Using Counting Processes to Estimate the Number of Ozone Exceedances: an Application to the Mexico City Measurements

Achcar, Jorge A.
Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo
Av. Bandeirantes, 3900
14049-900 - Ribeirão Preto - SP, Brazil
E-mail: achcar@fmrp.usp.br

Rodrigues, Eliane R.
Instituto de Matemáticas, Universidad Nacional Autónoma de México
Area de la Investigación Científica - Circuito Exterior - Ciudad Universitaria
México DF 04510, Mexico
E-mail: eliane@math.unam.mx

Tarumoto, Mario H.
Faculdade de Ciências e Tecnologia, Universidade Estadual Paulista Júlio de Mesquita Filho
Rua Roberto Simonsen, 305
19060-900 - Presidente Prudente - SP, Brazil
E-mail: tarumoto@fct.unesp.br

INTRODUCTION

One common problem experienced by inhabitants of large cities throughout the world is the exposure to ozone air pollution. Even though several measures have been implemented by the environmental authorities as an effort to control ozone level in Mexico City, this pollutant still presents high concentration levels. It is a well known fact that individuals exposed for a long period of time to a high concentration of ozone may experience serious health problems (see for instance Bell et al., 2005; O'Neill et al., 2004; and Loomis et al., 1996). Hence, to understand the behaviour of ozone is a very important issue.

In Mexico, the ozone environmental standard (NOM, 2002) states that a person should not be exposed, on average, for a period of one hour or more to a concentration of 0.11 parts per million (0.11ppm) or above. The threshold used in Mexico City to declare emergency alerts is 0.2ppm (see for instance http://www.sma.df.gob.mx). When this threshold is surpassed, measures are taken to bring the level down through actions that may reduce the emission of ozone precursors. Nevertheless, the main benefit of those emergency alerts is to warn the population about the high levels of ozone and prevent human exposure to the pollutant.

Many works modelling strategies to predict pollution emergency episodes have been considered in the literature. Among those strategies we have extreme value theory, multivariate analysis, Markov chains, stochastic volatility models and Poisson models (homogeneous and non-homogeneous). Itô et al. (2005) and Seinfeld (2004) present a review of some of the statistical methodologies commonly used in the study of environmental problems.

In this paper, we use a more general counting process than the Poisson process to estimate the probability that a given environmental threshold is surpassed a certain number of times in a time interval of interest. Two cases are considered. In one of them we keep the assumption of independent inter-occurrences times (present in Poisson models), but we change their distribution to a Gamma distribution. In the second case, we keep the Gamma distribution for the inter-occurrences times and remove the independence assumption. The distribution of those inter-exceedances times will depend

on some parameters that need to be estimated. That will be performed using a Bayesian point of view via a Markov chain Monte Carlo (MCMC) algorithm. The results presented here are applied to the ozone data provided by the monitoring network of the Metropolitan Area of Mexico City.

DESCRIPTION OF THE MODELS

Assume that there are K (> 0) days, $d_1, d_2, \ldots, d_K$, in which a given ozone environmental threshold is surpassed during the time interval $[0, T]$ ($T > 0$). Let $\mathbf{D} = \{d_1, d_2, \ldots, d_K\}$ be the set of observed data and let W_i, $i = 1, 2, \ldots$ denote the time between the ith and the $(i-1)$th exceedances. Let $N = \{N_t : t \in [0, T]\}$ be such that N_t records the number of times that a threshold exceedance occurred in the time interval $[0, t)$, $t \geq 0$. Define $S_n = \sum_{i=1}^{n} W_i$, $n \geq 0$. Hence, we may write, $P(N_t = n) = P(S_n \leq t) - P(S_{n+1} \leq t)$. Therefore, the distribution of S_n determines the distribution of N_t. Hence, if we have information on the behaviour of the W_i, $i = 1, 2, \ldots$, then we also have information on the behaviour of S_n, $n \geq 0$ and consequently on the behaviour of N. Two models are considered here for the counting process. They are described as follows.

Model I. First of all, we assume that the inter-occurrences times W_i, $i = 1, 2, \ldots, K$ are independent and identically distributed with a Gamma(α, β) common distribution with mean α/β and variance α/β^2. Hence, in here the vector of parameter to be estimated is $\boldsymbol{\theta}_I = (\alpha, \beta)$, $\alpha > 0$, $\beta > 0$.

Model II. In this model, we keep the assumption of identically distributed Gamma inter-exceedances times, but now we remove the independence assumption. In order to specify the model used here consider the following (see Sim, 1990). Let $Y = \{Y_t : t \geq 0\}$ be a Poisson process with mean $(p\,\beta\,t)$, $p \in (0, 1)$ and $\beta > 0$. Take X_i, $i = 1, 2, \ldots$ independent and identically distributed quantities with common distribution an Exponential(β), $\beta > 0$, with mean $1/\beta$ and variance $1/\beta^2$. Also, take E_i, $i = 1, 2, \ldots$ independent and identically distributed Gamma(β, α) random variables, $\alpha, \beta > 0$. Let W_i, $i = 1, 2, \ldots$ be the inter-occurrences times. Hence, define (see Sim, 1990, 1992), $W_i = \sum_{j=1}^{Y(W_{i-1})} X_j + E_i$, $i = 1, 2, \ldots$. Assuming that W_i, $i = 1, 2, \ldots$ is in equilibrium we have, from Sim (1990), that W_i has a Gamma(β, $\alpha\,(1-p)$) density function, i.e., $f_{W_i}(t) = ([\alpha\,(1-p)]^{\beta}\, t^{\beta-1}\, e^{-\alpha\,(1-p)\,t})/\Gamma(\beta)$, with $\alpha, \beta, t > 0$ and $p \in (0, 1)$. We also have (Sim, 1990) that the joint density function of W_i and W_{i+1} is

$$f_{W_{i+1}W_i}(s, t) = \left(\frac{s\,t}{p}\right)^{(\beta-1)/2} \frac{\alpha^{\beta+1}\,(1-p)^{\beta} e^{-\alpha\,(s+t)}}{\Gamma(\beta)}\, I_{\beta-1}\left(2\,\alpha\,[p\,s\,t]^{1/2}\right),$$

where $I_r(z)$ is the modified Bessel function of the first kind of order r. Therefore, the conditional density of W_{i+1} given W_i is

$$f_{W_{i+1}\,|\,W_i}(s\,|\,t) = \left(\frac{s}{p\,t}\right)^{(\beta-1)/2} \alpha\, e^{-\alpha\,(s+p\,t)}\, I_{\beta-1}\left(2\,\alpha\,[p\,s\,t]^{1/2}\right). \tag{1}$$

Hence, the vector of parameters to be estimated here is $\boldsymbol{\theta}_{II} = (\alpha, \beta, p)$, $\alpha > 0$, $\beta > 0$ and $p \in (0, 1)$.

Parameters will be estimated by a Gibbs sample drawn from the respective complete marginal conditional posterior distribution of each coordinate of the vector of parameters which are given as follows (from now on we take $d_0 = 0$).

In the case of Model I, the likelihood function of the model is given by

$$L(\mathbf{D}\,|\,\boldsymbol{\theta}_I) \;\propto\; \left(\prod_{i=1}^{K} f_{W_i}(d_i - d_{i-1})\right) P(W_{K+1} > T - d_K),$$

where $P(W_{K+1} > t) = 1 - \int_0^t f_{W_{K+1}}(s)\,ds$. Hence,

$$L(\mathbf{D}\,|\,\boldsymbol{\theta}_I) \;\propto\; \left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right)^{K} \left[\prod_{i=1}^{K}(d_i - d_{i-1})\right]^{\alpha-1} e^{-\beta\,d_K} \left[1 - \frac{\beta^{\alpha}}{\Gamma(\alpha)} \int_0^{T-d_K} s^{\alpha-1}\, e^{-\beta\,s}\,ds\right].$$

Therefore, the complete marginal conditional posterior distributions of the parameters α and β are, $P(\alpha \,|\, \beta, \mathbf{D}) \propto \psi_1(\alpha, \beta)$ and $P(\beta \,|\, \alpha, \mathbf{D}) \propto \psi_2(\alpha, \beta)$ where,

$$\psi_1(\alpha, \beta) = \exp\left[K\,\alpha\,\log(\beta) - K\,\log\left[\Gamma(\alpha)\right] + (\alpha - 1)\sum_{i=1}^{K}\log(d_i - d_{i-1}) + h_1(\alpha, \beta, T, d_K)\right]$$

and $\psi_2(\alpha, \beta) = \exp\left[K\,\alpha\,\log(\beta) - \beta\,d_K + h_1(\alpha, \beta, T, d_K)\right]$ and where we define $h_1(\alpha, \beta, T, d_K) = \log\left[1 - \int_0^t f_{W_{K+1}}(s)\,ds\right]$.

When we consider Model II, the likelihood function of the model is

$$L(\mathbf{D} \,|\, \boldsymbol{\theta}_{II}) \;\propto\; f_{W_1}(d_1)\left(\prod_{i=2}^{K} f_{W_i \,|\, W_{i-1}}(d_i - d_{i-1} \,|\, d_{i-1} - d_{i-2})\right)\, P(W_{K+1} > T - d_K \,|\, W_K = d_K - d_{K-1}),$$

where $P(W_{i+1} > x \,|\, W_i = t) = 1 - \int_0^x f_{W_{i+1} \,|\, W_i}(s \,|\, t)\,ds$. Hence, we have that

$$\begin{aligned} L(\mathbf{D} \,|\, \boldsymbol{\theta}_{II}) \;\propto\; & \frac{\alpha^{\beta+K-1}\,(1-p)^{\beta}}{\Gamma(\beta)}\left[\frac{d_1\,(d_K - d_{K-1})}{p^{K-1}}\right]^{(\beta-1)/2}\exp\left(-\alpha(d_K + p\,[d_{K-1} - d_1]\right) \\ & \left[\prod_{i=2}^{K} I_{\beta-1}\left(2\,\alpha\,[p\,(d_i - d_{i-1})\,(d_{i-1} - d_{i-2})]^{1/2}\right)\right] \\ & \left[1 - \int_0^{T-d_K} f_{W_{K+1} \,|\, W_K}(s \,|\, d_K - d_{K-1})\,ds.\right] \end{aligned}$$

Therefore, the complete marginal conditional distributions of the parameters are $P(\alpha \,|\, \beta, p, \mathbf{D}) \propto \psi_3(\alpha, \beta, p)$, $P(\beta \,|\, \alpha, p, \mathbf{D}) \propto \psi_4(\alpha, \beta, p)$ and $P(p \,|\, \alpha, \beta, \mathbf{D}) \propto \psi_5(\alpha, \beta, p)$, where

$$\begin{aligned} \psi_3(\alpha, \beta, p) &= \exp\left[(\beta + K - 1)\log(\alpha) - \alpha\,[d_k + p\,(d_{K-1} - d_1)] + h(\alpha, \beta, p, T, d_{K-1}, d_K)\right], \\ \psi_4(\alpha, \beta, p) &= \exp\left[(\beta + K - 1)\log(\alpha) + \beta\,\log(1-p) - \log\left(\Gamma(\beta)\right)\right. \\ &- \left.\frac{(\beta-1)}{2}\left[(K-1)\,\log(p) - \log(d_1) - \log(d_K - d_{K-1})\right] + h(\alpha, \beta, p, T, d_{K-1}, d_K)\right], \\ \psi_5(\alpha, \beta, p) &= \exp\left[\beta\,\log(1-p) - \frac{(\beta-1)(K-1)}{2}\log(p) - \alpha\,p\,(d_K - d_{K-1}) + h(\alpha, \beta, p, T, d_{K-1}, , d_K)\right] \end{aligned}$$

with $h(\alpha, \beta, p, T, d_{K-1}, d_K) = \sum_{i=2}^{K} h_2(\alpha, \beta, p, i) + h_3(\alpha, \beta, p, T, d_{K-1}, d_K)$, where we take $h_2(\alpha, \beta, p, i) = \log\left[I_{\beta-1}\left(2\,\alpha\,[p\,(d_i - d_{i-1})\,(d_{i-1} - d_{i-2})]^{1/2}\right)\right]$ and

$$h_3(\alpha, \beta, p, T, d_{K-1}, d_K) = \log\left[1 - \int_0^{T-d_K} f_{W_{K+1} \,|\, W_K}(s \,|\, d_K - d_{K-1})\,ds\right].$$

The prior distributions for all parameters, models, regions and data sets are taken to be Uniform distributions defined on appropriate intervals. Those intervals are considered known and will be specified later.

The model that provides the best fit to the data is chosen via the Deviance Information Criterion (DIC) (see for instance Spiegelhalter et al., 2002). The DIC can be estimated by using the generated MCMC sample. The smaller the DIC, the better the fit of the model to the data. Usually a difference of DIC between two models that is larger than 10 is a strong evidence in favor of the best model (Burhan and Anderson, 2002).

AN APPLICATION TO MEXICO CITY OZONE DATA

Nowadays in Mexico City, environmental emergency alerts are issued locally instead of declaring it in the entire city. Hence, the Metropolitan Area of Mexico City has been divided into five sections 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 http://www.sma.df.gob.mx). When the threshold 0.2ppm is surpassed in one or more of the regions, then an environmental emergency alert is issued only in those regions. Therefore, measures are taken only in those parts of the city instead of the whole city. In this paper, we have considered the same spatial division used by the Mexico City's environmental authorities to declare those alerts. We will analyse the data from all five regions. The threshold that we are going to consider is the Mexican standard for ozone, i.e., 0.11ppm.

The data used in the analysis (http://www.sima.gob.mx/simat/) correspond to nineteen years (from 01 January 1990 to 31 December 2008) of the daily maximum measurements in each region giving a total of $T = 6940$ measurements (for a description of how the data are obtained see for instance Achcar et al., 2008). The nineteen-year average measurements in regions NE, NW, CE, SE and SW are 0.1279, 0.1006, 0.1332, 0.1262 and 0.1503, respectively, with respective standard deviations given by 0.0579, 0.0401, 0.0556, 0.0479 and 0.0617. We also have that the threshold 0.11ppm was surpassed in 4147, 2925, 4675, 4616 and 5307 days in regions NE, NW, CE, SE and SW, respectively.

In order to perform the analysis we have split the data into two parts, from 01 January 1990 to 31 December 1999 and from 01 January 2000 to 31 December 2008. The main reason for doing so is that around the year 2000 we have that the last major restriction on private vehicles circulating in the Metropolitan Area was implemented. We also have that from around the year 2000 the daily maximum measurements present a clear decreasing behaviour. That is easily seen when we observe that, for instance, during the period 1990-1999, in regions NE, CE, SE and SW the average measurements were above 0.14 and during the period 2000-2008, the maximum value of the average measurements was achieved in region SW with a value of 0.1246. During the same period in regions NE, NW, CE and SE the average measurements range from 0.0923 (in region NW) to 0.1087 (in region CE). We also have that during the period 1990-1999, in regions NE, CE, SE and SW, in more than 75% of the days the threshold 0.11ppm was surpassed. In region NW we that the percentage of surpassings is 51.42% of the total days of the period 1990-1999. However, during the period 2000-2008, there was a decreasing of those percentages, being region SW the one where we still have a high percentage of days (66.51%) where the threshold 0.11ppm was surpassed. In the other regions the percentage ranges from 31.81% (in region NW) to 54.29% (in region SE).

The analysis will be performed for each region, model and set of data separately. In all models, regions, parameters and data sets, the estimation of the parameters was made through a sample obtained by using a Gibbs sampling algorithm. Ten chains were run for each parameter and samples were drawn after a burn-in period of 10000 steps. After the burn-in period each chain was run another 10000 steps and every 100th generated value was taken to be part of the sample. Hence, each chain produced a sample of size 100 and therefore, estimation of the parameters was made using a sample of size 1000. Convergence analysis of the algorithm was performed through visual inspection of the trace plots of each chain as well as using the Gelman-Rubin test (see Gelman and Rubin, 1992).

Regarding the hyperparameters of the prior distributions we have that in the case of either Model I or Model II for all regions and data sets we have that the parameters α and β have Uniform prior distributions U(0,10). When Model II is taken into account, then the parameter p has a U(0,0.5) prior distribution. When the DIC is used to select the model that best fit the data, we have that for all regions and data sets the selected model was Model II. Hence, we are going to report the estimated parameters only for that model. Therefore, in Table 1 we have the mean, standard deviation (indicated by SD) and the 95% credible interval for all parameters, regions and data sets when Model II is used.

Conclusion

In this paper we have considered two models for studying the behaviour of the time between consecutive surpassings of a given environmental threshold by a pollutant's concentration. One of the models assume independence between two consecutive such times and the other allows for a dependence. The models were applied to ozone measurements obtained from the monitoring network of Mexico City. The threshold considered was the Mexican ozone standard of 0.11ppm. The model selected to explain the behaviour of the data was the one that allows for the dependence between two consecutive times between surpassings. That result corroborates the day to day experience that when 0.11ppm is considered as a threshold, the inter-occurrences times are dependent. Even though the selected model is the same for both period considered (1990-1999 and 2000-2008), the difference in the behaviour of the measurements is also captured by the selected model. That is expressed in Table 1 where we can note the difference in the values of the estimated parameters. The values of α and β are larger when using the 1990-1999 data. When considering the parameter p we have that, with the exception of region SW, it is larger when considering the 2000-2008 data. That is reflected when we consider the graphical behaviour of the estimated and observed inter-occurrences times conditional densities. When we consider the data 1990-1999, we have that the estimated conditional densities underestimate, but not by much, the observed conditional densities when we consider regions NE, CE and SE and inter-occurrences times smaller than two days. When considering regions NW and SW and two days inter-occurrences time intervals, we have that the estimated conditional densities underestimate a lot the observed conditional densities in the case of region NW and provides a good estimation in the case of region SW. In the case of inter-occurrences times with length larger or equal to two, the fitting is reasonable for all regions. When considering the 2000-2008 we also have an underestimation of observed conditional densities by the estimated ones and inter-exceedances times smaller than two days. However, the underestimation is really bad for all regions, with the exception of region SW, where the fitting is good. In the case of larger inter-occurrence times the estimation is reasonable.

In conclusion, we have that in the case of ozone and using the threshold 0.11ppm a model that considers a dependent behaviour between two consecutive times between exceedances is a more adequate model. However, when considering Gamma inter-occurrences times, in some cases the estimated conditional densities of those times does not provide a good fit to the observed conditional densities

Table 1. Estimated parameters of Model II.

		Mean		SD		95% Credible interval	
		90-99	00-08	90-99	00-08	90-99	00-08
NE	α	3.4445	0.7249	0.1035	0.0335	(3.2727; 3.6149)	(0.6696; 0.7789)
	β	4.4811	1.4505	0.1181	0.0507	(4.2820; 4.6711)	(1.3663; 1.5382)
	p	0.0235	0.1436	0.0109	0.0189	(0.0066; 0.0418)	(0.1107; 0.1756)
NW	α	1.0140	0.4130	0.0365	0.0221	(0.9513; 1.0759)	(0.3773; 0.4480)
	β	1.7752	1.1502	0.0524	0.0443	(1.6827; 1.8608)	(1.0783; 1.2240)
	p	0.0989	0.1163	0.0154	0.0233	(0.0750; 0.1236)	(0.0772; 0.1551)
CE	α	3.9870	1.0620	0.0123	0.0389	(3.9645; 3.9988)	(0.9976; 1.1229)
	β	4.9451	1.8740	0.0376	0.0557	(4.8733; 4.9941)	(1.7860; 1.9669)
	p	0.0105	0.0827	0.0069	0.0150	(0.0012; 0.0232)	(0.0589; 0.1089)
SW	α	4.5652	2.1863	0.0792	0.0719	(4.4414; 4.6998)	(2.0774; 2.3062)
	β	4.9916	3.0730	0.0085	0.0883	(4.9707; 4.9998)	(2.9443; 3.2264)
	p	0.0657	0.0650	0.0147	0.0142	(0.0405; 0.0888)	(0.0434; 0.0885
SE	α	3.8651	1.2201	0.0664	0.0471	(3.7451; 3.9717)	(1.1428; 1.2969)
	β	4.9132	2.0899	0.0646	0.0666	(4.7878; 4.9930)	(1.9839; 2.2029)
	p	0.0147	0.0709	0.0089	0.0162	(0.0024; 0.03)	(0.0454; 0.0963)

when the present data is taken into account. In those cases perhaps considering a different form than the Gamma density could be more adequate, but this is the subject of another study.

Acknowledgements

This work was financially supported by the PAPIIT project number IN104110-3 of the DGAPA-UNAM, Mexico. JAA received partial financial support form CNPq-Brazil grant number 300235/2005-4. ERR thanks the Faculdade de Ciencias e Tecnologia of UNESP-Campus Presidente Prudente, Brazil, for the hospitality and support during the development of this work.

REFERENCES (RÉFERENCES)

1. Achcar, J.A., Fernández-Bremauntz, A.A., Rodrigues, E.R. and Tzintzun, G. (2008). Estimating the number of ozone peaks in Mexico City using a non-homogeneous Poisson model. *Environmetrics* **19**, 469-485.
2. Bell, M.L., Peng, R. and Dominici, F. (2005). The exposure-response curve for ozone and risk of mortality and the adequacy of current ozone regulations. *Environmental Health Perspectives* **114**, 532-536.
3. Burham, K.P. and Anderson, D.A. (2002). *Model Selection and Multivariate Inference: a Practical Information - Theoretic Approach*. Second Edition. Springer, New York.
4. Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Sciences* **7**, 457-511.
5. Itô, K., de León, S. and Lippman, M. (2005). Associations between ozone and daily mortality: a review and additional analysis. *Epidemiology* **16**, 446-457.
6. Loomis, D.P., Borja-Arbuto, V.H., Bangdiwala, S.I. and Shy, C.M. (1996). Ozone exposure and daily mortality in Mexico City: a time series analysis. *Health Effects Institute Research Report* **75**, 1-46.
7. NOM (2002). *Modificación a la Norma Oficial Mexicana NOM-020-SSA1-1993*. Diario Oficial de la Federación. 30 de Octubre de 2002, Mexico. (In Spanish.)
8. O'Neill, M.R., Loomis, D. and Borja-Aburto, V.H. (2004). Ozone, area social conditions and mortality in Mexico City. *Environmental Research* **94**, 234-242.
9. Seinfeld, J.H. (2004). Air pollution: a half century of progress. *American Institute of Chemical Engineers Journal* **50**, 1098-1108.
10. Sim, C.H. (1990). First-order autoregressive models for Gamma and Exponential processes. *Journal of Applied Probability* **27**, 325-332.
11. Sim, C.H. (1992). Point processes with correlated Gamma interarrival times. *Statistics and Probability Letters* **15**, 135-141.
12. Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion and rejoinder). *Journal of the Royal Statistical Society Series B* **64**, 583-639.