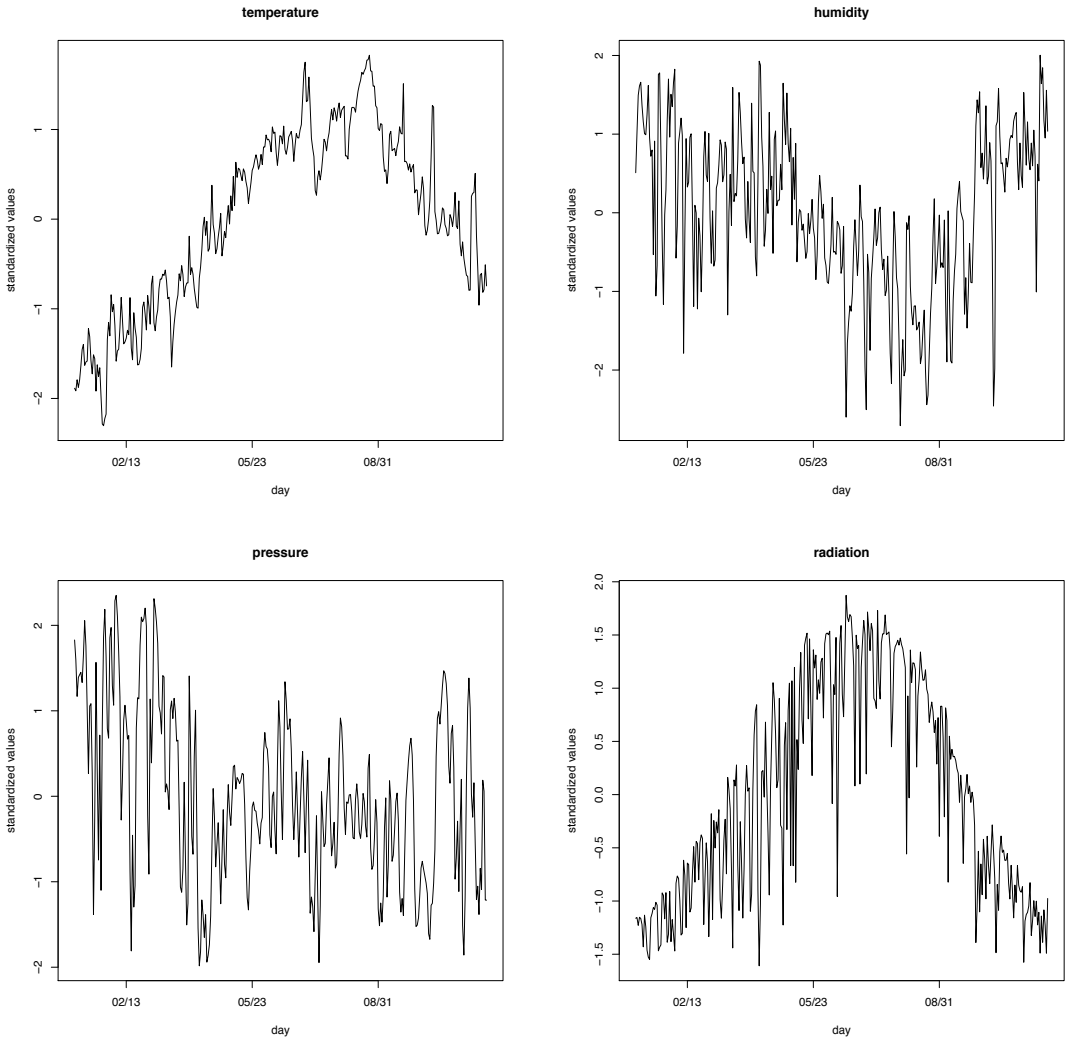


Figure 2: standardized values of daily averages of temperature (top left), humidity (top right), pressure (bottom left) and global radiation (bottom right), as recorded by the monitoring network of Rome in the period 1/3/2000 - 11/25/2000.

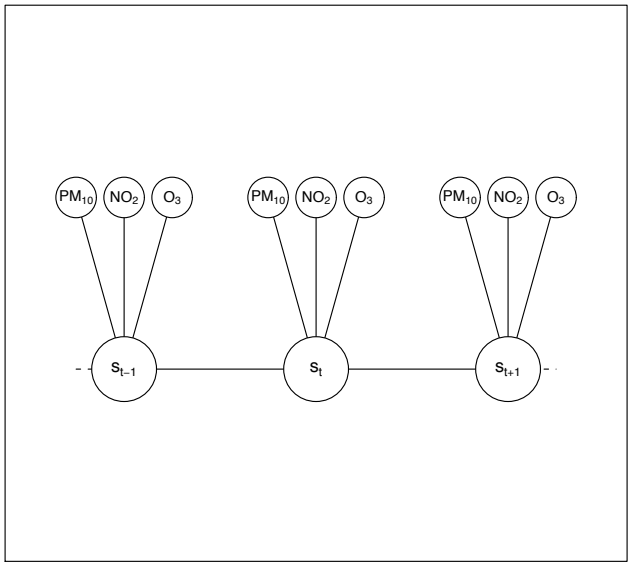


Figure 3: association graph of the hidden Markov model assumed for pollutants exceedances counts.

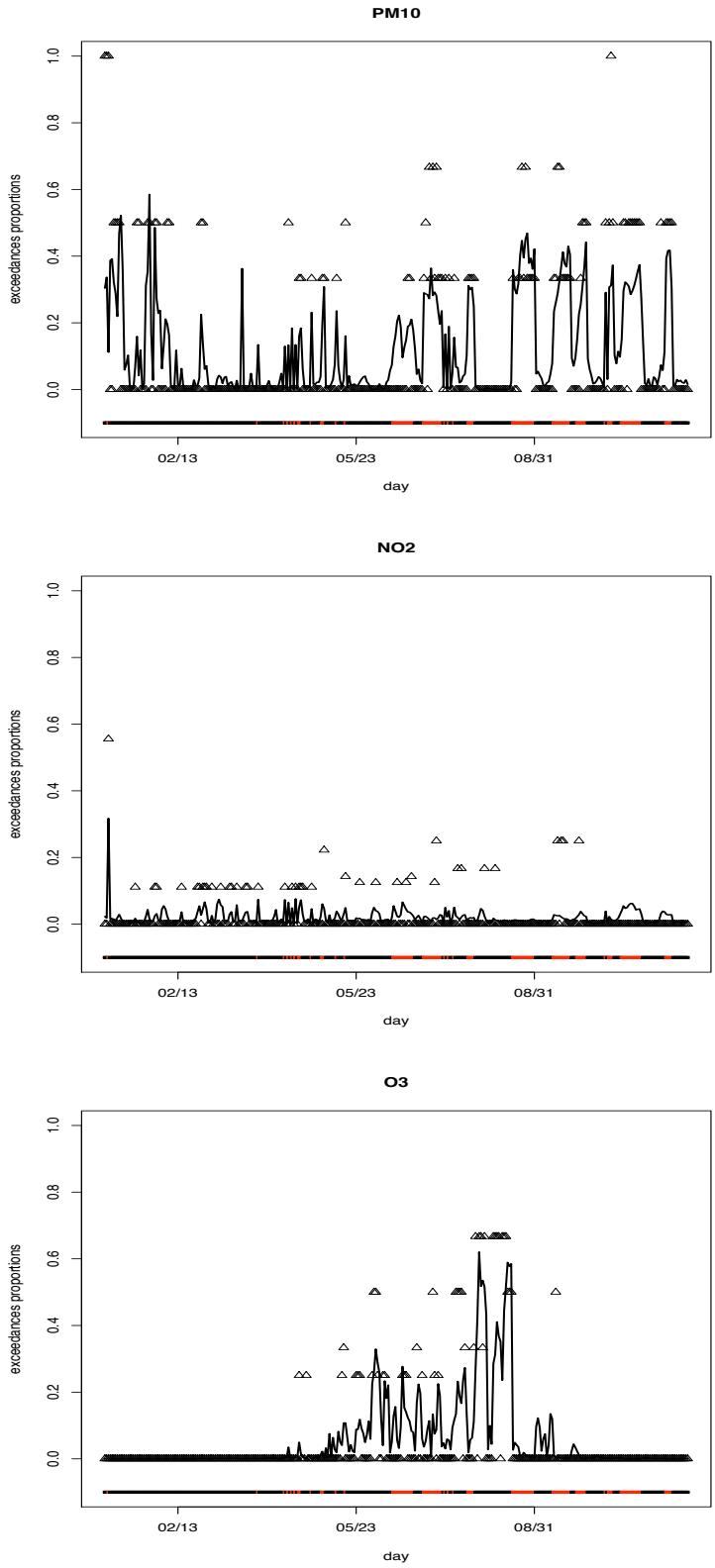


Figure 4: triangles: observed exceedances proportions (the number of station reporting an exceedance divided by the number of stations in operation, as recorded by the monitoring network of Rome in the period 1/3/2000 - 11/25/2000; solid line: the probability of an exceedance, as estimated by the hidden Markov model; bottom line: days classification as predicted by the model (red: state 0; black: state 1).

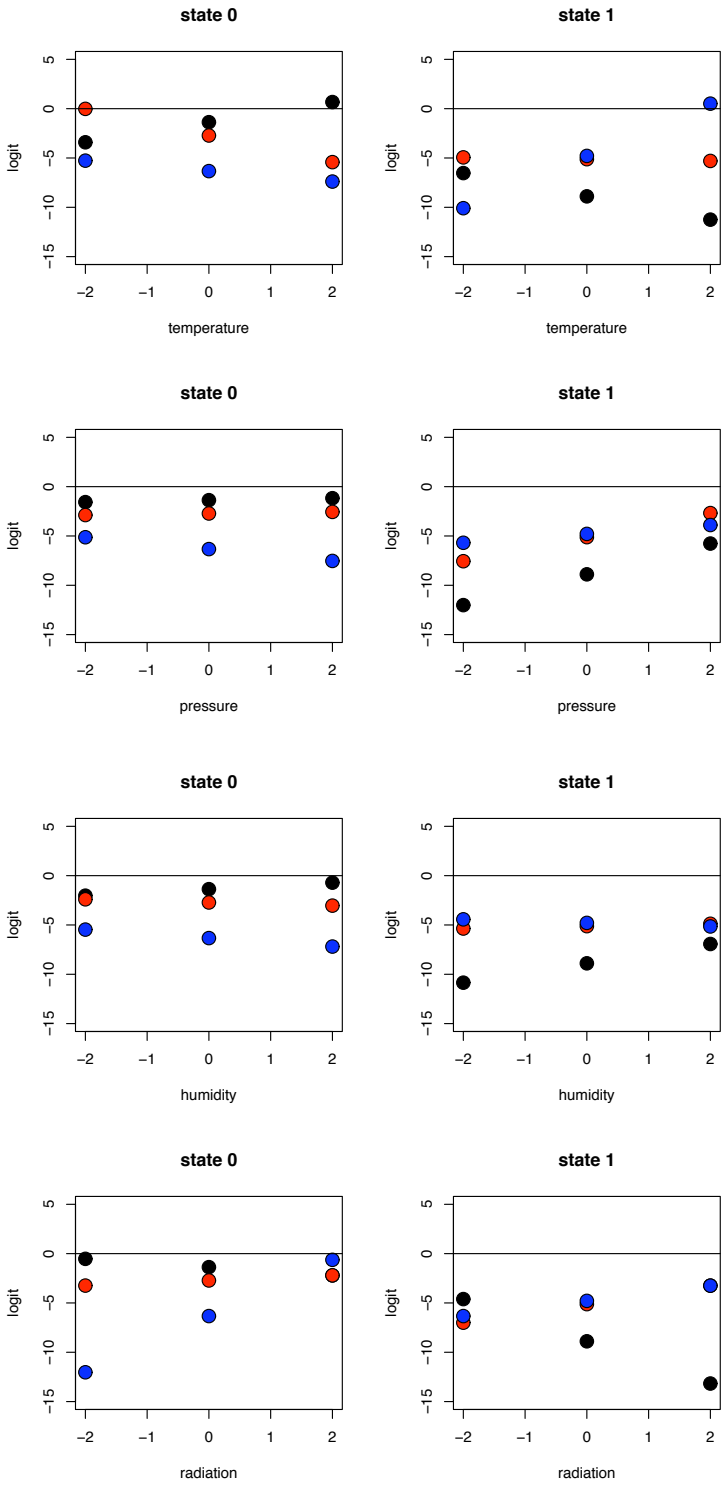


Figure 5: conditional log-odds of an exceedance probability, given the state, for particulate matter (black), nitrogen dioxide (red) and ozone (blue).

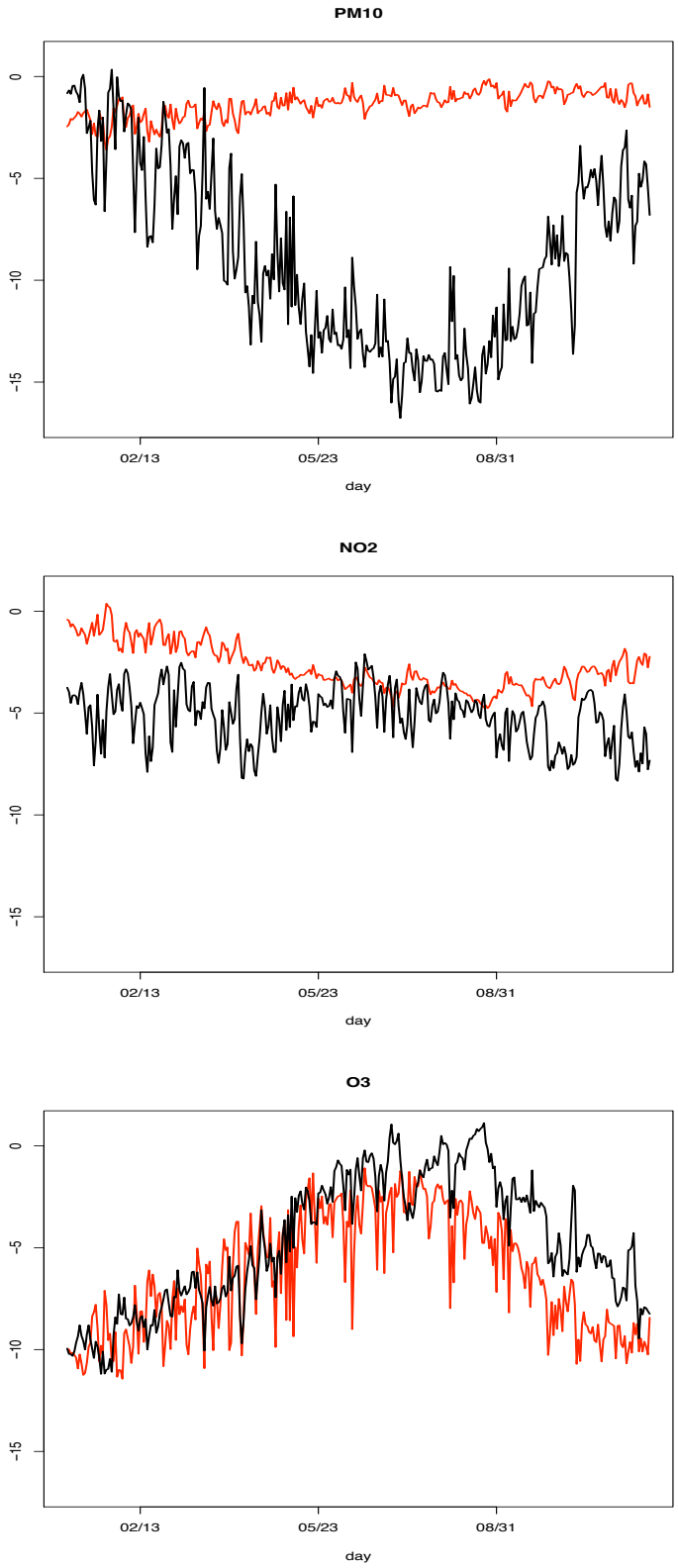


Figure 6: conditional linear predictors of particulate matter (top), nitrogen dioxide (middle) and ozone (bottom) episodes, given the latent state of the air (red: state 0; black: state 1), as estimated by a hidden Markov model.