10

A Non-Homogeneous Hidden Markov Model for the Analysis of Multi-Pollutant Exceedances Data

Francesco Lagona[1], Antonello Maruotti[2] and Marco Picone[3]
[1,2]*Dipartimento di Istituzioni Pubbliche, Economia e Società - Università di Roma Tre and GRASPA Research Unit of Roma Tre*
[3]*Dipartimento di Economia - Università di Roma Tre and GRASPA Research Unit of Roma Tre*
Italy

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 (World Health Organization, 2006). In urban areas, exceedances are usually recorded through a monitoring network, where concentrations of a number of pollutants are measured at different sites. Daily occurrences of exceedances of standards are routinely exploited by environmental agencies such as the US EPA and the EEA, to compute air quality indexes, to determine compliance with air quality regulations, to study short/long-term effects of air pollution exposure, to communicate air conditions to the general public and to address issues of environmental justice.

The statistical analysis of urban exceedances data is however complicated by a number of methodological issues. First, data can be heterogeneous because stations are often located in areas that are exposed to different sources of pollution. Second, data can be unbalanced because the pollutants of interest are often not measured by all the stations of the network and some stations are not in operation (e.g. for malfunctioning or maintenance) during part of the observation period. Third, exceedances data are typically dependent at different levels: multi-pollutants exceedances are not only often associated at the station level, but also at a temporal level, because exceedances may be persistent or transient according to the general state of the air and time-varying weather conditions may influence the temporal pattern of pollution episodes in different ways.

Non-homogeneous hidden Markov (NHHM) models provide a flexible strategy to estimate multi-pollutant exceedances probabilities, conditionally on time-varying factors that may influence the occurrence and the persistence of pollution episodes, and simultaneously accomodating for heterogeneous, unbalanced and temporally dependent data.

In this paper, we propose to model daily multi-pollutant exceedances data by a mixture of logistic regressions, whose mixing weights indicate probabilities of a number of air quality regimes (latent classes). Transition from one regime to another is governed by a non-homogeneous Markov chain, whose transition probabilities depend on time-varying meteorological covariates, through a multinomial logistic regression model. When these covariates are suitably chosen for measuring the amount of atmospheric turbolence, parameters of the multinomial logistic model indicate the influence of atmospheric stability on both the occurrence of typical pollution episodes and the persistence of these episodes. Conditionally on the latent class, exceedances are assumed independent and pollutant-specific exceedances probabilities depend on covariates that are proxies of the production of pollution. Because the information provided by these proxies is typically poor, latent classes accomodate for the influence of unobserved sources of pollution and, simultaneously, account for the dependence between multi-pollutant exceedances that were observed during the same day.

NHHM models generalize the class of homogeneous hidden Markov (HHM) models that are extensively discussed by MacDonald and Zucchini (1997). HHM models assume that the data are conditionally independent given the states of a (latent) homogeneous Markov chain and provide a flexible approach to model stationary categorical time series. An NHHM model is obtained as a generalization of a HHM model, by allowing the transition probabilities to be time-varying. On the other side, NHHM models generalize the class of mixtures of regression models with concomitant variables (Wang and Putermann, 1998), to allow for temporal dependence.

NHHM models have been already considered in the literature by several authors. Diebolt et al. (1994) have considered maximum likelihood estimation of the simple two-state Gaussian hidden Markov model with time-varying transition matrix. Applications of hidden Markov models with time-varying transitions include Durland and McCurdy (1994), Gray (1996), Peria (2002), Masson and Ruge-Murcia (2005), Kim et al. (2008), and Banachewicz et al. (2007). Wong and Li (2001) have considered a two-state non-homogeneous Markov switching mixture autoregressive model. All the above papers adopt classical inferential procedures. A Bayesian approach to inference for non-homogeneous hidden Markov model has been proposed by Filardo and Gordon (1998) and Meligkotsidou and Dellaportas (2010).

In environmental studies, NHHM models have found widespread application in meteorology and hydrology, in studies of climate variability or climate change, and in statistical downscaling of daily precipitation from observed and numerical climate model simulations (see, e.g., Zucchini and Guttorp 1991; Hughes and Guttorp 1994; Hughes et al. 1999; Charles et al. 1999; Bellone et al. 2000; Charles et al. 2004; Robertson et al. 2004; Betrò et al. 2008).

Fewer are the applications of homogeneous and non-homogeneous hidden Markov models in air quality studies, where this methodology has been mainly applied to study univariate pollutants concentrations under the assumption of normally-distributed data (Spezia, 2006; Dong et al., 2009) or to estimate exceedances probabilities (Lagona, 2005).

After describing the environmental data used in this study (Section 2), the specification of a NHHM for pollutants exceedances and the discussion of relevant computational details for estimation are outlined in Section 3. Section 4 illustrates an application to exceedances data of ozone, particulate and nitrogen dioxide, obtained from the monitoring network of Rome. Section 5 finally provides some concluding remarks.

2. Data

Our analysis is based on binary time series of occurrences and non occurrences of exceedances of air quality standards, as computed from hourly pollutants concentrations that are typically available from the monitoring network in an urban area.

station	type	PM_{10}	NO_2	O_3
Preneste	residential	34	2	
Francia	traffic		3	73
Magna Grecia	traffic		4	45
Cinecittà	residential	27	3	51
Villa Ada	residential	21	0	14
Castel Guido	rural	5	0	
Cavaliere	rural	3	0	
Fermi	traffic		22	64
Bufalotta	residential	16	0	18
Cipro	residential	7	2	31
Tiburtina	traffic		9	70
Arenula	residential		0	35

Table 1. Number of violations in 2009

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 (Italy) in 2009. These data are disseminated by the Environmental Protection Agency of the Lazio region (`www.arpalazio.net/main/aria/`). While six stations of the network are located in residential areas with moderate traffic, four stations are close to heavy traffic roads and two stations are located in rural areas.

Violations of air quality standards are defined differently for each pollutant, because most of the current legislation considers air quality standards separately for each pollutant. According to the most recent legislation, we recorded the day and the station 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 $\mu g/m^3$.

Table 1 displays the number of violations of the above standards, observed at the monitoring network in 2009. Empty cells indicate structural zeros, which are observed when a particular pollutant is not measured by the station. As expected, particulate and nitrogen dioxide exceed the standard in the neighborhood of traffic roads at a rate that is larger than that observed in residential areas, while most of the violations of ozone are observed in residential areas.

Although tables such as Table 1 are routinely reported to communicate the state of the air to the general public and to determine compliance with environmental regulations, these counts should be interpreted with caution, for a number of different reasons. First, some of the stations were not in operation during parts of the study period and hence the data are based on a time-varying number of stations. Second, the occurrence of exceedances is not only influenced by the location of the monitoring station but also by weather conditions. For example, global radiation and wind speed regulate the amount of atmospheric stability and can be responsible for stagnation or dispersion of pollutant concentrations. Atmospheric stability, i.e. the tendency of the atmosphere to resist or enhance turbulence, is related to both , global radiation and wind speed, leading to several stability classes. Stability classes are defined for different meteorological situations, characterized by wind speed and solar radiation (during the day) and can be classified according to the so-called Pasquill-Turner classification (Turner, 1994) As a result, these counts should be adjusted not only by the type of the station but also by weather conditions. This adjustment can be important when comparing exceedances data of several urban areas to address issues of environmental justice. We accordingly included daily means of wind speed and radiation into our analysis of exceedances data, as obtained by one of the most authoritative meteorological station in Rome (Collegio Romano, www.cra-cma.it/cromano.html).

3. A non-homogeneous hidden Markov model for binary data

Time series of exceedances data can be represented as a vector of $n \times H$ binary matrices, say $\boldsymbol{Y} = (\boldsymbol{Y}_t, t = 0, 1, \dots T)$, where the (i, h)th element y_{iht} of matrix $\boldsymbol{Y}_t$ is equal to 1 if the hth event occurred in unit i at time t and 0 otherwise, $i = 1 \dots n, h = 1 \dots H$.

We introduce a latent vector $\boldsymbol{s} = (s_t, t = 0, 1, \dots T)$, drawn from a vector $\boldsymbol{S} = (S_t, t = 0, 1, \dots T)$ of discrete random variables S_t that take K categorical values. The product sample space of $\boldsymbol{S}$, say $\mathbb{S}$, includes K^T vectors. Without loss of generality, we write the distribution of the observed data, say $P(\boldsymbol{Y})$, as a mixture of conditional multivariate distributions, say

$$p(\mathbf{Y}) = \sum_{\boldsymbol{s} \in \mathbb{S}} p(\mathbf{Y}, \boldsymbol{s}) = \sum_{\boldsymbol{s} \in \mathbb{S}} p(\mathbf{Y}|\boldsymbol{s}) p(\boldsymbol{s}).$$

As a result, the marginal covariance between two occurrences, say Y_{iht} and $Y_{jk\tau}$, is given by

$$\begin{aligned}\gamma(i, j, h, k, t, \tau) =& \mathbb{E} Y_{iht} Y_{jk\tau} - \mathbb{E} Y_{iht} \mathbb{E} Y_{jk\tau} \\ =& \sum_{\boldsymbol{Y}_{(i,h,t),(j,k,\tau)}} p(\boldsymbol{Y}) - \sum_{\boldsymbol{Y}_{(i,h,t)}} p(\boldsymbol{Y}) \sum_{\boldsymbol{Y}_{(j,k,\tau)}} p(\boldsymbol{Y}),\end{aligned}$$

where $\boldsymbol{Y}_{(i,h,t),(j,k,\tau)}$ indicates any matrix $\boldsymbol{Y}$ with $y_{iht} = y_{jk\tau} = 1$ and, analogously, $\boldsymbol{Y}_{(i,h,t)}$ ($\boldsymbol{Y}_{(j,k,\tau)}$) indicates any matrix $\boldsymbol{Y}$ with $y_{iht} = 1$ ($y_{jk\tau} = 1$). These covariances can be arranged in a $\boldsymbol{\Gamma}$ blocks-matrix $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_{t,\tau}; t, \tau = 0, 1 \dots T)$, whose diagonal blocks, $\boldsymbol{\Gamma}_{tt}$, describe the covariance structure between contemporary occurrences, while the off-diagonal blocks, $\boldsymbol{\Gamma}_{t\tau}$, describe the autocovariances and the cross-autocovariances of the multivariate time series.

The above mixture is called HHM model when

1. exceedances patterns are conditionally independent given the latent states (conditional independence assumption), namely

$$p(\mathbf{Y}|\boldsymbol{s}) = \prod_{t=0}^{T} p(\boldsymbol{Y}_t|\boldsymbol{s}) = \prod_{t=0}^{T} p(\boldsymbol{Y}_t|s_t) \tag{1}$$

2. and the latent vector $\boldsymbol{s}$ is sampled from a Markov chain, namely

$$p(\boldsymbol{s}) = \delta_s \prod_{t=1}^{T} p(s_{t-1}, s_t),$$

where $\delta_s = p(S_0 = s)$ and the transition probabilities $p(s_{t-1}, s_t) = P(S_t = s_t | S_{t-1} = s_{t-1})$ do not vary with time (homogeneity assumption).

In a HHM model, multivariate time series data are therefore modeled by a mixture of multivariate distributions, whose parameters depend on the stochastic evolution of a unobserved Markov chain. As a result, the hidden states of the chain can be interpreted as different regimes at which multivariate exceedances occur.

From a technical viewpoint, an HHM model greatly reduces the number of unknown parameters that drive the distribution of a multinomial time series. However, the K conditional distributions $p(\boldsymbol{Y}_t|S_t = s)$ still depend on $(2^{I \times J} - 1) \times K$ probabilities. Although a saturated log-linear re-parametrization of these probabilities is in principle possible, it would involve a model with high-order interactions that may be difficult to interpret. Moreover, estimation of saturated models can be unstable if data are unbalanced, as often happens with urban exceedances data. We therefore need to employ strategies to reduce the number of parameters. A parsimonious model that accounts for the multi-pollutant nature of the data and simultaneously allows for heterogeneous monitoring networks is a binary regression model, where pollutant-specific exceedances probabilities vary with the monitoring station. More precisely, we assume that

$$p(\boldsymbol{Y}_t|S_t = s) = \prod_{i=1}^{I}\prod_{j=1}^{J} \theta_{ijs}^{y_{ijt}} \left(1 - \theta_{ijs}\right)^{1-y_{ijt}}, \tag{2}$$

where θ_{ijs} is the conditional probability that pollutant j exceeds the standard at station i, under regime s. Probabilities θ_{ijs} can be re-parametrized in a number of different ways, depending on the purpose of the analysis and the availability of specific covariates on single stations. In our application, the following two-way logit model was exploited

$$\operatorname{logit} \theta_{ijs} = \beta_{0s} + \beta_{is} + \beta'_{js}, \tag{3}$$

where β_{0s} is a baseline parameter, while β_{is} and β'_{js} are respectively the station and the pollutant effects under regime s, with the identifiability constraints $\beta_{1s} = \beta'_{1s} = 0$, for each $s = 1 \ldots K$.

Parameters in equation (3) model exceedances data within multinomial regimes. In a HHM model, the temporal persistence of each regime during the period of interest is governed by the homogeneous (i.e., time-constant) transition probabilities of a latent Markov chain. 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 may have a significant influence on the persistence of exceedances, leading to a non-stationary behavior of exceedances data. Motivated by this, we extend the HHM model framework to allow for non-homogeneous transition probabilities that depend on a profile $\boldsymbol{x}_t$ of meteorological covariates. Specifically, we assume that latent vector $\boldsymbol{s}$ is drawn from a non-homogeneous Markov chain with distribution

$$p(\boldsymbol{s}) = \delta_s \prod_{t=1}^{T} p(s_{t-1}, s_t), \tag{4}$$

and exploit a multinomial logit model to re-parametrize the time-varying transition probabilities, as follows

$$p_t(s,k) = \frac{\exp\left(\gamma_{0ks} + \boldsymbol{x}_t^T \boldsymbol{\gamma}_{ks}\right)}{\sum\limits_{h=1}^{K} \exp\left(\gamma_{0hs} + \boldsymbol{x}_t^T \boldsymbol{\gamma}_{hs}\right)}, \tag{5}$$

where γ_{0ks} is a baseline regime-specific effect ($\gamma_{0ss} = 0$, for identifiability) and the vector γ_{ks} are regression coefficients that measure the effect of weather conditions on transition probabilities.

Combining (2) and (4), we propose to model a time series of multivariate exceedances data by the following marginal distribution:

$$p(\mathbf{Y}|\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\delta}) = \sum_{s_0} \delta(s) \sum_{s_1 \ldots s_T} \prod_{t=0}^{T} p_t(s_{t-1}, s_t) \prod_{i=1}^{I} \prod_{j=1}^{J} \theta_{ijs}^{y_{ijt}} \left(1 - \theta_{ijs}\right)^{1-y_{ijt}}, \tag{6}$$

known up to the parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}$. The above distribution modularizes the dependency structure of exceedance data, by separating temporal dependence, multivariate dependence, and non-stationary behavior. More precisely, the marginal covariance matrix of the multinomial time series $\boldsymbol{Y}$ can be viewed as a blocks-matrix $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_{t,\tau}; t, \tau = 0, 1 \ldots T)$, whose diagonal blocks, $\boldsymbol{\Sigma}_{tt}$, describe the association between contemporary exceedances, while the off-diagonal blocks, $\boldsymbol{\Sigma}_{t\tau}$, describe the autocovariances and the cross-autocovariances of the multivariate time series. In particular, the generic element of $\boldsymbol{\Sigma}_{tt}$ is the (marginal) covariance between the exceedances of two pollutants j and l, recorded at two stations i and m at the same time t, namely

$$\begin{aligned} \sigma_{ijlm}(t) =& p(y_{ijt} = 1, y_{lmt} = 1) \\ =& \sum_{k=1}^{K} \pi_k(t)\theta_{ijk}\theta_{lmk} - \left(\sum_{k=1}^{K} \pi_k(t)\theta_{ijk}\right)\left(\sum_{k=1}^{K} \pi_k(t)\theta_{lmk}\right), \end{aligned} \tag{7}$$

where

$$\pi_k(t) = p(S_t = k) = \sum_{\boldsymbol{s}_{0:t-1}} \delta_{s_0} \prod_{\tau=1}^{t-1} p_\tau(s_{\tau-1}, s_\tau) p_t(s_{t-1,k})$$

is the (time-varying) marginal probability for the latent chain of being in state k at time t. In general, for two different times t and τ, the generic element of matrix $\boldsymbol{\Sigma}_{t\tau}$ is given by

$$\begin{aligned} \sigma_{ijlm}(t, \tau) =& p(y_{ijt} = 1, y_{lm\tau} = 1) \\ =& \sum_{k,h}^{1\ldots K} \pi_{kh}(t, \tau)\theta_{ijk}\theta_{lmh} - \left(\sum_{k=1}^{K} \pi_k(t)\theta_{ijk}\right)\left(\sum_{h=1}^{K} \pi_h(\tau)\theta_{lmh}\right), \end{aligned} \tag{8}$$

where

$$\pi_{k,h}(t, \tau) = p(S_t = k, S_\tau = h) = \sum_{\boldsymbol{s}_{t:\tau-1}} \pi_k(t) \prod_{\tau'=1}^{\tau-1} p_{\tau'}(s_{\tau'-1}, s'_\tau) p_\tau(s_{\tau-1}, h)$$

is the joint probability for regimes k and h to act at times t and τ, respectively.

Examination of the above covariances clearly illustrates that model (6) includes, as particular cases, a number of simpler models that could be used for examining multivariate pollutants exceedances. For example, when the transition probability matrix takes a diagonal form, model (6) reduces to a simple mixture of K generalized linear models with concomitant variables (Wang and Putermann, 1998), which could be used when the data do not show a significant temporal dependency structure. When, additionally, $K = 1$, model (6) degenerates to a logistic regression model for multivariate pollutants exceedances (Kutchenhoff and Thamerus, 1996), which can be exploited under a strong homogeneity of the data.

We take a maximum likelihood approach to estimate the parameters of the proposed NHHM model. To account for the presence of missing values, our analysis is based on the maximization of the log-likelihood function that is obtained by marginalizing (6) with respect to the missing values, namely

$$l(\gamma, \boldsymbol{\beta}, \boldsymbol{\delta} | \boldsymbol{Y}_{\text{obs}}) = \sum_{\boldsymbol{Y}_{\text{mis}}} \log p(\boldsymbol{Y} | \gamma, \boldsymbol{\beta}, \boldsymbol{\delta}), \tag{9}$$

where $\boldsymbol{Y}_{\text{mis}}$ and $\boldsymbol{Y}_{\text{obs}}$ denote the arrays of the missing and observed values, respectively. We recall the conditional independence that in our NHHM model holds between exceedances within the same latent state. As a result, the contribution of each missing value to $l(\gamma, \boldsymbol{\beta}, \boldsymbol{\delta} | \boldsymbol{Y}_{\text{obs}})$ is equal to 1. By introducing a missing indicator variable ($r_{ijt} = 1$ if y_{ijt} is missing and 0 otherwise), the log-likelihood function that we maximize is thus finally given by

$$l(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\delta} | \boldsymbol{Y}_{\text{obs}}) = \log \sum_{s_0} \delta(s) \sum_{s_1 \ldots s_T} \prod_{t=0}^{T} p_t(s_{t-1}, s_t) \prod_{i=1}^{I} \prod_{j=1}^{J} \left(\theta_{ijs}^{y_{ijt}} \left(1 - \theta_{ijs}\right)^{1-y_{ijt}} \right)^{r_{ijt}}. \tag{10}$$

In the hidden-Markov-models literature, maximization of (10) is essentially based on the EM algorithm (see e.g. MacDonald and Zucchini, 1997; Cappé et al., 2005; and reference therein for details about the algorithm). As it stands, expression (10) is of little or no computational use, because it has K^{T+1} 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).

We point out that the estimation algorithm involves the iterative evaluation of the solutions of the weighted score equations:

$$\sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{s=1}^{K} Pr(S_{t+1} = k, S_t = j \mid \mathbf{Y}_{\text{obs}}, \hat{\boldsymbol{\theta}}) \frac{\partial \log p_t(s,k)}{\partial \gamma} = 0; \tag{11}$$

and

$$\sum_{t=1}^{T} \sum_{k=1}^{K} Pr(S_t = s \mid \mathbf{Y}_{\text{obs}}, \hat{\boldsymbol{\theta}}) \frac{\partial \log f(\mathbf{Y}_t \mid S_t = s)}{\partial \boldsymbol{\beta}}, \tag{12}$$

where $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\gamma})$ are the estimates found at a previous iteration. The above equations are the score equations of generalized linear models (GLM) with weights $Pr(S_{t+1} = k, S_t = j \mid \mathbf{y})$ and $Pr(S_t = s \mid \mathbf{y})$ respectively. Parameter estimates can therefore immediately be estimated by exploiting any GLM software, by simply including a component factor in a generalized linear model as suggested by Hinde and Wood (1987), which is conveniently (even if inefficiently) handled by augmenting the data matrix.

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 readily provide standard errors for parameters estimates.

We computed standard errors of parameter estimates using parametric bootstrap (Efron and Tibshirani, 1993), 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

$$\hat{se}_R = \left\{ \frac{1}{R-1} \sum_{r=1}^{R} [\hat{\kappa}(r) - \overline{\kappa}(R)]^2 \right\}^{1/2}, \tag{13}$$

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

In a general framework, there are at least three different methods for computing standard errors (and confidence intervals) of hidden Markov model parameters, namely likelihood profiling, bootstrapping and a method based on a finite difference approximation to the Hessian (Visser et al., 2000). In this paper we adopt the parametric bootstrap approach generating bootstrap samples according to the parametric model using the maximum likelihood estimates of the parameters. Our choice is due to both the simplicity of implementing the parametric bootstrap and the results produced by this procedure. As shown by Visser et al. (2000), in the context of long time series (i.e. $T > 100$) computing the exact Hessian is not feasible and, via a simulation study, it can be proved that likelihood profiling and bootstrapping produce similar results, whereas the standard errors from the finite-differences approximation of the Hessian are mostly too small.

However, in the general hidden Markov model framework assessing the uncertainty about parameters can be difficult, as bootstrapping typically relabels states in different ways: the role of states can be exchanged at each simulation. Problems due to label switching will be most acute when data are not informative about the transition matrices.

There are several possible solutions to this label switching problem, motivated by the literature on mixture distributions (see e.g. Richardson and Green, 1997; Celeux, 1998; Boys et al., 2000; Spezia, 2009). The label switching problem can be tackled by placing an identifiability constraint on some parameter. This operation can be risky if no information on the ordering constraint is available to the investigator; so, the parameter space can be truncated wrongly and, consequently, estimators are biased. Hence, the identifiability constraint can be placed when the regimes are well separated, only.

Bayesian information criterion (BIC) is used to compare the models. Selecting an NHHM model that minimizes the BIC provides a relatively parsimonious model that fits the data well. The final decision on how many and which of the resulting summary variables are to be included in the model is evaluated in terms of physical realism, distinctness of the weather state patterns and model interpretation.

4. Results

We have estimated a number of different NHHM models from the exceedances data described in Section 2, by varying the number K of states of the latent chain. This section presents the results obtained by a three-states model, which was chosen on the basis of the BIC statistic, the degree of separation of latent classes and the physical meaning of the parameters.

The posterior probabilities of the three states (Figure 1) show that the three latent classes are well separated and that days can be clustered according to their maximum posterior probabilities of class membership. The resulting classification is intuitively appealing (Figure 2): under state 1, pollution episodes are mainly characterized by ozone exceedances, while state 3 is dominated by exceedances of particulate matter and a few violations of the nitrogen dioxide standard; finally, state 2 clusters days with acceptable air conditions. We however remark that days are clustered by jointly modeling the exceedances probabilities of the three pollutants and simultaneously accounting for the type of monitoring station where violations of standards are observed. Table 2 shows the estimated effects on the conditional exceedances probabilities, for each state. Effects are displayed by taking the log-odds of a particulate exceedance in a moderate traffic station as baseline. Under state 1 the log-odds of an exceedance of ozone are greater than the log-odds of an exceedance of the other two pollutants. The situation is reversed under state 3, where particulate and nitrogen dioxide dominate the probability of a pollution episode. As expected, the exposure to pollution sources is strongly significant only when a pollution episode occur (i.e., under state 1 or 3). When, conversely, the quality of the air is acceptable, most of the stations are likely to report concentrations that are below the standard, regardless of the locations where measurements are made. However, when a pollution episode occurs, the expected number of violations depends on the distribution of the type of monitoring sites that are functioning. As a result, when a few violations occur, the model predicts a serious pollution episode only when exceedances are observed in locations that are exposed to low pollution sources. This explains why days with a similar number of exceedances are given a different class membership by the model (Figure 2).

The estimated transition probabilities of the latent chain, varying with weather conditions, are depicted in Figure 3, according to the origin state (columns) and the destination state (rows). Examining the pictures in the second row of the figure, we observe that a regime of acceptable air quality (state 2) is persistent during the whole year, as indicated by the large probabilities of remaining in state 2 (middle picture). As a result, the probabilities of moving from state 2 to a different state are generally low. As expected, while the probability of moving from state 2 to state 1 (ozone episodes) increases during the warm seasons, moderate probabilities of moving to state 3 (particulate and nitrogen dioxide episodes) increase during the cold seasons. The high variability of the probabilities of remaining in state 1 (first row, left) and in state 3 (third row, right) confirm that pollution episodes, as measured by the number of exceedances, were not persistent during the period of interest.

Figure 3 has been computed by exploiting the estimates of Table 3, which display the log-odds of conditional transition probabilities of the latent chain, by taking the probability of remaining in the same state as a reference. Examination of the second column of this table shows that the probability of moving from a state of good air quality to a pollution episode decreases at high wind speed, in keeping with the known role that the wind plays in the dispersion of pollutants. On the contrary, solar radiation has a positive effect on the probability to move to ozone episodes, occurring in summer, and a negative effect on the probability to move to episodes of particular matter and nitrogen dioxide, which occur during the cold seasons. The estimates of the first column of the table confirm that, when wind speed increases, the probability to move from state 1 to a state of acceptable air quality is much greater than that of moving to a ozone episode. Interestingly, global radiation has a negative effect on a transition from state 1. Particularly in winter, when state 1 is often reached, high

estimate	state 1	state 2	state 3
intercept	-3.4993	-4.0985	0.4709
	(0.3381)	(0.1592)	(0.0900)
low emission	-2.7563	-0.2737	-14.0231
	(0.4545)	(0.7789)	(3.4449)
ozone	3.9124	-2.2011	-18.2600
	(0.3724)	(0.6044)	(3.4433)
nitrogen dioxide	-1.6443	-1.7206	-3.3205
	(0.7845)	(0.3790)	(0.2037)

Table 2. Log-odds of exceedances probabilities (standard errors in brackets)

	destination		origin	
		state 1	state 2	state 3
	state 1	0	-11.3842	-19.1536
			(1.0768)	(2.3051)
intercepts	state 2	5.8137	0	-1.8993
		(1.6782)		(0.3318)
	state 3	9.5675	0.8748	0
		(2.3389)	(0.3204)	
	state 1	0	-0.5000	-0.7198
			(0.0384)	(0.0502)
wind speed	state 2	-1.1085	0	2.2761
		(0.6023)		(0.0925)
	state 3	-4.3735	-1.2730	0
		(1.1265)	0.0916	
	state 1	0	0.3808	0.6482
			(0.0348)	(0.0092)
global radiation	state 2	-0.1891	0	-0.1539
		(0.0195)		(0.0976)
	state 3	-0.4796	-0.1132	0
		(0.1211)	(0.0201)	

Table 3. Log-odds of transition probabilities (standard errors in brackets)

levels of solar radiation create in the atmosphere the phenomenon of thermal inversion (a layer of warm air settles over a layer of cold air) preventing the mixing up among the different layers of the air, due to the convective currents, and, as a result, preventing pollutants from rising and scattering. Finally, the estimates in column three of the table indicate the influence of weather conditions on the probability to move from state 3, which occurs in summer. As expected, while wind speed is associated with an increasing probability to return to a regime of clean air, an increase in the levels of global radiation negatively influences the chances for the system to return to a state of acceptable air conditions.

5. Discussion

Although most of the current legislation considers air quality standards separately for each pollutant, recent studies stress the importance of a joint examination of exceedances with respect to several air pollutants. When we face multivariate variables, and the primary focus

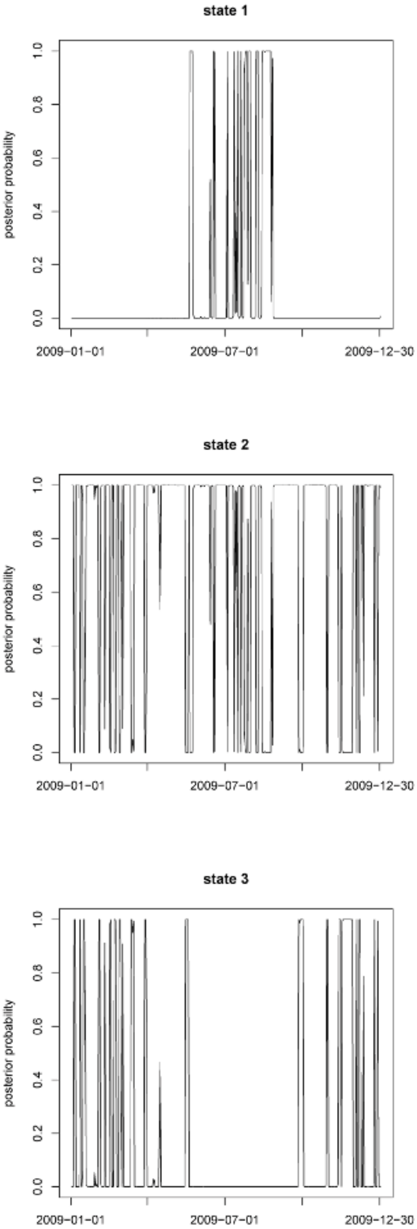


Fig. 1. posterior state probabilities, as estimated by a three-state non-homogeneous hidden Markov model.

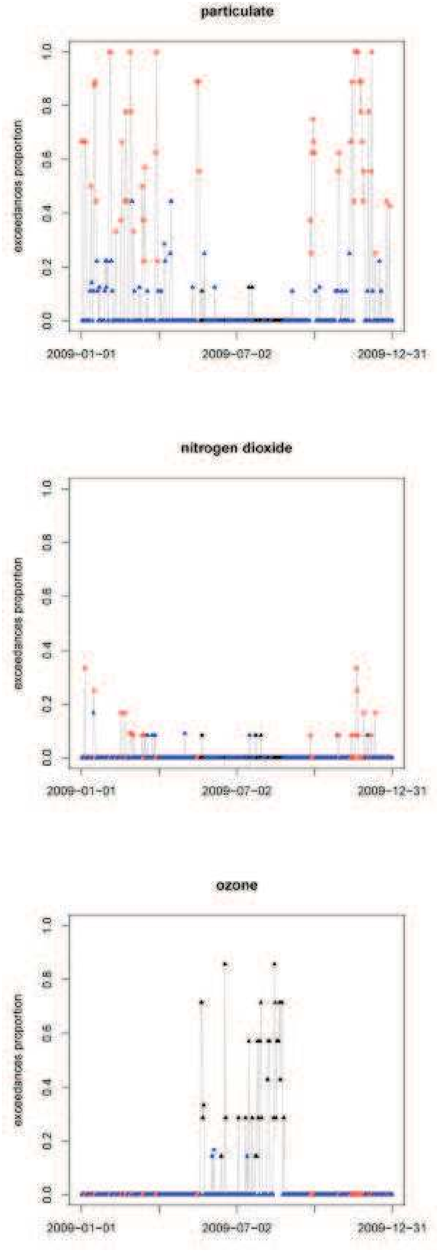


Fig. 2. observed exceedances proportions of three pollutants, clustered according to their posterior state probabilities, as estimated by a three-state non-homogeneous hidden Markov model; state 1 (black), state 2 (blue), state 3 (red).

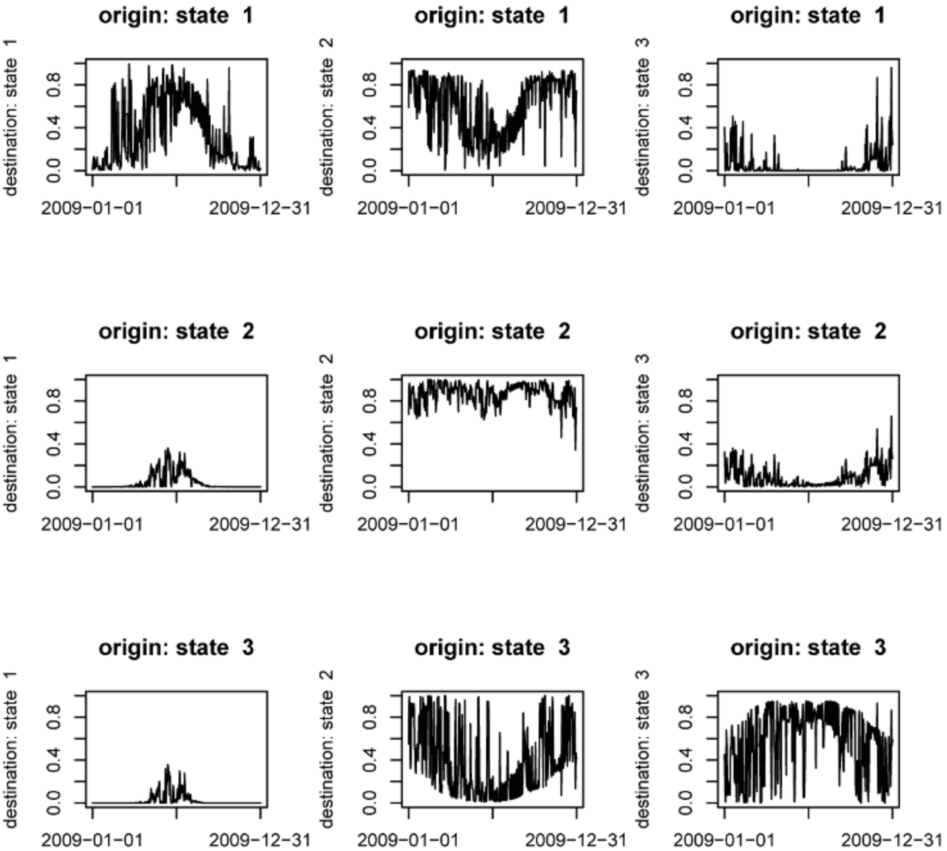


Fig. 3. Probabilities of transition from one state to another state, as estimated by a three-state non homogeneous hidden Markov model.

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.

To estimate multivariate exceedances probabilities, we have fitted a NHHM model to a time series of multivariate exceedances data. Non-homogeneous hidden Markov models are parsimonious specification of non-stationary time series and can be generalized along a number of dimensions, to accommodate continuous or discrete multivariate data and modularize the data dependence structure of the data according to the purpose of an analysis. In our case study, a NHHM model provides a parsimonious representation of a time series of multivariate exceedances by means of three latent regimes that are temporally persistent or transient, according to time-varying weather conditions. Estimates of the effects of factors that may influence both the occurrence and the persistence of specific exceedances are in terms of log-odds, which helps to communicate results to nonspecialists. The clear-cut separation of the three latent classes supports a model-based clustering of days into periods of severe pollution episodes and periods of reasonable quality of the air. Estimated transition probabilities allow to interpret the persistence of pollution episodes in terms of the general conditions of the weather in the area of interest.

The NHHM model presented in this paper provides a model-based clustering of days, according to different patterns of multi-pollutants exceedances probabilities. Estimated posterior probabilities of the two latent regimes can be then interpreted as an air quality index, which exploits maximum likelihood estimates to provide a daily summary of multivariate exceedances data. Model-based air quality indexes are certainly more difficult to explain to the general public than data-driven indexes that are based on a deterministic aggregation of the hourly measurements on each pollutant at every site in a monitoring network. However a data-driven approach (Bruno and Cocchi, 2002) does not use probabilistic assumptions on the data generating process, and, as a result, there are no obvious methods either to construct these indexes in the presence of missing data or to predict their values.

6. References

Banachewicz, K., Lucas, A. and Vaart, A. (2007). Modeling Portfolio Defaults Using Hidden Markov Models with Covariates. Econometrics Journal, 10, 1-18.

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

Bellone, E., Hughes, J.P., Guttorp, P. (2000). A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. Climatology Research, 15, 1-12.

Betrò, B., Bodini, A. and Cossu, Q.A. (2008). Using hidden Markov model to analyse extreme rainfall events in Central-East Sardinia. Envirionmetrics 19: 702-713.

Boys, R.J., Henderson, D.A. and Wilkinson, D.J.(2000). Detecting homogeneous segments in DNA sequence by using hidden Markov models. Journal of the Royal Statistical Society - Series C 49:269-285.

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 Statistics.

Celeux, G. (1998) Bayesian inference for mixture: the label switching problem. In COMPSTAT '98. Proc. Comiplutationial Statistics Conf. (eds R. W. Payne and P. J. Green), pp. 227–232. Heidelberg: Physica.

Charles, S.P., Bates, B.C., Hughes, J.P. (1999) A spatiotemporal model for downscaling precipitation occurrence and amounts. Journal of Geophysical Research - Atmospheres, 104, 31657-31669.

Charles, S.P., Bates, B.C., Smith, I.N., Hughes, J.P. (2004) Statistical downscaling of daily precipitation from observed and modelled atmospheric fields. Hydrological Processes, 18, 1373-1394.

Diebolt, F.X., Lee, J.H. and Weinbach, G.C.(1994). Regime switching with time varying transition probabilities. In C.P. Hargreaves, editor, Nonstationary Time Series Analysis and Cointegration, pp. 283–302.

Dong, M., Dong, Y., Kuang, Y., He, D., Erdal, S. and Kenski, D. (2009) PM2.5 concentration prediction using hidden semi-Markov model-based times series data mining. Expert Systems with Applications, 36: 9046-9055.

Durland, J. M. and McCurdy (1994). Duration-dependent transitions in a Markov model of U.S. GNP growth. Journal of Business and Economic Statistics, 12, 279-288.

Efron, B. and Tibshirani, R.J. (1993). An introduction to bootstrap. Chapman & Hall: New York.

Filardo, A. J. and Gordon, S. F. (1998). Business cycle durations. Journal of Econometrics, 85, 99-123.

Gray, S. F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. Journal of Financial Economics, 42, 27-62

Hinde, J.P. and Wood, A.T.A. (1987). Binomial variance component models with a non-parametric assumption concerning random effects. In Longitudinal Data Analysis, R. Crouchley (ed.). Averbury, Aldershot, Hants.

Hughes, J.P. and Guttorp, P. (1994). A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. Water Resources Research, 30:1535-1546.

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

Kim, C., Piger, J. and Startz, R. (2008). Estimation of Markov regime-switching regression models with endogenous switching. Journal of Econometrics, 143, 263-273.

Lagona, F. (2005). Air Quality Indices via Non Homogeneous Hidden Markov Models. Proceeding of the Italian Statistical Society Conference on Statistics and Environment, Contributed Papers, CLEUP, Padova, 91-94

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. New York.

Masson, P. and Ruge-Murcia, F. J. (2005). Explaining the transition between exchange rate regimes. Scandinavian Journal of Economics, 107, 261-278.

Meligkotsidou, L. and Dellaportas, P. (2010). Forecasting with non-homogeneous hidden Markov models. Statistics and Computing, in press

Peria, M. S. M. (2002). A regime-switching approach to the study of speculative attacks: A focus on the EMS cricis. Empirical Economics, 27, 299-334.

Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). Journal of the Royal Statistical Society - Series B 59: 731–792.

Robertson A.W., Kirshner S. and Smyth P. (2004) Downscaling of daily rainfall occurrence over Northeast Brazil using a hidden Markov model. Journal of Climate, 17(22):4407-4424.

Spezia L. (2006). Bayesian analysis of non-homogeneous hidden Markov models. Journal of Statistical Computation and Simulation, 76: 713-725.

Spezia, L. (2009). Reversible jump and the label switching problem in hidden Markov models. Journal of Statistical Planning and Inference, 139: 2305-2315

Turner, D.B. (1994). Workbook of Atmospheric Dispersion Estimates. Lewis Publishers.

Visser, I., Raijmakers, M.E.J. and Molenaar, P.C.M. (2000). Confidence intervals for hidden MArkov model parameters. British Journal of Mathematical and Statistical Psychology 53: 317-327.

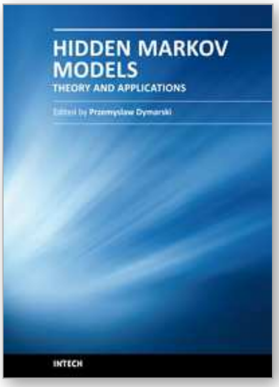
Wang, P. and Puterman, M.L. (1998). Mixed Logistic Regression Models. Journal of Agricultural, Biological, and Environmental Statistics, 3:175-200.

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

Wong, C.S. and Li, W.K. (2001). On a logistic mixture autoregressive model. Biometrika 88(3): 833-846.

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

Zucchini, W. and Guttorp, P. (1991). A hidden Markov model for space-time precipitaion. Water Resources Research, 27, 1917–1923.



Hidden Markov Models, Theory and Applications
Edited by Dr. Przemyslaw Dymarski

ISBN 978-953-307-208-1
Hard cover, 314 pages
Publisher InTech
Published online 19, April, 2011
Published in print edition April, 2011

Hidden Markov Models (HMMs), although known for decades, have made a big career nowadays and are still in state of development. This book presents theoretical issues and a variety of HMMs applications in speech recognition and synthesis, medicine, neurosciences, computational biology, bioinformatics, seismology, environment protection and engineering. I hope that the reader will find this book useful and helpful for their own research.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821