Time series forecasting of ozone levels in the Metropolitan Area of Monterrey, Mexico

S Iglesias-González 1, M Huertas1, I Hernández-Paniagua1,2 and A Mendoza4

1 Tecnologico de Monterrey, Escuela de Ingeniería y Ciencias, Av. Eugenio Garza Sada 2501, Monterrey, N.L., 64849, México
2 CONACYT-Consorcio CENTROMET, Camino a Los Olvera 44, Los Olvera, Corregidora, Querétaro, 76904, México

Email: mendoza.alberto@tec.mx (Corresponding author)

Abstract. Based on the limitations of air quality models to forecast air pollution, statistical models are convenient and suitable tools to predict pollutant concentrations. This research work proposes a SARMA model to forecast ozone maxima concentrations in five sites of the Metropolitan Area of Monterrey, Mexico. The design of the model includes diverse novel features: meteorological variables were considered as predictors, short-term ozone concentrations (4 times in the day) were forecasted, and appropriated transformations of meteorological parameters were included. The performance measures applied to assess the SARMA model demonstrated that this statistical model is reliable to predict short-term ozone concentrations and it is consistent with the general dynamics of the ozone formation in the metropolitan area.

Keyword: Seasonal ARMA, statistical models, meteorological variables, air quality.

1. Introduction

Forecasting is an integral and useful task to manage the urban air quality. Statistical and mathematical simplified models are common techniques use to forecast spatio-temporal pollutant concentrations. Artificial Neural Networks (ANN) have widespread acceptance to predict ozone (O₃) concentrations but are popular in applications where model interpretation has a secondary importance [1]. Other mathematical approaches to forecast ozone (O₃) are Fuzzy Time Series (FTS) [2] [3] and additive models [4]. The Autoregressive Integrated Moving Average (ARIMA) model [5] is a classical forecast technique used to analyze linear time series data. Diverse studies [6]-[8] have found that both ARIMA and ARMA models are reliable and capable of predicting short-term O₃ concentrations.

Tropospheric ozone (O₃) is a greenhouse gas and a criteria air pollutant that frequently exceeds the air quality standards in urban areas around the world [9]. This is a pollutant formed in the atmosphere by chemical reactions of volatile organic compounds (VOCs) and nitrogen oxides (NOx) in the presence of sunlight.

In order to reduce the exposure of urban population to pollution episodes, environmental authorities of the Mexican northern state of Nuevo Leon require a forecasting model to predict O₃ concentrations in the Monterrey Metropolitan Area (MMA). In response to this need, this paper describes the applications of a general Multiple Regression with Seasonal Autoregressive Moving Average (SARMA)
Errors model to estimate and forecast maximum O$_3$ concentrations, with a number of design and operational novel features.

2. **Material and methods**

The MMA is located in the Nuevo Leon State, at north-east of Mexico surrounded by mountains to the south and west, and flat terrain to the north-east, with an average altitude of 500 m.a.s.l. (Figure 1).

**Figure 1.** Monterrey Metropolitan Area and the selected monitoring sites.

2.1. **Monitoring Network and Data**

Within the MMA, the Environmental Monitoring Integrated System (SIMA) has thirteen air quality stations and is operated since November 1992. Ozone concentrations and seven meteorological variables (temperature, solar radiation, relative humidity, rainfall, pressure, wind speed and wind direction) are recorded every minute and summarized as hourly average. Only five sites (OBI, GPE, SNB, SNN, STA) were selected considering their data availability and long-term records, they are described in Table 1.

**Table 1.** Description of monitoring stations selected in the MMA.

| Site          | Code | Location         | Description                                                                 |
|---------------|------|------------------|-----------------------------------------------------------------------------|
| Guadalupe     | GPE  | 25°40.110’ N 100°14.907’ W | Urban background site in the La Pastora Park                                |
| San Nicolás   | SNN  | 25°44.727’ N 100°15.301’ W | Site surrounded by a large number of industries                            |
| Obispado      | OBI  | 25°40.561’ N 100°20.314’ W | Urban site near the city center of MMA                                       |
| Santa Catarina| STA  | 25°40.542’ N 100°27.901’ W | Urban site in a residential area                                            |
| San Bernabé   | SNB  | 25°45.415’ N 100°21.949’ W | Urban site downwind of industrial sources                                   |

The meteorological and air pollutant time-series data were analyzed and filtered prior to modeling O$_3$ concentrations. Six–hour maximum O$_3$ concentrations were predicted at 6:00, 12:00, 18:00 and 00:00 during daylight saving time (DST).

2.2. **General Statistical Approach**

The modeling approach employs a linear regression model to predict ground maximum O$_3$ levels, making use of meteorological variables as O$_3$ predictors. The residual part of the model is then modeled
through the family of Seasonal ARMA models which allows to understand the temporal structure of the series and the prediction of the series from past values. Six meteorological variables appropriately transformed were used as predictors: temperature (Temp), relative humidity (HR), solar radiation (SR), wind direction (WD) and wind speed (WS). The transformations were computed in order to linearize the relation with maximum O₃ levels.

The Box-Jenkins approach [5] was applied for building a time series model, whose general steps are summarized as follows:

a. Calculation of residuals from a multiple regression model.

b. Identification of a SARMA model.

c. Model fitting: Estimation of meteorological effects and SARMA parameters.

d. Model diagnostic.

e. Model selection.

2.3. The SARMA Model

The multiple regression model with multiplicative seasonal ARMA \((p, q) (P, Q)_s\) errors was formulated as

\[
\phi_p(B) \Phi_P(B^s) (y_t - x_t \beta) = \theta_q(B) \Theta_Q(B^s) \alpha_t \quad (1)
\]

where \(y_t\) and \(x_t\) represents the O₃ maxima and associated weighted average values of meteorological variables at time of the day \(t\), respectively; \(\phi_p(B) = 1 - \phi_1 B - \ldots - \phi_p B^p\), with \(B^p \equiv y_{t-p}\), being autoregressive operator of order \(p\) and \(\theta_q(B) = 1 - \theta_1 B - \ldots - \theta_q B^q\) is the moving average operator for the regular part, respectively; \(\Phi_P(B^s) = 1 - \Phi_1 B^s - \ldots - \Phi_P B^{ps}\) is the autoregressive operator of order \(P\), with being the number of periods per season, and \(\Theta_Q(B^s) = 1 - \Theta_1 B^s - \ldots - \Theta_Q B^{qs}\) is the moving average operator of order \(Q\) for the seasonal part, respectively; \(\alpha_t\) is usually assumed to be normal white noise with variance \(\sigma^2\). For the series being modeled, \(s\) is expected to be equal to four, i.e., a daily cycle effect. The effect of meteorological variables upon the maximum ozone concentration is represented by \(\beta\).

3. Results and discussion

3.1. Model Estimation

A single generic model by a SARMA \((3, 0) (3, 0)_4\) for each site was found to represent the maximum O₃ time series. Parameter estimates of the fitted models are given in Table 2. Results shows significant increasing effects mainly during springtime (March, April and May), while December presents decreasing deviations in most sites, this is consistent with the O₃ dynamics in the Monterey Metropolitan Area [10]. No indication of lack of stationarity is evident from the (P)ACF plots. Figure 2 shows the empirical (P)ACF computed from the historical observations, and the simulated (P)ACF computed from the generic model for the OBI site. The plot displays a typical exponential decay of ACF for both the regular and the seasonal component together with the corresponding cutoffs of PACF at first few lags.
Table 2. ARMA and Regression parameter estimates. Significance levels: 0.05 (*); 0.01 (**)  

|              | OBI  | GPE  | SNB  | SNN  | STA  |
|--------------|------|------|------|------|------|
| $\varphi_1$ | 0.320** | 0.308** | 0.338** | 0.354** | 0.350** |
| $\varphi_2$ | -0.002 | -0.012 | -0.015 | -0.018 | -0.032 |
| $\varphi_3$ | 0.069** | 0.093** | 0.044* | 0.092** | 0.045* |
| $\varphi_4$ | 0.244** | 0.219** | 0.205** | 0.226** | 0.193** |
| $\varphi_5$ | 0.035** | 0.070** | 0.034* | 0.064** | 0.067** |
| $\Phi_1$    | 0.053** | 0.057** | 0.066** | 0.090** | 0.028 |
| Intercept   | -24.408** | -17.008** | -20.564** | -8.427* | -36.244** |

February: 0.068 0.067 0.056 0.154 -0.003  
March: 0.182** 0.189** 0.148** 0.225** 0.081  
April: 0.289** 0.206** 0.223** 0.231** 0.136*  
May: 0.285** 0.181** 0.207** 0.241** 0.240**  
June: 0.087 -0.029 -0.053 -0.030 -0.013  
July: 0.024 -0.087 -0.059 0.014 -0.045  
August: -0.008** -0.076 -0.118 0.032 -0.072  
September: 0.232** 0.082 -0.079 0.028 0.055  
October: 0.243** 0.163** -0.025 -0.062 -0.022  
November: 0.094* 0.061 0.019 -0.024 0.25  
December: -0.169** -0.151** -0.060 -0.214** -0.138*  
06 to 11: -0.263** 0.024 0.155** 0.024 -0.057  
12 to 17: -0.014 0.165** 0.270** 0.155** 0.135**  
18 to 23: -0.044* 0.031 0.012 -0.057* -0.069**  

ln(Temp): 4.886** 3.629** 4.233** 2.080** 6.987**  
HR: 0.027** 0.014** 0.015** 0.020** 0.019**  
HR$^2$: -0.001** 0.000** 0.000** 0.000** 0.000**  
HR$^3$: 0.000** 0.000 0.000 0.000 0.000  
1/(WS+1): -2.980** -1.904** -1.932** -1.792** -2.409**  
SR: 1.029** 0.369** 0.819** 1.031** 0.788**  
SR$^2$: -0.548** -0.094 -0.408* -0.603** -0.355*  
WD: -0.003** -0.002** -0.002** -0.002** -0.002**  

Figure 2. Empirical and simulated (P)ACF for OBI site.
3.2. Model Diagnostics

The model is assessed by the usual Ljung–Box tests and ACF of residuals as are shown in Figure 3. Larger lags show significant ACF, indicating model inadequacy; however, ACF magnitudes at those lags are seen to be small, representing a marginal 5–7 former days effects on O$_3$ concentrations. The model accounts well for immediate previous times of the day and daily cycles effects.

![Figure 3. Residuals ACF and Ljung – Box tests.](image)

3.3. Model Performance

Table 3 shows the values of the performance measures used to assess the forecast models. Performance measures close to zero indicate a good forecasting. RMSE and MAE ranges from 9.58 to 10.89 and 6.87 to 8.15, respectively. These results are similar to the values report in other research works that apply ARIMA models to forecast O$_3$ levels [6] [11]. The MAPE tends to exhibit the largest values ranging from 25.65 to 32.31. The model shows negative values in all sites for MPE which indicates an overestimation of the maximum O$_3$ concentrations. The SNN site shows the lowest performance measures, suggesting that in this site the model performs better, however the differences between sites are not great. In general, the model performs similarly in all the sites, especially considering the MAE and RMSE.

Table 3. Performance measures for the models at each site.

| Performance Measure | Site |
|---------------------|------|
|                     | OBI  | GPE  | SNB  | SNN  | STA  |
| RMSE                | 10.1 | 10.1 | 10.89| 9.58 | 9.95 |
| MAE                 | 8.06 | 7.85 | 8.15 | 6.87 | 7.26 |
| MAPE                | 32.31| 27.65| 29.96| 25.65| 28.87|
| MPE                 | -9.29| -10.64| -8.53| -6.13| -7.28|
4. Conclusions
A multiple regression model with seasonal ARMA errors and meteorological parameters as predictors has shown to be adequate for forecasting the O₃ maxima at 4 times of the day (overnight, morning, afternoon and evening) in five sites of the Monterrey Metropolitan Area monitoring network in Mexico.

An important contribution of this research work is the feature of implement the meteorological variables as predictors and include the correct transformation of meteorological parameters in an ARMA model. This last characteristic allows to interpret the influences of each meteorological parameter in the formation of tropospheric O₃.

Acknowledgement
We thank to the Secretariat for Sustainable Development of the Nuevo León State for the public domain records.

References
[1] Chaloulakou A, Saisana M, and Spyrellis N. 2003. Comparative assessment of neural net-works and regression models for forecasting summertime ozone in Athens. Science of the Total Environment, 313(1-3):1–13, 2003.
[2] Dománska D. and Wojtylak M 2012 Application of fuzzy time series models for forecasting pollution concentrations. Expert Systems with Applications, 39(9):7673–7679.
[3] Dincer N and Akkus O 2018 A new fuzzy time series model based on robust clustering for forecasting of air pollution. Ecological Informatics, 43:157–164.
[4] Blanchard C, Hidy G, and Tanenbaum S 2014 Ozone in the southeastern United States: An observation-based model using measurements from the SEARCH network. Atmospheric Environment, 88:192–200.
[5] Box G, Jenkins G, Reinsel G, and Ljung G 2015 Time series analysis: forecasting and control. John Wiley & Sons.
[6] Kumar K, Yadav A, Singh M, Hassan H, and Jain V 2004 Forecasting daily maximum surface ozone concentrations in brunei darussalam—an arima modeling approach. Journal of the Air & Waste Management Association, 54(7):809–814.
[7] Beldjillali H, Lamri N, and Bachari N 2016 Prediction of ozone concentrations according the box-jenkins methodology for assekrem area. Applied Ecology and Environmental Sciences, 4(2):48–52.
[8] Rozita W, Mahiyuddin W, Jamil N, Seman Z, Ahmad N, Abdullah N, Latif M, and Sahani M 2018 Forecasting Ozone Concentrations Using Box-Jenkins ARIMA Modeling in Malaysia. American Journal of Environmental Sciences, 14(3):118–128.
[9] EEA. 2017 Outdoor air quality in urban areas. https://www.eea.europa.eu/airs/2017/environment-and-health/outdoor-air-quality-urban-areas. Acceded 12 January 2018.
[10] Hernandez-Paniagua I, Clemitshaw K, and Mendoza A 2017 Observed trends in ground-level O₃ in Monterrey, Mexico, during 1993-2014: comparison with Mexico City and Guadalajara. Atmospheric Chemistry & Physics, 17(14).
[11] Kumar U and Jain V 2010 ARIMA forecasting of ambient air pollutants (O₃, NO, NO₂ and CO). Stochastic Environmental Research and Risk Assessment, 24(5):751–760.