v. 1, n. 1, 2024

International Journal of
Exact
Sciences

OZONE STUDY IN MEXICO CITY: AN APPLICATION OF THE HIDDEN MARCOV PROCESSES

Blanca Rosa Pérez Salvador
Universidad Autónoma Metropolitana, Iztapalapa, Ciudad de México-México
https://orcid.org/0000-0002-9523-0700

Abstract: Ozone in the troposphere is a harmful gas for life on Earth, so its study and control is very important for us. In this paper, the ozone concentration data of Mexico City from 1992 to 2015 are analyzed as annual time series using hidden Markov models (HMM). It is considered that the different ozone generators form a sequence of states in a stochastic process that can be modeled with a Markov chain, and the ozone concentration measurements from the monitoring centers are random variables whose distribution function depends on the state of the Markov chain at time *t*. Two distribution functions were considered for the observed data, the normal and the Gamma, and the model parameters were estimated using EM method. One, two and up to seven states for the Markov chain were carried out. Finally, 222 models were estimated and the best model was selected for each year using the Bayesian information criterion, BIC. It is concluded that this model describes the data well.
Keywords: Markov Processes, Tropospheric Ozone, EM Method, Normal Distri- bution, Gamma Distribution.

INTRODUCTION

Ozone (O_3) is found in the Stratosphere, there it helps protect life on Earth from the sun's harmful ultraviolet rays. Ozone is also found in the Troposphere, there, according to Commission for Environmental Cooperation CEC (2008), Velázquez and Jiménez (2007), Sousa et al. (2007) and Filleul et al. (2006), even in small quantities, ozone has harmful effects on animals, plants and some materials. According to Davis et al. (2006), Scebba et al. (2006) Escobedo and Chacalo (2008), Tropospheric ozone affects the car- diovascular and respiratory systems, damages the eyes, reduces the yield of cultivated plants, deteriorates clothing made of cotton and synthetic materials, accelerates the fading of certain paints and coatings. Velázquez and Jiménez (2007) reports that tro- pospheric ozone is generated by photochemical reactions between nitrogen oxides (NO_x) and volatile organic compounds, (VOC), especially hydrocarbons.

The Mexican Official Norm (NOM-020-SSA1-2014) recommends maintaining concentrations less than 0.095 ozone units per million units of air (ppm) for the 1-hour average, and less than 0.070 units ppm for the 8-hour average.

OZONE LEVEL AS A HIDDEN MARKOV PROCESS

It is proposed to use homogeneous Hidden Markov Models (HMM), as an explora- tory statistical technique to analyze the sources of ozone contribution. A Hidden Markov Model according Cappé (2005) consists of two stochastic processes that occur simultaneously. The first of them is the sequence of states of a Markov Chain with k states X_t, and the second is a sequence of random variables O_t, whose distribution function depends on the state of the Markov Chain over time t.

In this model, the values of the second stochastic process, are observable and the sequence of states of the Markov Chain are kept hidden.

$$\begin{array}{ccccccccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & \dots & X_T \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \dots & \downarrow \\ O_1 & O_2 & O_3 & O_4 & O_5 & O_6 & O_7 & \dots & O_T \end{array}$$

The parameters of the Markov process are (1) the initial vector

$$\pi^T = (P(X_t = 1), P(X_t = 2), \dots,$$

$$P(X_t = k)) = (\pi_1, \pi_2, \dots, \pi_k),$$

(2) the transition matrix from one state to another

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{pmatrix}$$

with $p_{ij} = P\ (X_t = j|X_{t-1} = i)$, and (3) the parameters of the second stochastic process, $E(O_t|X_t = i)$ and $V\ (O_t|X_t = i)$.

Ozone levels are grouped into subsets of similar concentration which are called con- centration regimes or simply regimes. Each regime corresponds to a state in the HMMs and therefore each regime generates a distribution for the observations O_t. Of particular interest is the mean of the distribution for each state or regime, because it represents the concentration of different sources of pollutants; in this regard, Lenschow et al. (2001), establish that:

1. The mean of the first state or regime represents the real concentration of the pollutant, it is the background concentration.

2. The mean of the second state represents the concentration on days affected by an increase in contributions of anthropogenic origin, due to human activities carried out in the study region.

3. The third state average represents a concentration that exceeds the daily limit value established by official standards for some pollutant.

4. Finally, the mean in the fourth state of the Markov chain represents the concentration on those days affected by severe episodes and generally infrequent.

The average concentration per year, μ, is the ambient pollution of the place.

DATA

Data from http://www.aire.df.gob.mx/,the Mexico City Atmospheric Monitoring System website were used for the analysis; The average hourly concentration in IMECA units is reported there for the five regions of the city: northwest, northeast, center, sout- heast and southwest. With this information, daily averages were obtained. The analysis was done by year for the five regions, from 1992 to 2015, giving a total of 120 series analyzed.

PROCEDURE

Following Altman's proposal (2004), Pearson-type goodness-of-fit tests were first performed for each of the 120 series in order to determine whether the data satisfied the assumptions of the homogeneous Hidden Markov Models.

The HMMs were then fitted to the 120 series using the EM method as Bilmes (1998) indicated. Initial values for the iterative EM method were proposed, and after a certain number of iterations the final parameters corresponding to the HMM estimates were obtained. For the observations of the second stochastic process, two distributions were considered, the Normal and the Gamma.

Seven possible models were considered, with one, two, three, four, five, six and seven states in the Markov Chain associated with the HMM. The best model selection was done with Bayesian Information Criterion, BIC.

CASE *K* STATES

To start the EM method, in each annual data series, the following values were given:

1. The initial vector π is π = (1/*k*, 1/*k*, 1/*k*, ..., 1/*k*),

2. The initial state sequence is done as: if

$$\min_{O_t} -(i-1)(\max_{O_t} - \min_{O_t})/k \leq O_t < \min_{O_t} + i(\max_{O_t} - \min_{O_t})/k \quad \text{then} \quad X_t = i$$

International Journal of Exact Sciences ISSN 2966-1153 DOI https://doi.org/10.22533/at.ed.153112401078

RESULTS FROM NORMAL DISTRIBUTION ARE

$$\hat{\pi} = (1, 0)$$

$$P = \begin{pmatrix} 0.964 & 0.036 \\ 0.052 & 0.948 \end{pmatrix}$$

$$\hat{\mu}_1 = 28.398 \quad \hat{\mu}_2 = 42.675$$
$$\hat{\sigma}_1 = 6.458 \quad \hat{\sigma}_2 = 7.328$$

Number of observations in state 1=216

Number of observations in state 1=149

State 1 observations range: (9:5847:50)

State 2 observations range: (26:4561:33)

M = 34:14 SD = 9:77

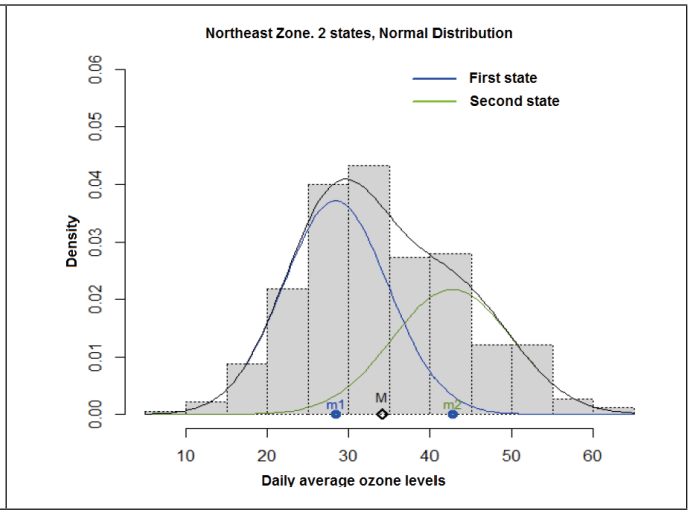

Fig 1. Results of the best model with 2002 data and Normal distribution

$$\hat{\pi} = (0, 1, 0, 0)$$

$$P = \begin{pmatrix} 0.763 & 0.219 & 0.017 & 0.000 \\ 0.082 & 0.799 & 0.119 & 0.000 \\ 0.000 & 0.215 & 0.698 & 0.087 \\ 0.000 & 0.000 & 0.146 & 0.854 \end{pmatrix}$$

$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$
21.05	29.37	38.18	48.14

$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_3$	$\hat{\sigma}_4$
4.056	3.58	3.71	5.999

Number of observations in state 1 =53

Number of observations in state 2 =163

Number of observations 3n state 3 =97

Number of observations 4n state 4 =53

State 1 observations range: (8:25; 28:87)

State 2 observations range: (21:58; 38:16)

State 3 observations range: (31:58; 47:00)

State 4 observations range: (37:29; 67:70)

M = 33.39 SD = 9.27

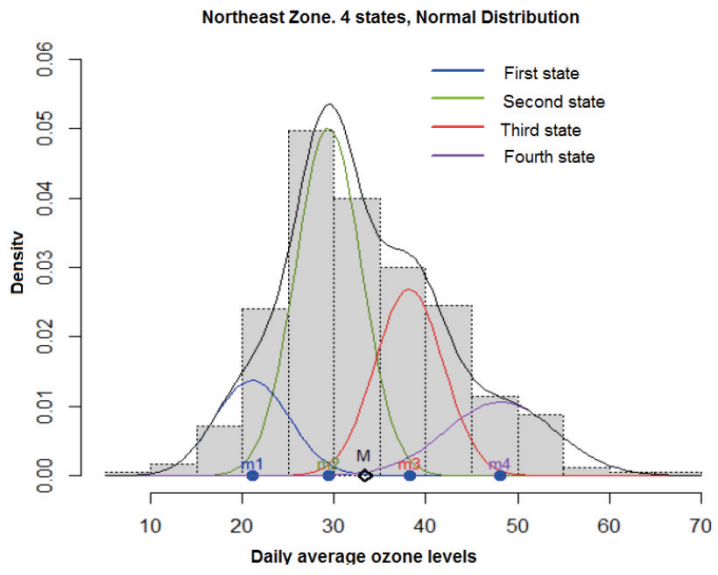

Fig 2. Results of the best model with 2012 data and Normal distribution

$$\hat{\pi} = (1, 0, 0)$$

$$P = \begin{pmatrix} 0.862 & 0.138 & 0.00 \\ 0.111 & 0.760 & 0.129 \\ 0.00 & 0.261 & 0.739 \end{pmatrix}$$

$$\hat{\mu}_1 = 18,894 \quad \hat{\mu}_2 = 27.996 \quad \hat{\mu}_3 = 40.224$$
$$\hat{\sigma}_1 = 4,324 \quad \hat{\sigma}_2 = 4.912 \quad \hat{\sigma}_3 = 6.343$$

Number of observations in state 1 =135

Number of observations in state 2 =156

Number of observations in state 3 =74

State 1 observations range: (7:79; 30:20)

State 2 observations range: (17:04; 38:37)

State 3 observations range: (26:66; 58:57)

M = 27.23 SD = 9.25

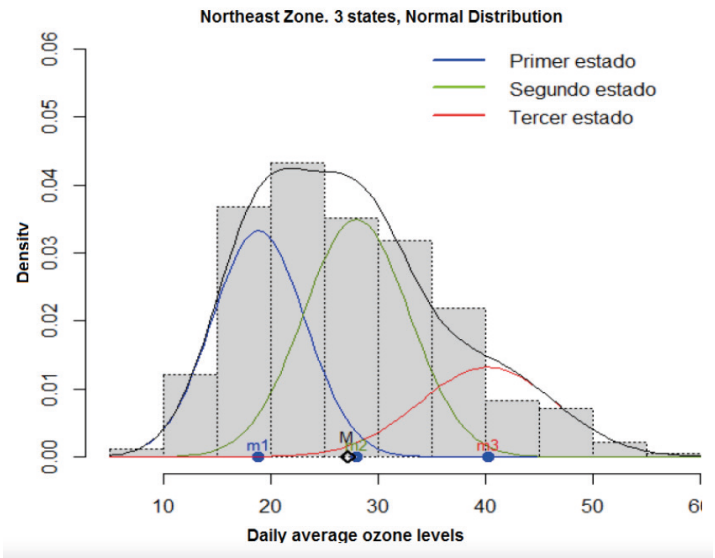

Fig 3. Results of the best model with 2015 data and Normal distribution

RESULTS FROM GAMMA DISTRIBUTION ARE

$$\hat{\pi} = (0, 1, 0, 0)$$

$$P = \begin{pmatrix} 0.7310.2440.0250.000 \\ 0.0870.9050.0080.000 \\ 0.0000.0370.8140.149 \\ 0.0000.0000.1610.839 \end{pmatrix}$$

$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$
27.54	30.50	62.71	38.12 .
$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_3$	$\hat{\sigma}_4$
0.889	1.194	1.123	1.326

Number of observations in state 1 =42

Number of observations in state 2 =147

Number of observations 3n state 3 =88

Number of observations 4n state 4 =88

State 1 observations range: (13:45; 31:95)

State 2 observations range: (22:16; 59:16)

State 3 observations range: (31:40; 68:26)

State 4 observations range: (53:80; 94:66)

M = 46.41 SD = 17.22

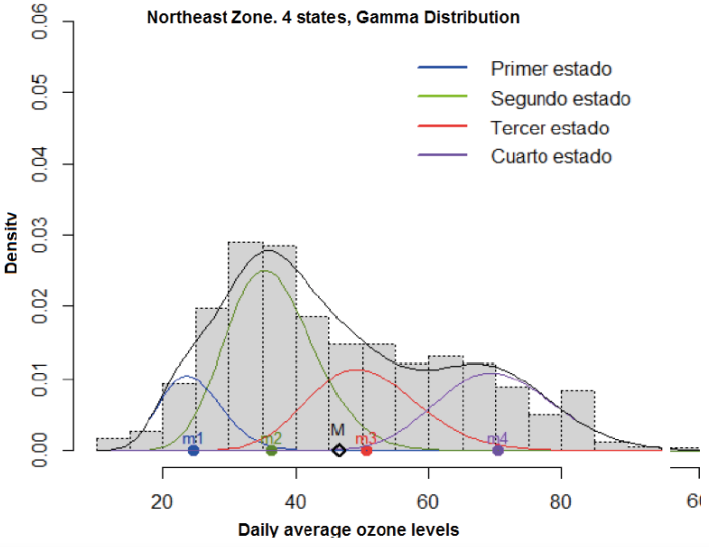

Fig 4 Results of the best model with 1998 data and Gamma distribution

$\hat{\pi} = (0, 1)$

$$P = \begin{pmatrix} 0.979 & 0.021 \\ 0.038 & 0.962 \end{pmatrix}$$

$\hat{\mu}_1 = 21.52 \quad \hat{\mu}_2 = 36.22$
$\hat{\sigma}_1 = 1.187 \quad \hat{\sigma}_2 = 1.051$

Number of observations in state 1=222

Number of observations in state 1=143

State 1 observations range: (12:45; 47:29)

State 2 observations range: (22:66; 59:33))

M = 30:56 SD = 8:47

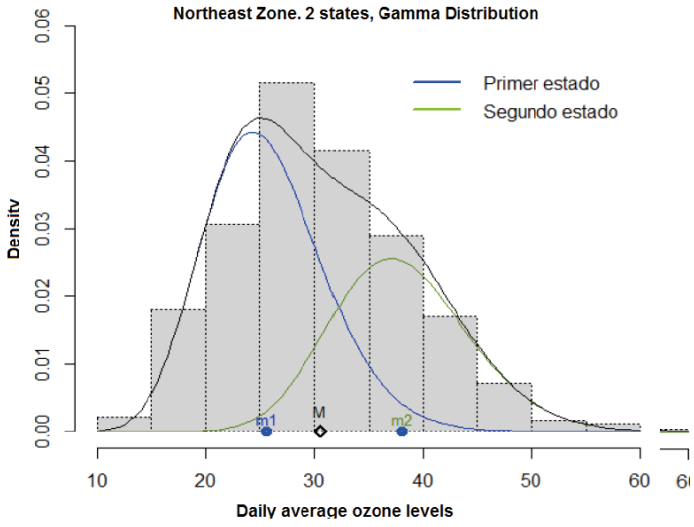

Fig 5. Results of the best model with 2006 data and Gamma distribution

$\hat{\pi} = (0, 1, 0)$

$P = ()$

$\hat{\mu}_1 = 30.40 \quad \hat{\mu}_2 = 38.79 \quad \hat{\mu}_3 = 38.32$
$\hat{\sigma}_1 = 0.669 \quad \hat{\sigma}_2 = 0.773 \quad \hat{\sigma}_3 = 1.128$

Number of observations in state 1 =62

Number of observations in state 2 =169

Number of observations 3n state 3 =134

State 1 observations range: (10:12; 26:04)

State 2 observations range: (19:20; 44:41)

State 3 observations range: (27:54; 64:66)

Rango observaciones estado 1: (10:12; 26:04)

Rango observaciones estado 2: (19:20; 44:41)

Rango observaciones estado 3: (27:54; 64:66)

M = 32.92 SD = 10.03

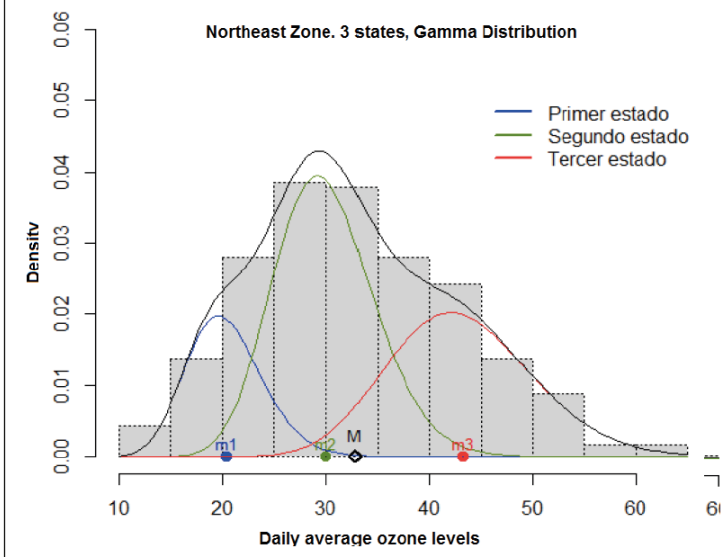

Figure 6. Results of the best model with 2014 data Gamma distribution

3. The transition probability p_{ij} was calculated as follows: the number of times that the state *i* transitioned to state *j* was counted and this result is divided by the number of times that state *i* was visited.

4. Finally, the observed data were grouped into *k* subsets according to the sequence of states, for each state the initial parameters of the distribution (Normal or Gamma) are calculated by maximum likelihood method.

The algorithm for the calculation described was implemented with the R software, https://www.r-project.org/.

RESULTS

Only in 9 of the 120 series the goodness of fit test were rejected, because HMM were not suitable for modeling their behavior. It is important to note that these 9 series correspond to the first years of registration, whose records had many missing data. The analysis was done with the remaining 111 series, with the two distributions (Normal and Gamma), thus obtaining a total of 222 models. In this section only 6 of these 222 models are shown, three from the Normal distribution for the years 2002, 2012, and 2015 and three from the Gamma distribution for the years 1998, 2006, and 2014. All of them with data from the northeast area because in this area the best models the number of states is more variable.

In addition to the parameters of the HMM model, it is of interest to analyze the relationship between background pollution $\hat{\mu}_1$ and ambient pollution μ, that is, μ/m_1. When this ratio is close to 1, it means that pollution in that year and in that area is almost entirely explained by background pollution. These values are shown by area in the following tables.

Zone	$\hat{\mu}$	$\hat{\mu}_1$	$\hat{\mu} - \hat{\mu}_1$	$\frac{\hat{\mu}}{\hat{\mu}_1}$	%	*Range*
Northwest	40.97	27.95	13.01	1.42	0.71	0.55 – 0.85
Northeast	38.44	26.28	12.16	1.41	0.72	0.54 – 0.84
Center	37.45	23.83	13.62	1.49	0.69	0.53 – 1.00
Southwest	47.8	28.88	18.93	1.59	0.64	0.52 – 0.80
Southeast	42.35	28.79	13.56	1.48	0.7	0.46 – 1.00

Table 1. Averages of $\hat{\mu}$ and $\hat{\mu}_1$ of the best annual models with Normal distribution.

Zone	$\hat{\mu}$	$\hat{\mu}_1$	$\hat{\mu} - \hat{\mu}_1$	$\frac{\hat{\mu}}{\hat{\mu}_1}$	%	*Range*
Northwest	40.98	30.2	10.78	1.37	0.73	0.56 – 0.88
Northeast	38.47	28.15	10.32	1.37	0.73	0.63 – 0.84
Center	38.23	27.07	11.16	1.43	0.7	0.54 – 0.82
Southwest	47.84	31.28	16.56	1.54	0.66	0.44 – 0.84
Southeast	42.93	30.82	12.11	1.42	0.71	0.50 – 0.93

Table 2. Averages of $\hat{\mu}$ and $\hat{\mu}_1$ of the best annual models with Gamma distribution.A test of paired differences indicates that μ and $\hat{\mu}_1$ are significantly the same for the Normal and Gamma distributions. The following tables show this.

Zone	t estatitics	P-value	Decision
Northwest	t=1.7664	0.0906	H_0 is not rejected
Northeast	t=1.4809	0.1522	H_0 is not rejected
Center	t=1.8235	0.0812	H_0 is not rejected
Southwest	t=0.8196	0.4208	H_0 is not rejected
Southeast	t=1.7845	0.0875	H_0 is not rejected

Table 3. Paired means test for the contribution of background pollution.

H_o : Contribution of background pollution, μ_1, is the same under assumption of a HMM with gamma and Normal observations.

Zone	t estatistics	P-value	Decision
Northwest	t=-0.2726	0.7876	No se rechaza Ho.
Northeast	t=-1.9756	0.0603	No se rechaza Ho.
Center	t=-1.7426	0.0947	No se rechaza Ho.
Southwest	t=-1.0518	0.3038	No se rechaza Ho.
Southeast	t=-1.0181	0.3192	No se rechaza Ho.

Table 4. Paired means test for the estimate of mean ambient pollution.

H_o : The mean level of ambient pollution, μ, is the same when it is assumed that the daily average levels are distributed according to a HMM with Gaussian observations and gamma observations.

CONCLUSIONS

This document illustrates how a HMM describes aspects of average daily ozone levels in the five Mexico City zones. Due to serious impact of this pollutant on human health, it is important to analyze the problem in order to evaluate public policies aimed at maintaining good air quality.

The main aspects observed in the application of HMM to ozone time series in Mexico City were:

1. Between 64 and 72 percent of air pollution is explained by μ_1, making it an important indicator of actual exposure to ozone.

2. The average annual ozone estimate, $\hat{\mu}$, in the five zones of Mexico City indicates a decrease of 66.6 percent over the observation time. This shows that from 1992 to 2015 the actions to reduce ozone pollution have been effective.

3. It is concluded that when the average daily pollution level is between (0, 63.63), the best models has two or three states, and if the pollution is higher on several days of the year, four states are required.

4. In average, the elements on transition matrix diagonal are almost 1, so it is more likely to remain in the same state the next day. That is, it is unlikely that the next day of low pollution, there will be high pollution and vice versa.

5. The BIC criterion indicated that the Normal best describes the data observed from 1992 to 2000, and that in later years it was better to use the gamma distribution.

HMMs are a tool for analyzing time series of pollutants; because

1. It allows to establish a relationship between the states of a Markov chain and the pollution regimes, and is also a tool to analyze the contribution of pollutant sources.

2. In particular, it allows to characterize the background contamination associated with the first state of the Markov chain.

3. It is possible to know the contribution of each pollution regime with respect to ambient pollution, by exploring the relationship between the mean of each regime m_i and the annual mean, μ, of the pollutant data series.

4. The probability of change between the different states of the Markov chain can be established, therefore, the probability of change between the different pollution regimes.

5. They represent an individual statistical analysis, which stands out for its solid theo- retical support, which also has the ease of its interpretation and the reproducibility of its results; as well as the fact that the modeling can be carried out using free software such as R.

It would be interesting to incorporate other variables into the model such as ambient temperature, wind speed, other particles, etc., and model through a multi-variate distribution HMM on the observations; this could offer a different approach to the variables that influence ozone formation.

This analysis can be performed for other air pollutants such as sulfur dioxide, carbon monoxide, nitrogen dioxide and suspended particles, which could provide more information on the background pollution as a whole and on the relevance of each in the metropolitan air quality index.

REFERENCES

[1] Altman, R.M (2004). Assessing the Goodness-of-Fit of Hidden Markov Models. Bio- metrics, Vol. 60(2), pp. 444-450.

[2] Bilmes, J.A. (1998) A gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Interna- tional Computer Science Institute Berkeley CA, pp. 1-13.

[3] Cappé, O., Moulines, E. y Rydén, T. (2005) Inference in hidden Markov models. Springer, New York.

[4] Comisión para la Cooperación Ambiental, CEC. (2008). Aire y Atmósfera. Ozono troposférico. El Mosaico de América del Norte panorama de los problemas ambien- tales más relevantes. pp 1-4. http://www3.cec.org/islandora/es/item/986-north-american-mosaic-overviewkey-environmental-issues-es.pdf

[5] Davis, D.D., and Orendavici, T. (2006). Incidence of ozono symptons on vegetation within National Wildlife Refuge in New Jersey, USA. Environmental Pollution, Vol. 143(3), pp. 555-564.

[6] Escobedo, F., and Chacalo, A. (2008) Estimación preliminar de la descontaminación atmosférica por el arbolado urbano de la Ciudad de México. Interciencia, Vol 33(1). pp 29.

[7] Filleul, L., Cassadou, S., Médina, S., Fabres,P., Lefranc, A., Eilstein, D., Le Tertre, A., y Ledrans, M. (2006). The relation between temperature, ozone, and mor- tality in nine French cities during the heat wave of 2003. Environmental Health Perspectives, 114 (9), pp. 1344-1347.

[8] Lenschow, P., Abraham, H.J., Kutzner, K., Lutz, M., Preub, J.D., Reichenbacher, W. (2001) Some ideas about the sources pp PM10. Atmosferic Environmental", Vol. 35, pp. S23-S33.

[9] Scebba, F., Canaccini, F., Castagna, A., Bender, J.,Weigel, H., y Ranieri, A. (2006) Physiological and biochemical stress responses in grassland species are influenced by both early-season ozone exposure and interspecific competition. Environmental Pollution, 142 (3), pp 540-548

[10] Sousa, S.I.V., Martins, F.G., Alvim-Ferraz, M.C.M., y Pereira, M.C. (2007) Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environmental modelling and software, 22(1), pp. 97-103.

[21] Velázquez, F. C., and Jiménez, A. S. (2007) La contaminación por ozono tro- posférico. El caso de Motril Granada. Observatorio Mediambiental. Vol 10, pp. 265-280.