SPACE-TIME ANALYSIS OF MEXICO CITY OZONE LEVELS

Gabriel Huerta, Centro de Investigación en Matemáticas, México
Bruno Sansó, Universidad Simón Bolívar, Venezuela
Jonathan R. Stroud, University of Chicago and Argonne National Laboratory, U.S.A.

Jonathan R. Stroud, 5734 S. University Ave., Chicago, IL 60637, U.S.A.
E-mail: stroud@galton.uchicago.edu

Abstract:

We consider hourly readings of ozone concentrations over Mexico City and propose a model for spatial as well as temporal interpolation and prediction. The model is based on regressing the observed readings on a set of meteorological variables, such as temperature and humidity. A few harmonic components are added to account for the main periodicities that ozone presents during a given day. The model incorporates spatial covariance structure for the observations and the parameters that define the harmonic components. Using the Dynamic linear model framework, we show how to compute smoothed means and predictive values. The methodology is illustrated with observations corresponding to September of 1997.
Key Words: Ground-level ozone; Spatio-temporal models; Bayesian inference; Dynamic linear models; state-space models.

1. Introduction

Studying levels of tropospheric ozone is important to understand and improve air quality in major urban areas. Environmental experts and authorities have a special interest in ozone because of its impact on diminishing health, deteriorating materials and damaging vegetation. According to environmental standards, pure air should contain less than 1% compound of ozone and exceedingly levels may produce eye irritation, aggravate respiratory and cardiovascular diseases.

We concentrate on analyzing tropospheric ozone for Mexico City, the most polluted city in the world. Located on the bottom of a valley, with approximately 20 million habitants, Mexico City has maintained high levels of pollution during several years mainly due to huge amounts of motor vehicle and industrial activity. In 1986, the authorities of the city recognized the magnification of the problem and implemented a network of monitoring stations to measure ozone, carbon-Mon oxides and hydrocarbons. The network is named Red Atomática de Monitoreo Ambiental de la Ciudad de México (R.A.M.A.). Currently, the stations that form part of the R.A.M.A. function during the 365 days of the year with occasional interruptions for calibration. Each station takes measures of pollutants automatically, second by second, and the corresponding averages per hour are reported to the public. The units of the measurements are in parts per million, that is, the amount of concentration of the substance in a volume, where the volume is divided into one million parts.

In this paper, we consider the spatio-temporal analysis of ozone time series obtained at some of the stations of the R.A.M.A. Each data point is an hourly concentration of ozone in ppm. We also consider hourly measurements of the meteorological variables, temperature, humidity and wind velocity. Our goal is to propose a statistical model that forecasts temporally, interpolates spatially and show its performance at both levels. We elaborate our models within the Bayesian paradigm using Dynamic Linear Models as in West and Harrison (1997). We strongly believe our modeling approach could assist in the implementation of an environmental contingency strategy.

Previous analyses of ground-level ozone data for multiple sites, modeled jointly, appears in the paper by Carroll *et al.* (1997), which uses a spatially homogeneous and temporally stationary space-time model to study ozone exposure in Texas. Their model includes temperature, wind speed and wind direction as covariates. Also, Guttorp *et al.* (1994) built a space-time model for tropospheric ozone via the spatial deformation method of Sampson and Guttorp (1992), and placed it in a temporal framework by adding a stationary AR process at each station. On the other hand, there is work that considers multiple sites but modeled separately. For example, Rao *et al.* (1997) and Milanchus *et al.* (1998) consider an iterative moving-average filter that decom-

poses ozone into a baseline, trend and a seasonal variation site by site. An extensive and critical review of different approaches of meteorological adjustment and spatio-temporal estimation of ozone are discussed in Thompson *et al.* (1999). Other general approaches of space-time modeling appear in Stroud *et al.* (1999), Sansó and Guenni (2000), Tonellato (1997), Wikle *et al.* (1999), Berliner *et al.* (1999), Mardia *et al.* (1998), among others.

In the next section, we describe the data under study. In Section 3, we consider the periodicities of the ozone series using a standard Bayesian regression tool. In Section 4, we present our space-time model for ozone. In the final section, we present the results based on the model and related discussion.

2. Data description

We consider hourly averages of ozone in ppm measured during 1997 at 19 different monitoring stations scattered irregularly in Mexico City. For 10 of these 19 stations, we also have hourly measurements corresponding to three meteorological variables: temperature (in degrees centigrades), wind speed (meters/second) and relative humidity (in percent). Refer to Figure 1 to locate the stations in a map that includes the metropolitan area of Mexico City. All the data were provided by the Instituto Nacional de Ecología of Mexico.

Hourly ozone time series for the month of September 1997 appear in Figure 2. The series correspond to 5 monitoring stations. The first two frames are for stations nearby the downtown area: Merced (MER) and Hangares (HAN). The middle frame is for station Benito Juarez (BJU), which is located close to the center of the map of Figure 1. The last two frames correspond to stations in the south side of the city, Pedregal (PED) and Tlalpan (TLA). In general, we notice a diurnal cycle of ozone and usually a very high peak during the early afternoon hours, between 1 pm and 4 pm. This high peak is associated to the daily maximum temperature and the motor-vehicular activity in the city during the morning and early afternoon hours. Also, there is a smaller but frequent nocturnal peak. Parts of the series are missing and usually correspond to late evening-early morning hours. We do not notice any obvious weekly patterns or weekend effects but there are changes from one day to the other that suggest that, even after considering daily cycles, there is lack of stationarity in the series. Additionally, the figure shows that the spatial pattern of ozone in Mexico City is complex. Notice that the levels of Merced are consistently lower than those of Hangares, which

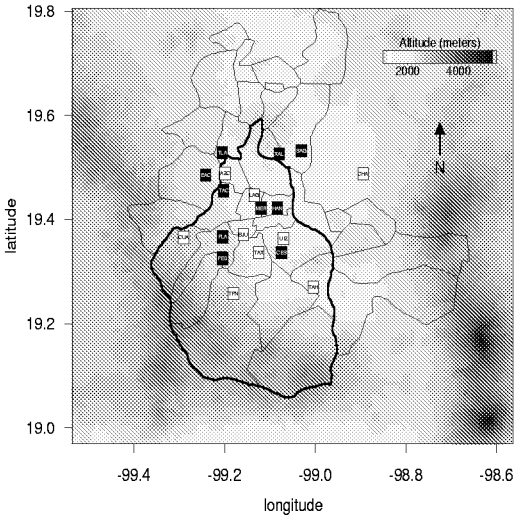

Figure 1: Locations of 19 Monitoring Stations. Black squares represent the 10 stations considered in the present analysis, while white squares represent additional stations.

is spatially close, during the whole month.

Figure 3 summarizes the data of ozone with plots of hourly medians over 1997 for each of the 19 monitoring stations. Each frame roughly represents the Northwest (NW), Northeast (NE), Southwest (SW) and Southeast (SE) of Mexico City. The hourly medians exhibit the daily cycle of ozone and clearly show that the high peak is reached at different hours of the afternoon across stations. Particularly, for the stations grouped within the NW, NE and SW, the variability of median level across nearby stations is important. In a similar display, Figure 4 presents the hourly medians over 1997 for temperature (top frames) and relative humidity (bottom frames) for the 10 stations where meteorology is available. The stations were grouped into two subregions, one that represents the stations closer to the interior of the valley that surrounds Mexico City (right panels) and the other, that represents the stations closer to the mountain side (left panels). Both median temperature and median relative humidity have daily cycles, are negatively correlated and exhibit less spatial variability compared to median ozone levels.

As is usual for ozone measurements, the distribution of the data has an asymmetric shape that suggests the use of a transformation to justify the use of models based on the normal distribution. The two most common transformations in the literature for

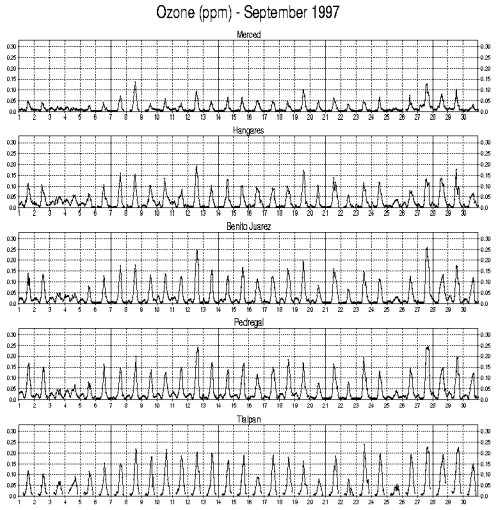

Figure 2: Hourly ozone levels for five monitoring stations corresponding to September 1997

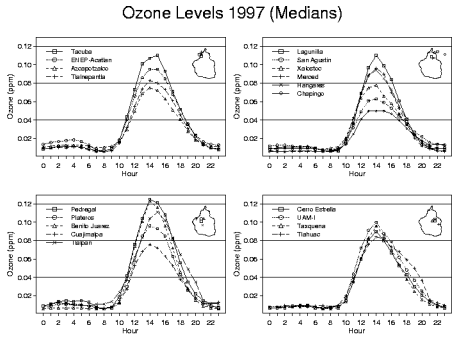

Figure 3: Median ozone levels per hour for 1997.

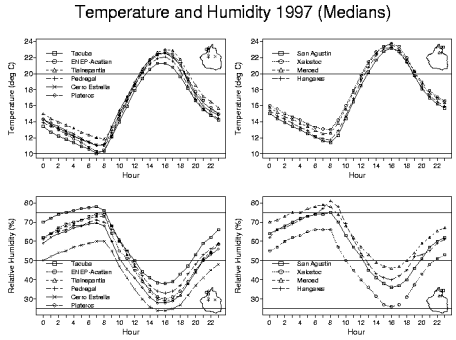

Figure 4: Median temperatures (top frames) and relative humidities (bottom frames) per hour for 1997.

ozone data are the square root and the natural logarithm. Thompson *et al.* (1999) report a summary of the transformations used by different authors that analyze ozone series. In this paper, we consider a log transformation for the data. This is supported on analyses of the distribution of the observed values as well as the behavior of the residuals of the models that we propose in the following sections.

3. Inference on Periodicities

As can be seen from the descriptive analysis of the data, an accurate specification of the cyclical behavior is a key feature for modeling. A statistical approach to make inferences on periodicities is the *Bayesian periodogram* introduced by Bretthorst (1988). The Bayesian periodogram is defined as the marginal log-likelihood of the regression model

$$Y_t = a\ \cos(2\pi t/\lambda) + b\ \sin(2\pi t/\lambda) + \epsilon_t,$$

marginalized with respect to the reference prior $p(a, b, \sigma^2) \propto 1/\sigma^2$, where t indexes time, $\epsilon_t \sim N(0, \sigma^2)$ and λ is the underlying periodicity or wavelength.

Figure 5 shows the Bayesian periodograms for the ozone time series of September 1997 and measured at the monitoring stations that have meteorology. The figure has a common range of values for λ, between 0 and 50 time units. We observe that the general pattern of all the periodograms is similar and that the data has distinctive cycles with wavelengths of 12 and 24 hours. Some of the stations have a smaller peak at eight hours. We also evaluated the Bayesian periodogram for values of λ greater than 50 and we could not find any other relevant peaks.

4. Space-Time Model

Let Y_{it} denote the observed log ozone concentration, for each station $i = 1, \ldots, S$ and time $t = 1, \ldots, T$ and let X_{ijt} be the j-th covariate at time t and station i. Then define $Z_{it} = (1, X_{i1t}, \ldots, X_{rit})'$ and let β_t be the corresponding $(r+1)$-dimensional vector of covariate coefficients. Z_{it} may include meteorology and a spatial trend can be modeled by making some terms functions of the coordinates of the stations, for example, a first or a second order polynomial on the coordinates of latitude and longitude.

Let α_{it} denote the q-dimensional vector of seasonal coefficients for station i corresponding to a seasonal component S_t' consisting of sine and cosine terms. The specification of the periodicities of

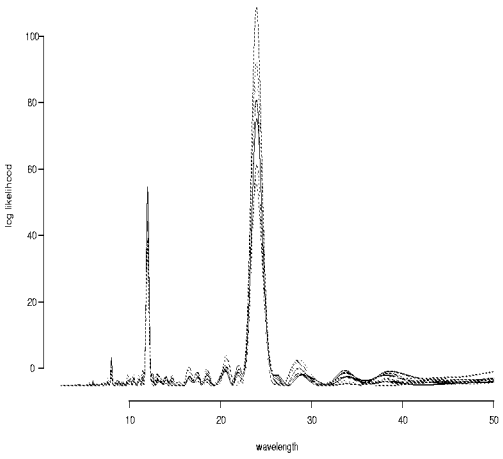

Figure 5: Bayesian periodogram of ozone concentrations at ten monitoring stations in Mexico City, during the first two weeks of September 1997.

the seasonal component may follow a Bayesian periodogram analysis. For instance, the results of the previous section lead to a seasonal term

$$S_t' = \left(\cos(\frac{\pi t}{12}), \sin(\frac{\pi t}{12}), \cos(\frac{\pi t}{6}), \sin(\frac{\pi t}{6})\right)$$

which implies $q = 4$.

The general space-time model that we propose for ozone data is given by

$$Y_{it} = Z_{it}'\beta_t + S_t'\alpha_{it} + \varepsilon_{it}$$

where the observation errors, ε_{it}, are spatially correlated with Gaussian distribution and a covariance matrix of the form $\sigma^2 V_1$. Note that the coefficients related to the covariates are assumed equal for all stations, while we are assuming that each station has its own set of seasonal parameters. This is justified by the results obtained with a simplified version of the model that was fitted to the data station by station . This general space-time model clearly has a very high number of parameters, since at each time t there are q parameters for each station, r common ones plus the parameters that define the spatial correlation.

A substantial reduction in the number of parameters is achieved by assuming that the amplitudes of each cyclical component are different, but the phases are very similar between stations and almost constant in time. This assumption is supported by a univariate models fitted station by station.

We can thus consider a modification of the model given by

$$Y_{it} = Z_{it}'\beta_t + S_t'\alpha_{it} + \varepsilon_{it}$$

but now α_{it} is a vector of dimension $q/2$ and $S_t' = \left(\cos(\frac{\pi t}{12}) + a_1 \ \sin(\frac{\pi t}{12}), \ldots \cos(\frac{\pi q t}{12}) + a_{q/2} \ \sin(\frac{\pi q t}{12})\right)$.

Thus $\alpha_{jit}^2(1 + a_j^2)$ is the amplitude of the j-th periodicity of the i-th station at time t, and $\tan^{-1}(a_j)$ is its amplitude. At the current stage of modeling, we are using temperature, humidity, wind-speed anomalies and a second order polynomial on latitude and longitude as the defining covariates for Z_{it}'.

Additionally, the parameters in the model evolve in time according to random-walk evolutions, i.e.,

$$\beta_t = \beta_{t-1} + \omega_{1,t}; \ \omega_{1,t} \sim N(0, \mathbf{W}_{1,t}),$$
$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} + \omega_{2,t}; \ \omega_{2,t} \sim N(0, \mathbf{W}_{2,t})$$

where the vector $\boldsymbol{\alpha}_t$ is the concatenation of the vectors $(\alpha_{j1t}, \alpha_{j2t}, \ldots, \alpha_{jSt})$; $j = 1, \ldots, q$, which are the parameters corresponding to the j-th periodicity.

Furthermore, at the observation level, we assume that the covariance matrix of the errors has the form $\sigma^2 exp(-D/\lambda_\epsilon)$, or $V_1 = exp(-D/\lambda_\epsilon)$, where D is the matrix of euclidean distances between monitoring stations. $\mathbf{W}_{1,t}$ is specified with a *discount factor* approach and $\mathbf{W}_{2,t}$ as a block diagonal matrix with blocks of the form $\tau_j exp(-D/\lambda_j); j = 1, \ldots, q$.

This spatio-temporal model can be easily written in the state-space form notation of West and Harrison (1997). Thus, conditional on the hyperparameters that define the covariance structure, the filtering and recurrence equations of the DLM produces predictive values and restropective inferences for observed values. Formal Bayesian inference on the hyperparameters leads to the *Forward Filtering Backward Simulation* algorithm which is computationally very intensive for a high-dimensional state vector. Alternatively, the hyperparameters may be estimated via *Empirical Bayes* by maximizing the marginal log-likelihood over an extensive grid of values. Then, the standard filtering and recurrence equations may be applied to update the state-vector of the DLM conditional on these MLE estimates.

5. Results and Discussion

The space time model was used to study the data of the first two weeks of September 1997. The empirical Bayes estimate for σ^2 is 0.08 and for λ_ϵ is 0.001. The covariance for the regression evolution was specified with three discount factors $(\delta_1, \delta_2, \delta_3)$. δ_1 corresponds to the terms of the second order polynomial in latitude and longitude, δ_2 is related to the covariates temperature and relative humidity, while δ_3 is

related to the wind-speed anomalies. The empirical Bayes estimates of the discount factors are 0.825, 0.875 and 0.95 respectively. Additionally, the empirical Bayes estimates for τ_1 and τ_2 are both equal to 0.00015. For λ_1 and λ_2 the estimates are both 0.1.

Under these specifications, we produce Figure 6 which has information at two levels for 5 monitoring stations. From September 1 until the afternoon of September 13, filtered means and 95% probability bands (solid lines) are plotted with the observations (black circles). Furthermore, from the evening hours of September 13 until the end of September 15, we present forecast means with the 95% predictive probability intervals (solid lines) and the actual observed values of ozone (white circles). Notice that in the restrospective sense, the model represents the cyclical patterns and non-stationarities of the data adequately. On the other hand, the predictive intervals become explosive as time progresses, so the model is only useful for short-term forecasting.

Maps of hourly ozone levels for September 2, 1997 appear in Figure 7. The region of inference is the convex hull of the points that represent the stations where we have ozone data. At each hour, unknown values of the covariates were taken as averages of all the known values at other stations. Thus, the DLM was fitted with all the 19 stations and the hourly filtered means for September 2 smoothed using the *Splus* function *interp*. The resulting map seems to be consistent with the cyclical behavior of the data and theories about the dispersion of ozone in Mexico City. The pollutant builds at around 10-11 a.m., the peak hours are between 2 and 3 p.m. The levels decrease at about 5-6 p.m. We detected some peculiar boundary effect at late night-early morning hours like 4 a.m. This is due to a combination of missing information at those hours in distant locations and the use of a second order polynomial as a spatial mean function.

Further developments consist in formulating a model that will produce a complete Bayesian analysis on the hyperparameters and formally interpolates the meteorology.

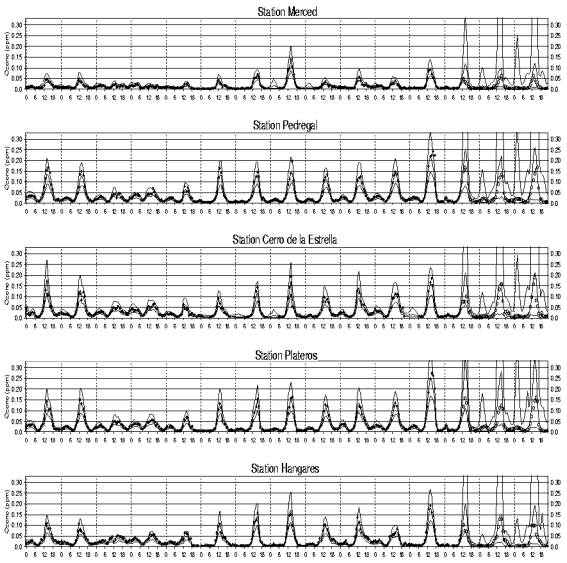

Figure 6: Data, retrospective means and predictive values with 95% probability bands for 5 monitoring stations.

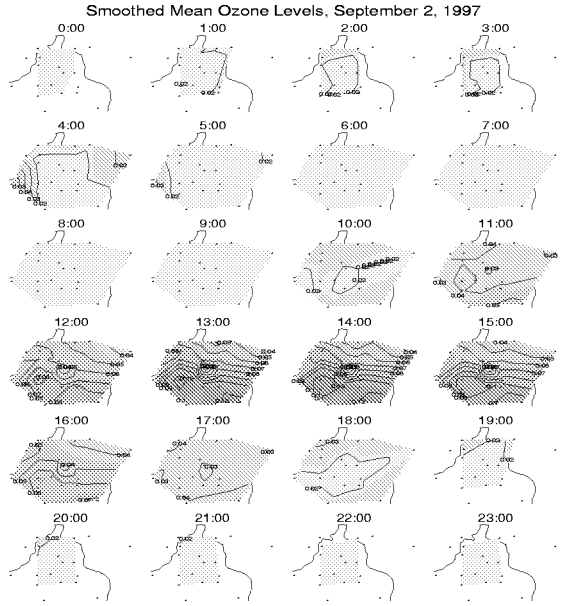

Figure 7: Hourly smoothed means of ozone levels for September 2, 1997.

Bibliography

Berliner, L.M., Royle, A.J., Wikle, C.K. and Milliff, R.F. (1999) Bayesian methods in the atmospheric sciences. In *Bayesian Statistics 6, Oxford* (eds J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M Smith), pp. 83–100. Oxford University Press.

Bretthorst, L. (1988) *Bayesian Spectral Analysis and Estimation.* New York: Springer Verlag.

Carroll, R., Chen, R., George, E.I., Li, T.H., Newton, H.J., Schmiediche, H. and Wang, N. (1997) Ozone exposure and population density in Harris County, Texas. *J. Am. Statist. Ass.*, **92,** no. 438, 392–415.

Guttorp, P., Meiring, W. and Sampson, P.D. (1994) A space-time analysis of ground-level ozone data. *Environmetrics*, **5,** 241–254.

Mardia, D.V., Goodall, C., Redfern, E. and Alonso, F.J. (1998) The kriged kalman filter. *Test*, **7,** no. 2, 217–285.

Milanchus, M.L., Rao, T.S. and Zurbenko, I.G. (1998) Evaluating the effectiveness of ozone management efforts in the presence of meteorological variability. *Journal of the Air and Waste Management Association*, **48,** 201,215.

Rao, S.T., Zurbenko, I.G., Neagu, R., Porter, P.S., Ku, J.Y. and Henry, R.F. (1997) Space and time scales in ambient ozone data. *Bulletin of the American Meteorlogical Society*, **78,** 2153–2166.

Sampson, P. and Guttorp, P. (1992) Nonparametric estimation of nonstationary spatial covariance structure. *J. Am. Statist. Ass.*, **87,** 108–119.

Sansó, B. and Guenni, L. (2000) A non-stationary multisite model for rainfall. *J. Am. Statist. Ass.*, **95,** no. 452.

Stroud, J., Müller, P. and Sansó, B. (1999) Dynamic models for spatio-temporal data. Technical Report Working Paper 99-20. Institute of Statistics and Decision Sciences, Duke Unversity, Durham, North Carolina.

Thompson, M.L., Reynolds, J., Cox, L.H. and Guttorp, P. (1999) A review of statistical methods for the meteorological adjustment fo tropospheric ozone. Technical Report. N.R.C.S.E, University of Washington and E.P.A.

Tonellato, S. (1997) Bayesian dynamic linear models for spatial time series. Technical Report Rapporto di ricerca 5/1997. Dipartimento di Statistica – Università Ca' Foscari di Venezia, Venice, Italy.

West, M. and Harrison, P.J. (1997) *Bayesian Forecasting and Dynamic Models*, 2nd edn. New York: Springer.

Wikle, C., Berliner, M. and Cressie, N. (1999) Hierarchical Bayesian space-time models. *J. Env. Ecol. Statist.*, **5,** 117–154.