International Journal of Pure and Applied Mathematics
Volume 48 No. 3 2008, 315-331

A TRANS-DIMENSIONAL MCMC ALGORITHM TO ESTIMATE THE ORDER OF A MARKOV CHAIN: AN APPLICATION TO OZONE PEAKS IN MEXICO CITY

Luis J. Álvarez[1], Eliane R. Rodrigues[2 §]
[1]Instituto de Matemáticas – UNAM - Unidad Cuernavaca
Av. Universidad, S/N – Lomas de Chamilpa
Morelos, Cuernavaca, 62210, MEXICO
e-mail: lja@matcuer.unam.mx
[2]Instituto de Matemáticas – UNAM
Area de la Investigación Científica
Circuito Exterior, Ciudad Universitaria
México, D.F., 04510, MEXICO
e-mail: eliane@math.unam.mx

Abstract: In this paper the problem of estimating the order $K \geq 0$ of a Markov chain is addressed. In order to do so, we assume that ozone peaks follow a time-homogeneous Markov chain of order K. This order is estimated using a trans-dimensional Markov chain Monte Carlo (MCMC) algorithm. Once K is estimated it is possible to obtain estimates for the transition matrix of the chain. Results are applied to ozone data provided by the Mexico City monitoring network. Prediction about the probability of having an ozone peak in a given time into the future given some present and/or past conditions may be obtained using the estimated transition matrix of the chain.

AMS Subject Classification: 60J20, 62F15, 62M99, 92F05
Key Words: trans-dimensional MCMC, Bayesian inference, inference in stochastic processes, ozone peaks

Received: May 26, 2008

§Correspondence author

1. Introduction

Nowadays, a common problem in large cities is the constant violation of air quality standards due to high concentration of pollution. Those high levels of pollutants can be very hazardous to human health (see [7], [27], [31], and references therein). In particular, it is well known that for ozone levels above 0.11 parts per million (0.11ppm) a very sensitive population (newborn and elderly) experience a deterioration in their health. Therefore, being able to estimate the probability of occurrence of such events when the threshold is 0.11ppm or higher is considered is of great importance for policy makers. If there is a high probability of a violation of a given standard, then environmental authorities could implement preventive measures. In this way emergency situations could be avoided and/or the population in general could be advised to take protective measures.

Several methods have been used in order to predict the violation of an air quality standard. Among them we may refer to [26], [39], [35] and [36] when the interest resides in applying extreme values theory to perform predictions. However, other techniques may also be used to study this type of problems. We may quote, multivariate analysis (see [24]), neural networks (see [1], [15], [23]), Poisson models (see [30], [28], [32]), and time series analysis (see [31]).

During the last decades some researchers have been looking at the possibility of using Markov chain models to predict the occurrence of a violation of the ozone environmental standard. Among them we have [5] and [29]. However, a common assumption of these works is that the Markov chain ruling the occurrence of a violation of the air quality standard has order one. Under this hypothesis, the transition probabilities of the chain are estimated using the maximum likelihood method.

Aiming to relax the assumption on the order of the Markov chain, [3] consider it a random variable and use a *maximum a posteriori* method to estimate it. Once the order is estimated, the transition probabilities are calculated using the value that maximises their marginal posterior distribution. The theoretical part of that work was applied to the ozone data collected during the year 2003 from the monitoring network of the Metropolitan Area of Mexico City. The results obtained show that the order of the chain may depend not only on the region considered but also on the environmental standard taken into account. Furthermore, the state space of the order of the chain could vary. An implication of the latter is that several supervised repetitions of the calculations using different state spaces for the order K should be performed in order to obtain

an appropriate state space that would allow a conclusive result. As an example of this situation, consider the case (see [3]) where we have the threshold $L = 0.11$ppm. Consider also the ozone data that represents the daily maximum values given by the five most representative monitoring stations of the Metropolitan Area of Mexico City during the year 2003. If the state space of the order K was chosen to be $\{0, 1, 2\}$, then we would have that the value of K that maximises its marginal posterior distribution is 2. However, letting the state space be $\{0, 1, 2, 3, 4\}$, we have that the maximum is in fact attained at $K = 3$ (see [3]).

Therefore, one possible option to avoid the type of situation described above could be to take the state space of K very large and calculate the value of the posterior distribution of K for each value in the state space. However, we could have the values of the probabilities very close to each other making it more difficult to decide what the appropriate order K is. Another option is to let K vary freely and settle on the value that is more representative of its true value. The later is the methodology pursued in the present work.

The novelty here is that instead of using covariance methods or any other classic method to estimate the order of the chain, we use a trans-dimensional MCMC algorithm (see [13] and [21]). The advantage of using this approach is that the fewer supervised runs of the algorithm would be necessary. Furthermore, even though the state space can be a large one, there will not be the problem of having very similar probabilitites (due to the association that should be made to all values in the state space of the chain) for the several possible values that the order may assume. This is due to the fact that even though there are many possibilities for the value of the order of the chain, only those that were in fact very significant would be the ones that would be visited by the MCMC algorithm. Hence, the presence of diluted probabilities values associated with elements of the state space would be due only to their similar importance in the path followed by the chain. Note that this algorithm allows the order of the chain to vary and using the usual empirical measures its distribution may be inferred. After the order of the chain is estimated we take as its transition probabilities those that maximise their marginal posterior distributions.

This work is presented as follows. In Section 2 a mathematical description of the model is given as well as the introduction of the parameters to be estimated. Section 3 contains the Bayesian formulation of the problem. A trans-dimensional Markov chain Monte Carlo algorithm (trans-dimensional MCMC) to sample from the posterior distribution is presented in Section 4. In Section

5 we apply the results to the ozone data registered by the monitoring network of the Metropolitan Area of Mexico City from 1 January 1998 to 31 December 2004. Section 5 also presents an analysis of the convergence of the Monte Carlo algorithm. Finally, in Section 6 some comments about the methodology and results obtained are given.

Since the only pollutant taken into account here is ozone, from now on we drop the "ppm" from the notation.

2. Mathematical Formulation of the Model

Let $S \subseteq \{0, 1, \ldots\}$ be the state space of a random variable K and let L be a fixed positive real number. The value L represents the threshold we are interested in knowing if it has been surpassed or not by a given pollutant. Denote by N the number of observed measurements and let it be such that $K \leq N$ with probability one. For $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_N)$ a sequence of daily maximum measurements of a given pollutant, define a sequence of random variables $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$ associated to $\mathbf{Z}$ in the following way. For $i = 1, 2, \ldots, N$

$$Y_i = \begin{cases} 1, & \text{if } Z_i \leq L\,, \\ 2, & \text{if } Z_i > L, \end{cases} \tag{1}$$

i.e., Y_i, $i = 1, 2, \ldots, N$ indicates if the environmental standard of interest has been violated or not on the i-th day.

Remark. Note that we could have a sequence of thresholds, say $L_1 < L_2 < \cdots < L_M$ and the sequence $\mathbf{Y}$ could assume values from one up to $M+1$ depending on which subinterval the measurement was. However, for simplicity we have chosen to work with just one threshold.

Assume that the sequence $\mathbf{Y}$ is ruled by a time homogeneous Markov chain of order K, indicated by $X^{(K)} = \{X_n^{(K)} : n = 0, 1, \ldots\}$, whose state space is given by $\chi_1^{(K)} = \{(x_1, x_2, \ldots, x_K) : x_i \in \{1, 2\},\, i = 1, 2, \ldots, K\}$.

In order to obtain a manageable transition matrix of $X^{(K)}$, we are going to associate $\chi_1^{(K)}$ to the set $\chi_2^{(K)} = \{0, 1, 2, \ldots, 2^K - 1\}$ using the function $f : \chi_1^{(K)} \to \chi_2^{(K)}$ given by $f((x_1, x_2, \ldots, x_K)) = \sum_{l=0}^{K-1} (x_{l+1} - 1)\, 2^l$.

Remarks. 1. Unless otherwise stated, from now on we use the set $\chi_2^{(K)}$ to represent the state space of $X^{(K)}$. We will also use the notation $\mathbf{x} = (x_1, \ldots, x_n) \leftrightarrow \overline{m}$ to indicate that $\mathbf{x} \in \chi_1^{(K)}$ corresponds to $\overline{m} \in \chi_2^{(K)}$.

2. For observations $(y_1, y_2, \ldots, y_N)$, if the present state of the chain is $X_n^{(K)} = (y_{n+1}, y_{n+2}, \ldots, y_{n+K}) \leftrightarrow \overline{m}$, $n = 0, 1, \ldots, N-K-1$, then the next state is $(y_{n+2}, \ldots, y_{n+K}, y_{n+K+1}) \leftrightarrow \overline{m}' \in \chi_2^{(K)}$ with probability different of zero. Hence, $\overline{m}' \in \chi_2^{(K)}$ occurs if and only if the observation following the sequence $y_{n+1}, y_{n+2}, \ldots, y_{n+K}$ is y_{n+K+1}. This allows us to use a reduced transition matrix of $X^{(K)}$. This matrix is indicated by $P^{(K)} = \left(P_{\overline{m}j}^{(K)}\right)_{\overline{m}\in\chi_2^{(K)}, j\in\{1,2\}}$ and is defined by

$$P_{\overline{m}j}^{(K)} = P(Y_{n+K+1}^{(K)} = j \,|\, X_n^{(K)} = (y_{n+1}, \ldots, y_{n+K}) \leftrightarrow \overline{m}),$$

for $n = 0, 1, \ldots, N-K-1$; $\overline{m} \in \chi_2^{(K)}$ and $j \in \{1, 2\}$. Note that $P^{(K)}$ is a $2^K \times 2$ stochastic matrix.

Therefore, we have that the parameter to be estimated is $\theta = (K, P^{(K)}) \in \Omega_K$, where $\Omega_K = \{K\} \times (\Delta_2)^{2^K}$ and $\Delta_2 = \{(x_1, x_2) \in \mathbb{R}^2 : x_i \geq 0, i = 1, 2;\ x_1 + x_2 = 1\}$. In general, $\theta \in \Omega = \bigcup_{K\in S} \Omega_K$.

3. Bayesian Formulation of the Model

Once the order of a Markov chain is known, there are several ways of estimating its transition probabilities. One classical way of doing so is through the maximum likelihood method and χ^2 tests (see for example [16] among others). An alternative method is the use of Bayesian estimators.

Instead of using directly the complete posterior distribution $P(K, P^{(K)} \,|\, \mathbf{Y})$ we are going to use the marginal posterior distributions of the parameters to obtain information about them. That is, we are going to use the marginal posterior distribution of K, i.e., $P(K \,|\, \mathbf{Y})$, and the marginal posterior distribution of $P^{(K)}$ given K, i.e., $P(P^{(K)} \,|\, K, \mathbf{Y})$. In order to do so, we use the fact that

$$P(K \,|\, \mathbf{Y}) \propto L(\mathbf{Y} \,|\, K)\, P(K)\,, \tag{2}$$

where $P(K)$ is the prior distribution of K and $L(\mathbf{Y} \,|\, K)$ is the marginal likelihood function given by (see [3])

$$L(\mathbf{Y} \,|\, K) \propto \prod_{\overline{m}\in\chi_2^{(K)}} \left[\frac{\Gamma(n_{\overline{m}1}^{(K)} + \alpha_{\overline{m}})\,\Gamma(n_{\overline{m}2}^{(K)} + \beta_{\overline{m}})}{\Gamma(n_{\overline{m}1}^{(K)} + \alpha_{\overline{m}} + n_{\overline{m}2}^{(K)} + \beta_{\overline{m}})} \frac{\Gamma(\alpha_{\overline{m}} + \beta_{\overline{m}})}{\Gamma(\alpha_{\overline{m}})\,\Gamma(\beta_{\overline{m}})}\right], \tag{3}$$

where $n_{\overline{m}j}^{(K)}$ records the number of transitions such that the state corresponding to $\overline{m} \in \chi_2^{(K)}$ is followed by the observation $j \in \{1, 2\}$.

Following suggestions appearing in the literature (see for instance [21]) we

assume that the *prior distribution of the order* K is a truncated Poisson distribution with parameter $\lambda > 0$, i.e.,

$$P(K) \propto \frac{\lambda^K}{K!} I_S(K) \tag{4}$$

where $I_A(x) = 1$ if $x \in A$ and is zero otherwise (other distributions may also be considered).

In order to obtain information about $P(P^{(K)} \,|\, K, \mathbf{Y})$ we assume that the lines $P_{\overline{m}}^{(K)}$, $\overline{m} \in \chi_2^{(K)}$ of the transition matrix $P^{(K)}$ are independently sampled from a Beta prior distribution with parameters $\alpha_{\overline{m}}$, $\beta_{\overline{m}} > 0$, $\overline{m} \in \chi_2^{(K)}$. Additionally, from [3], we have that

$$\begin{aligned} P(P^{(K)} \,|\, K, \tilde{\mathbf{Y}}) \;\; = \;\; & \prod_{\overline{m} \in \chi_2^{(K)}} \left[\frac{\Gamma(n_{\overline{m}1}^{(K)} + \alpha_{\overline{m}} + n_{\overline{m}2}^{(K)} + \beta_{\overline{m}})}{\Gamma(n_{\overline{m}1}^{(K)} + \alpha_{\overline{m}})\, \Gamma(n_{\overline{m}2}^{(K)} + \beta_{\overline{m}})} \right. \\ & \left. \left(P_{\overline{m}1}^{(K)}\right)^{n_{\overline{m}1}^{(K)} + \alpha_{\overline{m}} - 1} \left(1 - P_{\overline{m}1}^{(K)}\right)^{n_{\overline{m}2}^{(K)} + \beta_{\overline{m}} - 1} \right]. \end{aligned}$$

Therefore, if the value that maximise (2) is $K = k$, then the transition probabilities that maximise each row of the transition matrix $P^{(K)}$ is

$$P_{\overline{m}1}^{(k)} = \frac{n_{\overline{m}1}^{(k)} + \alpha_{\overline{m}} - 1}{n_{\overline{m}1}^{(k)} + n_{\overline{m}2}^{(k)} + \alpha_{\overline{m}} + \beta_{\overline{m}} - 2} = 1 - P_{\overline{m}2}^{(k)} \tag{5}$$

(see for example [17]).

The aim here is to obtain a sample $K_1, K_2, \ldots, K_J$ (J sufficiently large) of K and the usual empirical measures to estimate $P(K \,|\, \mathbf{Y})$. The algorithm presented here allows the order K to change to either $K + 1$ or $K - 1$, or to stay with the same value. This is achieved through a trans-dimensional MCMC method presented in the next section.

4. A Trans-Dimensional MCMC Algorithm

There are several works addressing the problem of estimating the order K of a Markov chain. Among them, we have [6] (pioneer in studying this type of problem) using likelihood methods (see also [16], [20] and [25]); [4] using χ^2 tests. In [40] we have the use of Akaike's information criterion (see [2]) to estimate this order. [18] use Bayes factor. Another work using Bayesian methodology, mainly the *maximum a posteriori* estimate is [3].

The present paper differs from previous work when we use a trans-dimensional

MCMC algorithm to estimate the order K of the chain from its marginal posterior distribution. The results are applied to the sequence recording the occurrence of a violation of a given threshold for ozone in the Metropolitan Area of Mexico City.

Trans-dimensional MCMC type algorithms (see [13] and [21]) have been used in several applications from the theoretical point of view to molecular biology (see for example [3], [8], [9], [10], [12], [14], [22], [33], [34], [37] and [38], and references therein). In the present work we use this type of algorithm to estimate the order of a sequence recording the violations of an environmental standard.

The algorithm used in the present work is described as follows. Assume that the present output is K. At time of a transition an independent choice is made when attempting one of the following three moves:

(a) increase by one the order of the chain (birth move);

(b) decrease by one the order of the chain (death move);

(c) do not change the order of the chain;

with probabilities b_K, d_K and r_K, respectively, depending only on the current order K and satisfying $b_K + d_K + r_K = 1$. The probabilities b_K and d_K are such that,

$$b_K = c \min\left\{1, \frac{P(K+1)}{P(K)}\right\} \quad \text{and} \quad d_K = c \min\left\{1, \frac{P(K-1)}{P(K)}\right\},$$

with $c > 0$ a suitable constant subject to $b_K + d_K < 1$. Note that the birth and death probabilities defined this way are such that the reversibility condition $d_{K+1}\, P(K+1) = b_K\, P(K)$ is satisfied.

The acceptance probabilities of each move are given as follows.

(a) Birth move: If a birth move is chosen, then increase the order of the chain by one. Accept the change with probability $\alpha(K, K+1) = \min\left\{1, \frac{L(\mathbf{Y} \mid K+1)}{L(\mathbf{Y} \mid K)}\right\}$.

(b) Death move: If a death move is chosen, then decrease the order of the chain by one. Accept the change with probability $\alpha(K, K-1) = \min\left\{1, \frac{L(\mathbf{Y} \mid K-1)}{L(\mathbf{Y} \mid K)}\right\}$.

If neither of those moves is chosen, then the chain stays with the same order. This move, when chosen, is always accepted.

Remarks. Note that changes of one in the order of the chain, reflects on the transition matrix by either increasing or decreasing the number of rows of $P^{(K)}$ by a factor of 2, respectively. However, since we are using the marginal posterior distribution of K, the changes occurring in $P^{(K)}$ do not appear explicitly in the

probability of accepting the changes of the dimension of the model.

After producing a sample $\{K_j, j = 1, 2, \ldots, J\}$ (for J sufficiently large) we may use the fact that the probabilities

$$P(K = k \,|\, \mathbf{Y}) = \frac{1}{J} \sum_{j=1}^{J} I_{\{k\}}(K_j), \quad k \in S\,,$$

where $I_{\{a\}}(X) = 1$ if $X = a$ and it is zero otherwise, estimate the posterior distribution of K. Once this posterior distribution is estimated, we use its mode as an estimator of the order of the chain. After that, the transition probabilities of the chain may be estimated using (5).

5. Simulation

In this section we give a brief description of the data used in the analysis. We also set the values of the parameters that are used in the sampling procedure given by the Monte Carlo algorithm. Finally, we present the results and the convergence assessment of the algorithm. Since the only pollutant considered here is ozone, from now on we omit the notation ppm.

5.1. Description of the Data

The data was obtained from the monitoring network of the Metropolitan Area of Mexico City (www.sma.df.gob.mx/simat/). The Metropolitan Area is split into five regions or sectors corresponding to the Northeast (NE), Northwest (NW), Centre (CE), Southeast (SE) and Southwest (SW) and the ozone monitoring stations are placed throughout the city (see for instance [1] and [3]). The measurements are obtained minute by minute and the averaged hourly result is reported at each station. The daily maximum measurement for a given region is the maximum over all the maximum averaged values recorded hourly during a 24-hour period by each station placed in the region.

The data used in the analysis corresponds to seven years, ranging from 1 January 1998 to 31 December 2004 (inclusive), of the daily maximum measurement of each region as well as the overall daily maximum measurements for the city (the latter is indicated by MAMC). In the case of MAMC, the daily maximum measurement is obtained by taking the maximum among the daily maximum values provided by regions NE, NW,CE, SE and SW.

The seven-year average measurements in regions NE, NW, CE, SE and SW

are 0.097, 0.121, 0.126, 0.143 and 0.122, respectively, with standard deviations 0.035, 0.049, 0.046, 0.052 and 0.042. For MAMC we have an average measurement of 0.154 with standard deviation of 0.05. During the period ranging from 1 January 1989 to 31 December 2004, the Mexican ozone standard of 0.11 was exceeded 816, 1046, 1631, 1581 and 1863 days in regions NE, NW, CE, SE and SW, respectively; and the daily peaks were double the Mexican standard 4, 64, 57, 38 and 178 days in regions NE, NW, CE, SE and SW, respectively (see [1]). Hence, when considering the measurements given by MAMC we have that during the same period there were 2063 days in which the threshold of 0.11 was surpassed and 237 days in which the measurements were above 0.22. Analysis were performed for each region and data set separately.

5.2. Setting the Parameters

Using the setting presented in Sections 2, 3 and 4 we have the following. The number of observed values is $N = 2557$. There are several choices for the parameter $\lambda > 0$ in (4). By a preliminary analysis of the data we reached the conclusion that one suitable choice could be $\lambda = 1$. This choice of λ was made based on the plots of the covariance function of the sequence of measurements for several lags for all zones including MAMC. On average, the maximum of the covariance function was for lag equal to one. The values of L used in (1) are $L = 0.11, 0.17, 0.23$. The reason for selecting these values is the following. The threshold 0.11 is the Mexican standard, hence its importance. The threshold 0.23 is approximately the standard used in Mexico City of declaring an emergency situation. We would like to call attention to the following. The value 0.11 is frequently surpassed in Mexico City and the value 0.23 is rarely exceeded. Therefore, we have decided to include a third threshold, namely $L = 0.17$ because it is an intermediate value between those two extremes. We would like to point out that during the period of time considered here, this threshold of 0.17 was exceeded in 95, 429, 433, 313 and 774 in regions NE, NW, CE, SE and SW, respectively. When considering the MAMC case the threshold 0.17 was surpassed in 980 days (see [1]).

The values of $(\alpha_{\overline{m}}, \beta_{\overline{m}})$, $\overline{m} \in \chi_2^{(K)} = \{0, 1, \ldots, 2^K - 1\}$ vary between 3 and 8 depending on $n_{\overline{m}i}^{(K)}$, $\overline{m} \in \chi_2^{(K)}$, $i \in \{1, 2\}$. The assignment of the values of $\alpha_{\overline{m}}$ and $\beta_{\overline{m}}$, for each $\overline{m} \in \chi_2^{(K)}$, was made using a function that associates to the maximum of $n_{\overline{m}i}^{(K)}$, $i = 1, 2$, the parameter 8 and assigns to the remaining parameter a value in $\{3, 4, 5, 6, 7\}$ depending on how distant it is from the maximum of $n_{\overline{m}i}^{(K)}$, $i = 1, 2$. When $n_{\overline{m}i}^{(K)} = 0$, then automatically the value 3 is

assigned to the parameter corresponding to it. For the constant c appearing in the trans-dimensional moves, we take $c = 0.35$. This choice was made to have the value of c as close as possible to the maximum value such that $d_K + b_K < 1$. In (4) we take $S = \{0, 1, \ldots, 13\}$. We use the values $K_0 = 1, 3, 5$ to initialise the algorithm.

5.3. Convergence Assessment

Since the model used here is a time homogeneous model we have split the data set into time homogeneous segments. The length of the segments varied according to the region and/or threshold considered. Hence, we have the following. If we take $L = 0.11$, then for regions CE, SW and for MAMC we have two time homogeneous segments. One containing the first 900 observations and another with the remainder of the data. When considering regions NE, NW and SE we have that the time homogeneous segments are: the first 400 observations, from observation 401 to 1000 and from observation 1001 to the end of the data set; the first 900 observations, from observation 901 to 1100 and from observation 1101 to the end of the data set; and the first 800 observations, from observation 801 to 1200 and from observation 1201 to the end of the data set, respectively. If $L = 0.17$ is considered, then for regions CE and NW we have two time homogeneous segments. The first one containing the first 900 observations and the second one containing the remaining data. Region SW and the case for MAMC have four time homogeneous segments described as follows. The first 300 observations, from observation 301 to 800, from observation 801 to 1200 and from observation 1201 to the end of the data set; and the first 500 observations, from observation 501 to 1000, from observation 1001 to 1500 and the one containing the reminder to the data, respectively. In region SE two time homogeneous segments were found. One containing the first 750 observations and another containing the remaining data. Region NE was composed of only one time homogeneous segment. If $L = 0.23$, then in all cases the sequences were time homogeneous.

Once we have identified the time homogeneous parts of the sequence $\mathbf{Y}$, we have performed preliminary runs in order to obtain an estimate of the transition matrix of the algorithm. Using the empirical estimator of the transition matrix of the chain recording the order K, we have obtained its largest eigenvalue smaller than one. Note that for a Markov chain with state space $\mathcal{E}$, n-step transition matrix $P^{(n)} = \left(P_{xy}^{(n)}\right)_{x,y\in\mathcal{E}}$, stationary distribution $\nu(\cdot)$ and with ρ the largest eigenvalue of $P^{(1)}$ that is smaller than one, we have that the total

variation distance $d(P_{x\cdot}^{(n)}, \nu)$ between $P_{x\cdot}^{(n)}$ and ν is such that $d(P_{x\cdot}^{(n)}, \nu) \leq \rho^n$. We have estimated the value of n such that $\rho^n \approx \epsilon$ for a given $\epsilon > 0$. Therefore, taking $\epsilon = 10^{-2}$ we have obtained that the convergence diagnostics would be reliable if we applied a data thinning to every 300 observations. Based on the results obtained we have decided to perform 6×10^5 iterations in order to assess the convergence to the marginal posterior distribution of K.

Convergence assessment was made following [12]. Therefore, we apply the chi-square test to the samples obtained using the algorithm given in Section 4 and taking into account three different initial values of the chain. (Other approaches can also be used, see for example [11], [19] and [38], among others.) In all cases convergence was achieved from the first 300 steps of the simulation. A sample of size 2000 was taken to perform estimation of the distribution of the order K.

5.4. Estimating the Order K of the Chain $X^{(K)}$ Produced by the Mexico City Ozone Data

In Table 1 we have the estimated probability distribution of the order of the Markov chain for each region and for the MAMC data when the thresholds $L = 0.11, 0.17$ and 0.23 are considered. We use (i), (ii), (iii) and (iv) next to the name of a region to indicate the time homogeneous segment of that region to which the values in that specific row in the table correspond.

In order to see how the methodology described in this paper can be used, take region SW for instance. This region is the one where the threshold 0.17 is surpassed in more days than any other (with exception of the data MAMC, but this is no longer used to declare emergency situations). Then we have that for threshold $L = 0.17$ the order K that has the highest probability varies from four to six depending on the period considered. When considering approximately the first four years of observations (i.e., 1998-2001) the order K tends to be higher than when considering later years. If $L = 0.11$ is considered, then in the first three years (approximately) the dependence seems to be very low. That phenomenon could be explained by the fact that in earlier years the threshold $L = 0.11$ would be exceeded practically every day and therefore the dependence on previous states could be disguised. However, in later years the level of ozone pollution has experienced a decrease. Hence the dependence of previous measurements would become clearer.

Predictions about the probability of having an ozone peak in a given day into the future given past measurements can be made in the following way.

		$K=0$	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$	$K=9$	$K=10$
NE	0.11 (i)	0	0	0	0	0.01	0.06	0.8	0.12	0.01	0	0
	0.11 (ii)	0	0	0	0	0	0.84	0.14	0.02	0	0	0
	0.11 (iii)	0	0	0	0	0	0	0	0.87	0.12	0.01	0
	0.17	0	0.03	0.26	0.26	0.06	0.33	0.06	0	0	0	0
	0.23	0.68	0.19	0.1	0.03	0	0	0	0	0	0	0
NW	0.11 (i)	0	0.02	0.05	0.28	0.2	0.22	0.2	0.03	0	0	0
	0.11 (ii)	0	0.02	0.01	0.78	0.16	0.03	0	0	0	0	0
	0.11 (iii)	0	0	0	0	0	0	0	0	0.89	0.1	0.01
	0.17 (i)	0	0	0	0	0	0	0.85	0.13	0.02	0	0
	0.17 (ii)	0	0.07	0.65	0.22	0.06	0	0	0	0	0	0
	0.23	0.73	0.15	0.08	0.03	0.01	0	0	0	0	0	0
CE	0.11 (i)	0	0	0	0	0	0	0.85	0.13	0.02	0	0
	0.11 (ii)	0	0	0	0	0	0	0	0.88	0.11	0.01	0
	0.17 (i)	0	0.05	0.47	0.15	0.05	0.23	0.04	0.01	0	0	0
	0.17 (ii)	0	0.12	0.18	0.06	0.12	0.44	0.07	0.01	0	0	0
	0.23	0.3	0.4	0.22	0.06	0.02	0	0	0	0	0	0
SW	0.11 (i)	0	0.59	0.27	0.1	0.03	0.01	0	0	0	0	0
	0.11 (ii)	0	0.01	0.04	0.02	0	0	0.05	0.76	0.11	0.01	0
	0.17 (i)	0	0	0	0	0	0.84	0.14	0.02	0	0	0
	0.17 (ii)	0	0.03	0.02	0	0	0.3	0.56	0.08	0.01	0	0
	0.17 (iii)	0	0	0	0	0	0.01	0.86	0.12	0.01	0	0
	0.17 (iv)	0	0	0	0.04	0.6	0.13	0.2	0.02	0.01	0	0
	0.23	0	0.57	0.3	0.1	0.02	0.01	0	0	0	0	0
SE	0.11 (i)	0	0.54	0.26	0.1	0.02	0.02	0.06	0	0	0	0
	0.11 (ii)	0	0.01	0.01	0	0.14	0.72	0.11	0.01	0	0	0
	0.11 (iii)	0	0	0	0	0	0	0	0.89	0.1	0.01	0
	0.17 (i)	0	0	0	0	0.32	0.56	0.1	0.02	0	0	0
	0.17 (ii)	0	0	0	0.77	0.18	0.05	0	0	0	0	0
	0.23	0.9	0.04	0.02	0	0	0	0	0	0	0	0
MAMC	0.11 (i)	0	0.58	0.3	0.1	0.02	0	0	0	0	0	0
	0.11 (ii)	0	0.01	0	0	0.1	0.75	0.13	0.01	0	0	0
	0.17 (i)	0	0.07	0.04	0.05	0.05	0.66	0.11	0.02	0	0	0
	0.17 (ii)	0	0	0	0	0	0	0.87	0.11	0.02	0	0
	0.17 (iii)	0	0	0	0	0.03	0.01	0.84	0.1	0.02	0	0
	0.17 (iv)	0	0.01	0.12	0.04	0.26	0.5	0.06	0.01	0	0	0
	0.23	0	0.24	0.1	0.14	0.42	0.08	0.02	0	0	0	0

Table 1: Probability function of the order K of the Markov chain $X^{(K)}$. Note that, even though $S = \{0, 1, \ldots, 13\}$, in the table are displayed only those states that were visited by the trans-dimensional algorithm.

First, one can see what the order of the dependence of the sequence of ozone peaks is. After that one may use expression (5) to make the desired prediction. As an example, consider region SW. Take for instance the more recent years measurements and the threshold $L = 0.11$. In this case (see Table 1, row corresponding to SW, 0.11 (ii)), we have that the value of K that maximises $P(K \mid \mathbf{Y})$ is $K = 7$, with a probability of 0.76. The corresponding state space of the chain $X^{(7)} = X$ is $\chi_2^{(7)} = \{0, 1, \ldots, 2^7 - 1\}$. Therefore, if we want to know what the probability of having the threshold $L = 0.11$ suprpassed in two days into the future given that today and in the past six days there were such violations. Hence, we want to know what is the probability of having state $(2, 2, 2, 2, 2, 2, 2)$ followed by a state $(2, 2, 2, 2, 2, 2, x)$ which is followed by a state $(2, 2, 2, 2, 2, x, 2)$. That is, we are interested in knowing what is the probability of having the state $\overline{m} = 127$ followed by an observation $x = 1$ or $x = 2$ and then observing $y = 2$. Therefore, we want $P_{\overline{m}1} P_{\overline{m}'2} + P_{\overline{m}2} P_{\overline{m}''2}$

where $\overline{m}' = (2, 2, 2, 2, 2, 1, 2)$ and $\overline{m}'' = (2, 2, 2, 2, 2, 2, 2)$. These probabilities may be calculated using the expression (5).

6. Discussion

The algorithm considered here in order to estimate the order K of a Markov chain is of fast convergence when applied to the ozone data provided by the monitoring network of Mexico City.

Note that the way the algorithm is constructed there is no need for updating the transition matrix of the Markov chain in each step. Therefore, few calculations are necessary during the implementation and execution of the algorithm.

Since covariates such as wind speed and direction and temperature are not included explicitly in the methodology applied here, it could be used as a first approach when making prediction in cities where the monitoring network is not as complete as the one in Mexico City. Even though covariates are not taken into account, differences in the ozone behaviour in different parts of the city are reflected in the results. This support the decision makers when they take action of declaring emergency situations regionally instead of declaring it at city level.

An improvement that could be made to the algorithm is to consider the parameter λ of the prior distribution of the order K, also a random variable with a suitable prior distribution. We could then estimate λ using the Bayesian approach instead of using the mean value of all the values obtained for the several regions through the covariance values.

It would be interesting to consider a version of the algorithm presented here that would allow the explicity inclusion of the covariates.

Acknowledgements

The second author was partially funded by CONACyT Grant Number 45684-F. The authors thank David Flores for providing a subroutine that allows infinite precision calculations.

References

[1] J.A. Achcar, A.A. Fernández-Bremauntz, E.R. Rodrigues, G. Tzintzun,

Estimating the number of ozone peaks in Mexico City using a non-homogeneous Poisson model, *Environmetrics* (http://www.interscience.wiley/10.002/env.890) (2007).

[2] H.A. Akaike, New look at the statistical model identification, *IEEE Transaction on Automatic Control*, **19** (1974), 716-723.

[3] L.J. Álvarez, A.A. Fernández-Bremauntz, E.R. Rodrigues, G. Tzintzun, Maximum a posteriori estimation of the daily ozone peaks in Mexico City, *Journal of Agricultural, Biological, and Environmental Statistics*, **10** (2005), 276-290.

[4] T.W. Anderson, L.A. Goodman, Statistical inference about Markov chains, *Ann. Math. Stat.*, **28** (1957), 89-110.

[5] J. Austin, H. Tran, A characterization of the weekday-weekend behavior of ambient ozone concentrations in California, In: *Air Pollution* **VII**, WIT Press, UK (1999), 645-661.

[6] M.S. Bartlett, The frequency goodness of fit test for probability chains, *Proc. Camb. Phil. Soc.*, **47** (1951), 86-95.

[7] M.L. Bell, A. McDermontt, S.L. Zeger, J.M. Samet, F. Dominici, Ozone and short-term mortality in 95 US urban communities, 1987-2000, *Journal of the American Medical Society*, **292** (2004), 2372-2378.

[8] R.J. Boys, D.A. Henderson, On determining the order of Markov dependence of an observed process governed by a hidden Markov model, *Special Issue of Scientific Programming*, **10** (2002), 241-251.

[9] R.J. Boys, D.A. Henderson, D.J. Wilkinson, A comparison of reversible jump MCMC algorithms for DNA sequences segmentation using hidden Markov models, *Comp. Sci. and Statist.*, **33** (2001), 35-49.

[10] S.P. Brooks, Trans-dimensional Markov chains and their applications in Statistics, In: *COMPSTAT 2002 Proceedings in Computational Statistics*, Springer, USA (2002), 91-102.

[11] S.P Brooks, P. Giudici, Markov chain Monte Carlo convergence assessment via two-way analysis of variance, *Journal of Computational and Graphical Statistics*, **9** (2000), 266-285.

[12] S.P. Brooks, P. Giudici, A. Philippe, Nonparametric convergence assessment for MCMC model selection, *Journal of Computational and Graphical Statistics*, **12** (2003), 1-22.

[13] B.P. Carlin, S. Chib, Bayesian model choice via Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society Series B*, **57** (1995), 473-484.

[14] J.M. Castelloe, D.L. Zimmerman, Convergence assessment for reversible jump MCMC samplers, *Technical Report*, **313**, Department of Statistics and Actuarial Sciences, University of Iowa, USA (2003).

[15] A.C. Comrie, Comparing neural network and regression models for ozone forecasting, *J. Air and Waste Manage.*, **47** (1997), 653-663.

[16] D. Dacunha-Castelle, M. Dulfo, *Probability and Statistics*, Volume **II**, Springer-Verlag, USA (1986).

[17] M. Evans, T. Swartz, *Approximating Integrals via Monte Carlo and Deterministic Methods*, Oxford Statistical Sciences Series, **20**, Oxford University Press, UK (2000).

[18] T.-H. Fan, C.-A. Tsai, A Bayesian method in determining the order of a finite state Markov chain, *Communications in Statistics - Theory and Methods*, **28** (1999), 1711-1730.

[19] Y. Fan, S.P. Brooks, A. Gelman, Convergence assessment of Monte Carlo simulation via score statistics, *Technical Report*, Statslab, University of Cambridge, UK (2003).

[20] I.J. Good, The likelihood ratio test for Markov chains, *Biometrika*, **42** (1955), 531-533.

[21] P.J. Green, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Journal of the Royal Statistical Society Series B*, **57** (1995), 711-732.

[22] P.J. Green, Trans-dimensional Markov chain Monte Carlo, In: *Highly Structured Stochastic Systems* (Ed-s: P.J. Green, N.L. Hjort, S. Richardson), Oxford University Press, UK (2005), 179-198.

[23] R. Guardani, C.A.O. Nascimento, M.L.G. Guardani, M.H.R.B. Martins, J. Romano, Study of atmospheric ozone formation by means of a neural

network based model, *J. Air and Waste Management Assoc.*, **49** (1999), 316-323.

[24] R. Guardani, J.L. Aguiar, C.A.O. Nascimento, C.I.V. Lacava, Y. Yanagi, Ground-level ozone mapping in large urban areas using multivariate analysis: application to the São Paulo Metropolitan Area, *J. Air and Waste Management Assoc.*, **53** (2003), 553-559.

[25] P.G. Hoel, A test for Markov chains, *Biometrika*, **41** (1954), 430-433.

[26] J. Horowitz, Extreme values from a nonstationary stochastic process: an application to air quality analysis, *Technometrics*, **22** (1980), 469-482.

[27] K. Itô, S. de León, M. Lippman, Associations between ozone and daily mortality: a review and additional analysis, *Epidemiology*, **16** (2005), 446-457.

[28] J.S. Javits, Statistical interdependencies in the ozone national ambient air quality standard, *J. Air Poll. Control Assoc.*, **30** (1980), 58-59.

[29] L.C. Larsen, R.A. Bradley, G.L. Honcoop, A new method of characterizing the variability of air quality-related indicators, In: *Air and Waste Management Association's International Specialty Conference of Tropospheric Ozone and the Environment*, California, USA (1990).

[30] M.R. Leadbetter, On a basis for "peak over threshold" modeling, *Statistics and Probability Letters*, **12** (1991), 357-362.

[31] D.P. Loomis, V.H. Borja-Arbuto, S.I. Bangdiwala, C.M. Shy, Ozone exposure and daily mortality in Mexico City: a time series analysis, *Health Effects Institute Research Report*, **75** (1996), 1-46.

[32] A.E. Raftery, Are ozone exceedance rate decreasing?, Comment of the paper "*Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone*" by R. L. Smith, *Statistical Sciences*, **4** (1989), 378-381.

[33] S. Richardson, P.J. Green, On Bayesian analysis of mixture with an unknown number of components (with discussion), *J. R. Stat. Soc. Series B*, **59** (1997), 731-792.

[34] C.P. Robert, T. Rydén, D.M. Titterington, Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo

method, *Journal of the Royal Statistical Society Series B*, **62** (2000), 57-75.

[35] E.M. Roberts, Review of statistics extreme values with applications to air quality data. Part I. Review, *Journal of the Air Pollution Control Association*, **29** (1979), 632-637.

[36] E.M. Roberts, Review of statistics extreme values with applications to air quality data. Part II. Applications, *Journal of the Air Pollution Control Association*, **29** (1979), 733-740.

[37] S.A. Sisson, Trans-dimensional Markov chains: a decade of progress and future perspectives, *Journal of the American Stat. Assoc.*, **100** (2005), 1077-1089.

[38] S.A. Sisson, Y. Fan, A distance-based diagnostic for trans-dimensional Markov chains, *Statistics and Computing*, **17** (2007), 357-367.

[39] R.L. Smith, Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone, *Statistical Sciences*, **4** (1989), 367-393.

[40] H. Tong, Determination of the order of a Markov chain by Akaike's information criterion, *Journal of Applied Probability*, **12** (1975), 488-497.