Predicting Hourly Ozone Pollution in Dallas-Fort Worth Area Using Spatio-Temporal Clustering

Mahdi Ahmadi[1], Yan Huang[2], Kuruvilla John[3]

[1]University of North Texas, 1155 Union Circle #311098
Denton, TX 76203-5017 USA
Telephone: 940-565-2400
Email: MahdiAhmadi@unt.edu

[2]University of North Texas, 1155 Union Circle #311098
Denton, TX 76203-5017 USA
Telephone: 940-369-8353
Email: Yan.Huang@unt.edu

[2]University of North Texas, 1155 Union Circle #311098
Denton, TX 76203-5017 USA
Telephone: 940-565-4302
Email: Kuruvilla.John@unt.edu

Abstract

Ground level ozone is one of the air pollutants with negative impacts on human health and environment. The complexity of physical process of ozone formation makes it difficult to predict ozone concentration accurately. In this work clustering techniques and multiple regression analysis are used to construct a simply interpretable forecasting model. Time series of ozone and meteorological variables in Dallas-Fort Worth area for 12 years at 14 monitoring stations are acquired and processed. First, k-means cluster analysis is performed on ozone time series to specify data driven ozone seasons at each station. Next, spatial hierarchical clustering is performed to find ozone zones in the area during each ozone season recognized in the previous step. Finally, multiple linear regression between meteorological variables and ozone in each zone is developed. For liner forecasting temperature, solar radiation, wind speed, and previous ozone values are used since ozone is strongly auto-correlated. Monitoring stations in each temporal and spatial cluster show consistent behavior which makes it possible to perform ozone forecasting even when a station is off. Results show high accuracy of ozone forecasting and yet simple to interpret the link between meteorology and ozone behavior. Also, clustering results are useful to understand the temporal and spatial patterns of the ozone dynamics in the area.

1. Introduction

Ozone is a highly reactive chemical species with proven negative impacts on humans. Ground level ozone pollution is formed by a chain of photochemical reactions in the presence of nitrogen oxides (NO_X) and reactive volatile organic compounds (VOCs). Tropospheric ozone formation is a complex process (Seinfeld and Pandis, 2012) and displays strong seasonal and diurnal patterns with higher concentration during summer and in the afternoon. Ozone prediction is a difficult task due to the complexity of the formation process and spatio-temporal variation in both meteorological factors and precursors.

Over the decades, several statistical techniques have been developed to account for the effect of meteorological factors and predict ozone (Lou Thompson et al., 2001, Schlink et al., 2003). The simplest model is multiple linear regression that assumes an additive linear relationship to link ozone concentration to meteorological factors (Feister and Balzer, 1991, Korsog and Wolff, 1991, Dueñas et al., 2002, Fiore et al., 1998, Walker, 1985, Kuntasal and Chang, 1987, Zeldin et al., 1990, Cassmassi and Bassett, 1993, Abdul-Wahab et al., 1996, Katsoulis, 1996). Although linear models are easy to interpret but they are generally poor in accuracy. Also methods such as principal component analysis (PCA), artificial neural networks and clustering techniques are proposed to reduce the dimensionality of the problem to make prediction and interpretation easier (Al-Alawi et al., 2008, Sousa et al., 2007, Lengyel et al., 2004, Guardani et al., 2003, Kovač-Andrić et al., 2009, Kaburlasos et al., 2007, Bruno et al., 2004, Sahu et al., 2007, Sahu and Bakar, 2012, Austin et al., 2014). Ozone forecasting can be performed more effectively once the temporal and spatial patterns are quantified. In this paper clustering techniques and multiple linear regression analysis are used to explain the patterns of ozone in Dallas-Fort Worth (DFW) area and also to develop a linear model for ozone prediction.

2. Method

The objective of this work is to perform clustering of ozone time series and spatial clustering of ozone monitors to produce better input for linear regression analysis. The flow chart of the data mining tasks is presented in fig 1. Accordingly, following procedure is performed:

Step I: Acquiring dataset

Step II: Pre-processing data

Step III: performing *K*-means cluster analysis on time series of 8-hr average daily maximum ozone. The goal is to recognize ozone seasons.

Step IV: evaluating the overlap between ozone seasons at different monitoring stations to determine the best split for the entire monitoring network (i.e. regional ozone seasons)

Step V: performing hierarchical cluster analysis on 1-hr ozone time series of all monitoring stations for each ozone season. The goal is to determine the best spatial clustering (ozone zones) of in each ozone season.

Step VI: developing multiple linear regression model for each ozone zone in each season.

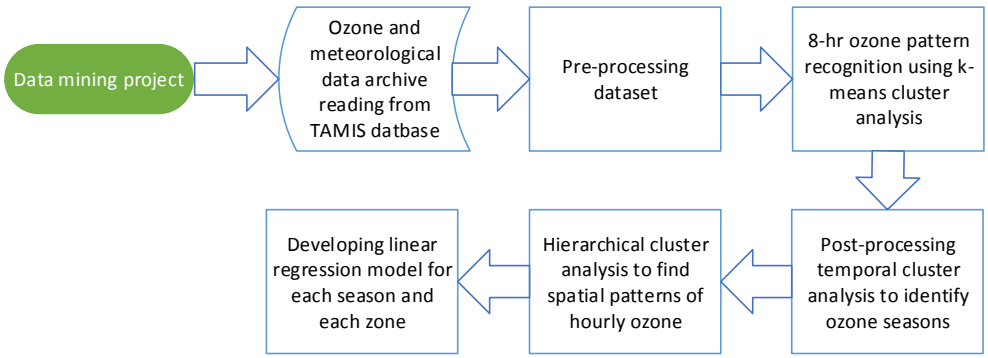


Figure 1. Flow chart of the data mining tasks

3. Dataset

Dallas-Fort Worth (DFW) Metroplex is chosen as the study area. Measurement data collected by TCEQ (Texas Commission on Environmental Quality) CAMS (Continuous Air Monitoring Stations) were used for entry to dataset. The map of CAMS in DFW area is shown in fig. 2. The dataset includes 1-hr measurement time series of ozone (O_3), ambient temperature (T), solar radiation (SR), wind speed (W) for 12 years (2002-2013). The dataset approximately includes 5886720 total entries. Time series of variables at C13 CAMS are presented in figs 2 to 7.

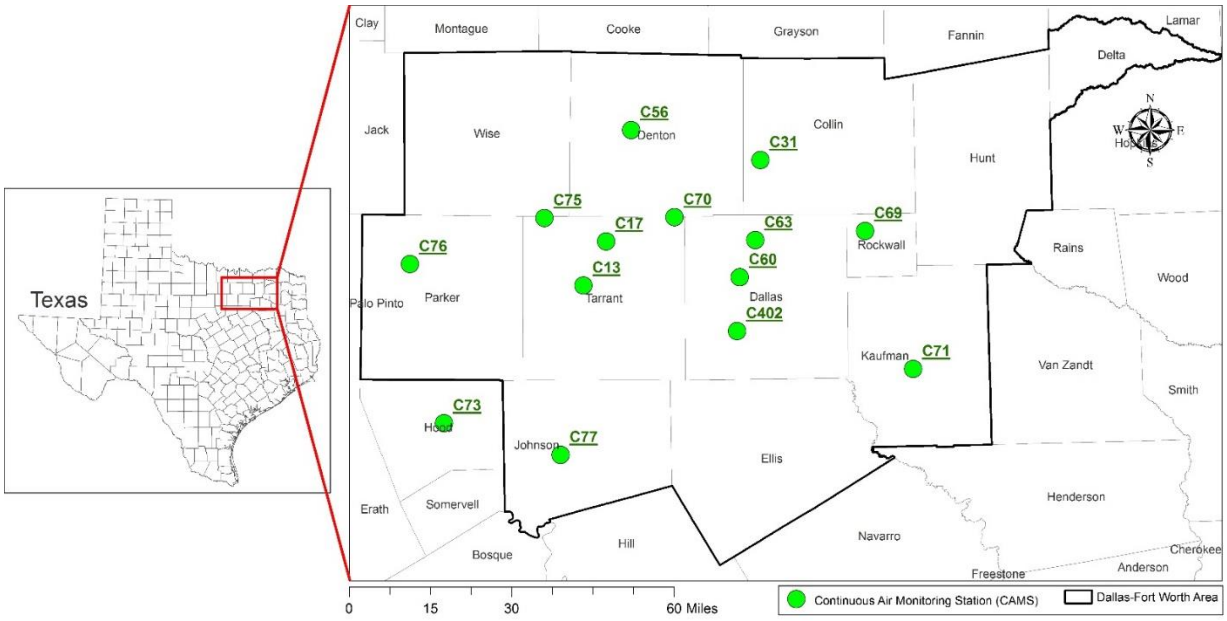


Figure 2. Map of the study area and locations of Continuous Air Monitoring Stations (CAMS)

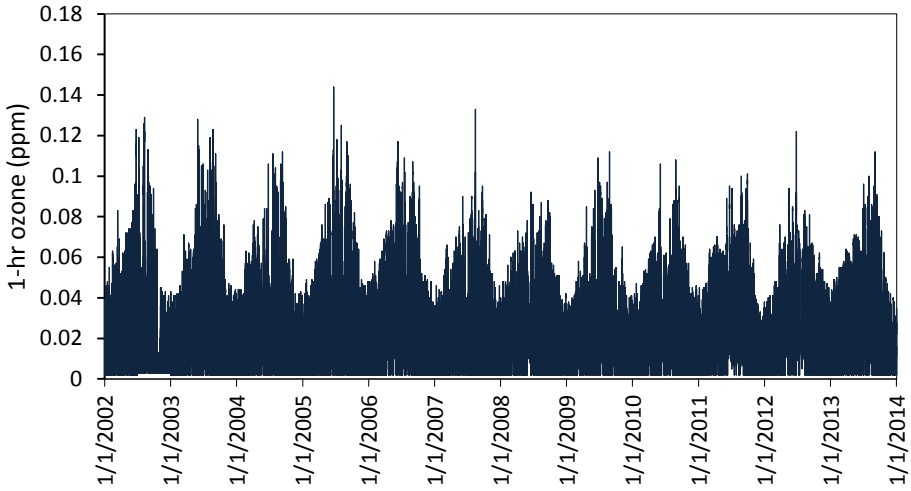


Figure 3. Time series of 1-hr ozone measured at C13 CAMS

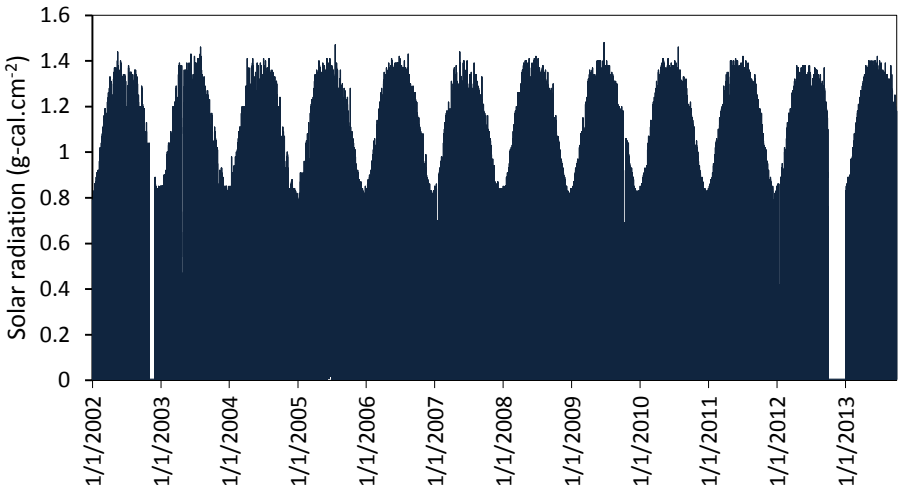


Figure 4. Time series of 1-hr solar radiation measured at C13 CAMS

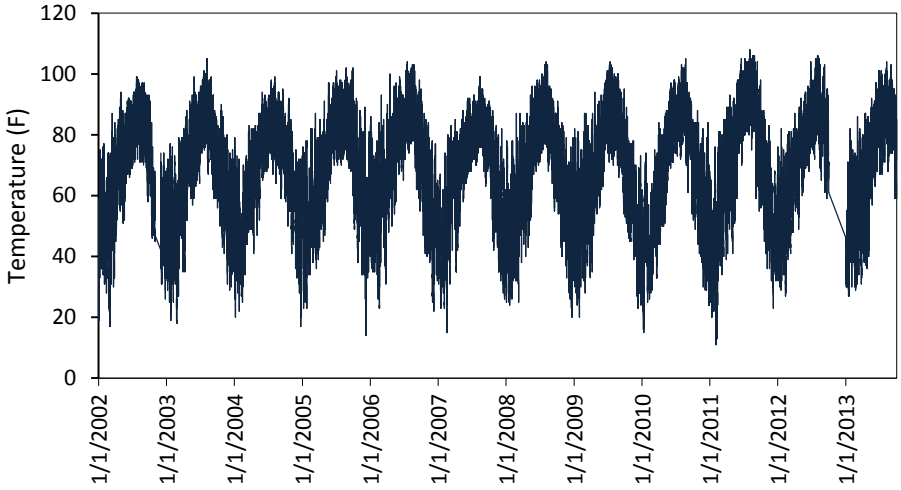


Figure 5. Time series of 1-hr temperature measured at C13 CAMS

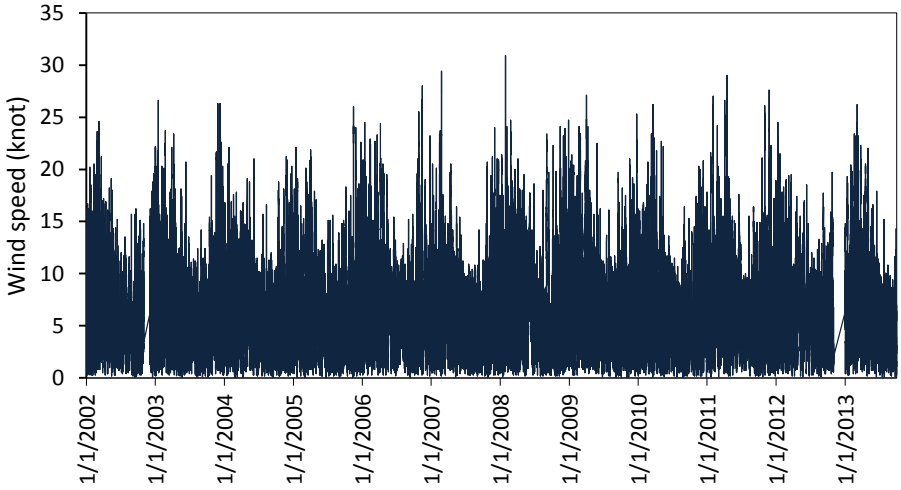


Figure 6. Time series of 1-hr temperature measured at C13 CAMS

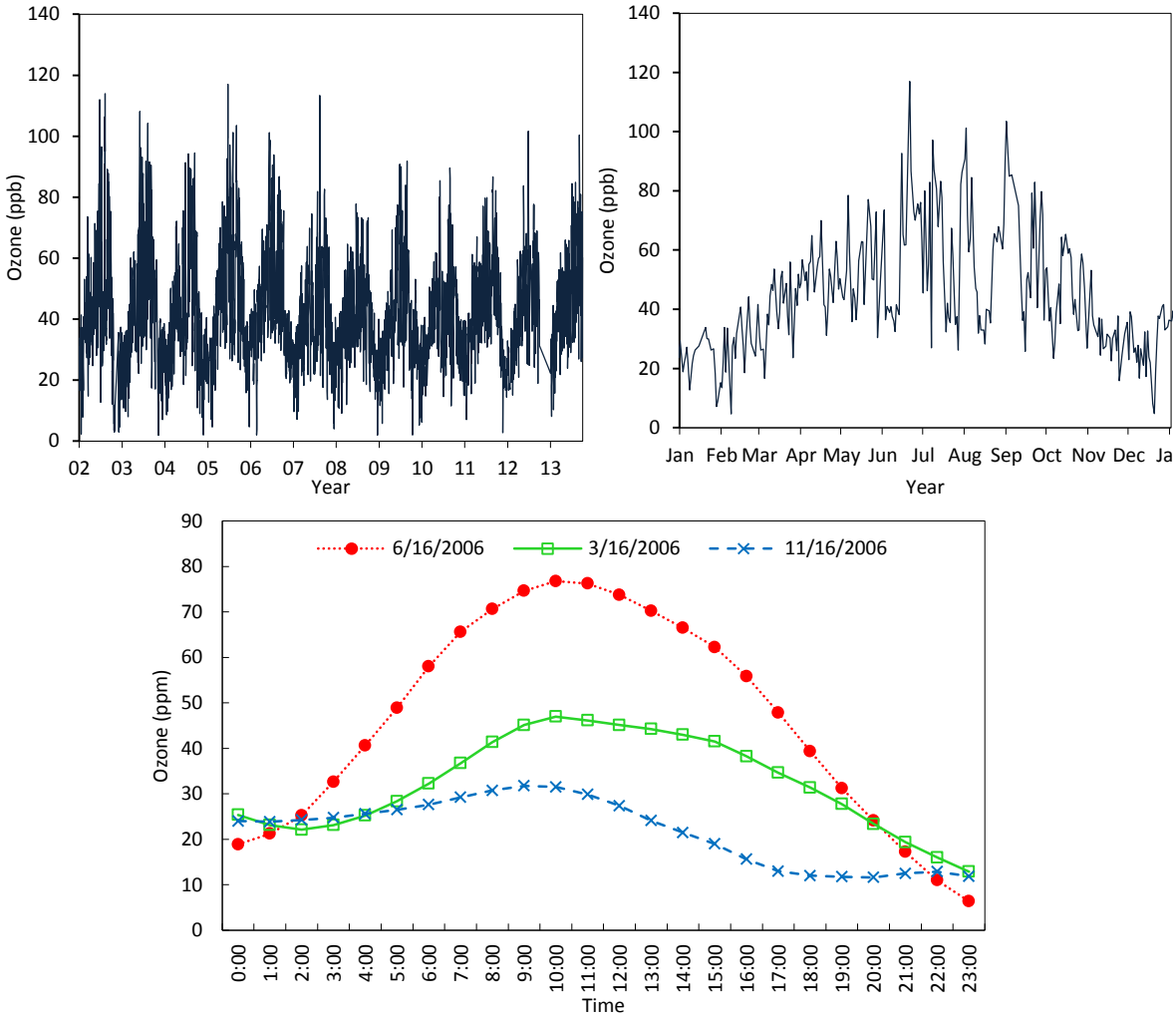


Figure 7. Time series of 8-hr average daily maximum ozone (top left); time series for one random year (top right); 8-hr average ozone profile for three random days measured at C13 CAMS

4. Data Mining

After steps I and II (data acquisition and pre-processing) for step III, simple k-means clustering analysis was performed on time series of 8-hr average daily maximum ozone. Clustering analyses were performed with different number of clusters (k) and two distance function (Euclidean and Manhattan distance) to find the optimum arrangement. Results show no significant difference between Euclidean and Manhattan distance functions. To select number of clusters (k) three main criteria were considered (1) a solution with reasonable within-cluster sum of square error (SSW); (2) clusters with minimum variability in each cluster; (3) high interpretable solution based on the knowledge of the ozone pollution in the area. Result of the clustering analysis for C13 CAMS is shown in fig. 8. Plot box representations of ozone data in each temporal cluster are shown in figs. 9 to 12. In Step (IV) the results are used to produce three ozone season clusters based on their share in each month of the year: low, moderate, and high (shown in Table 1).

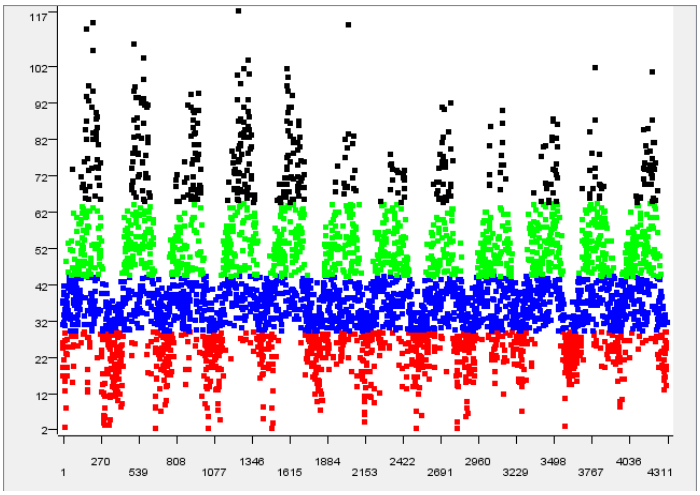


Figure 8. Seasonal clusters of ozone using k-means method at C13 CAMS

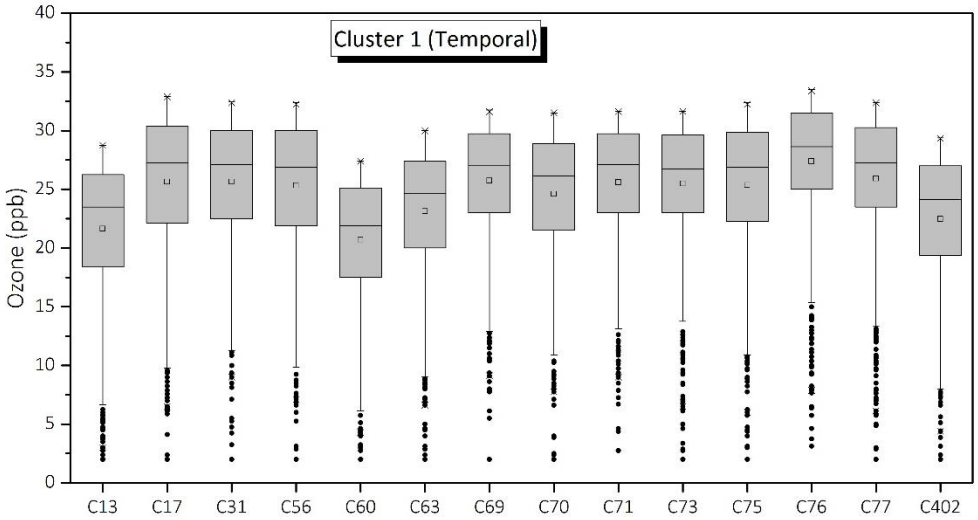


Figure 9. Box plot presentation of CAMS in temporal cluster #1

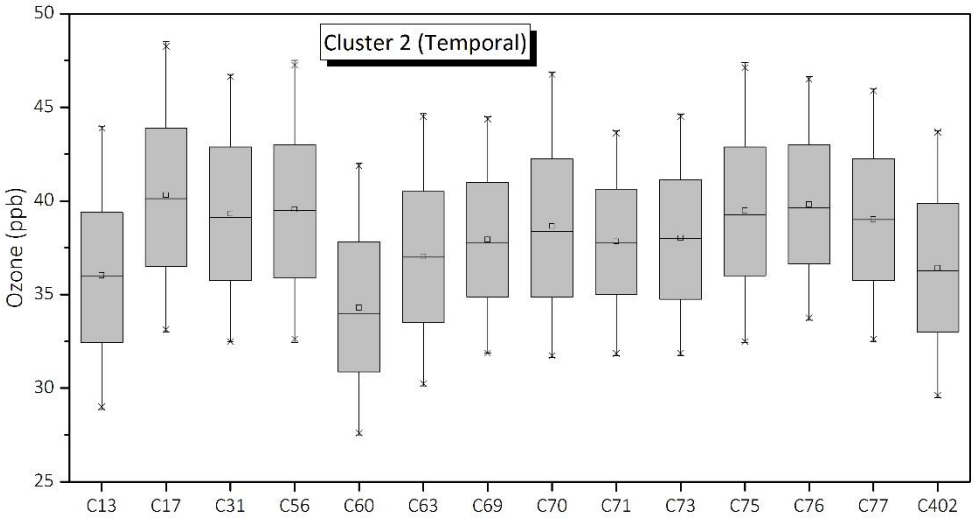


Figure 10. Box plot presentation of CAMS in temporal cluster #2

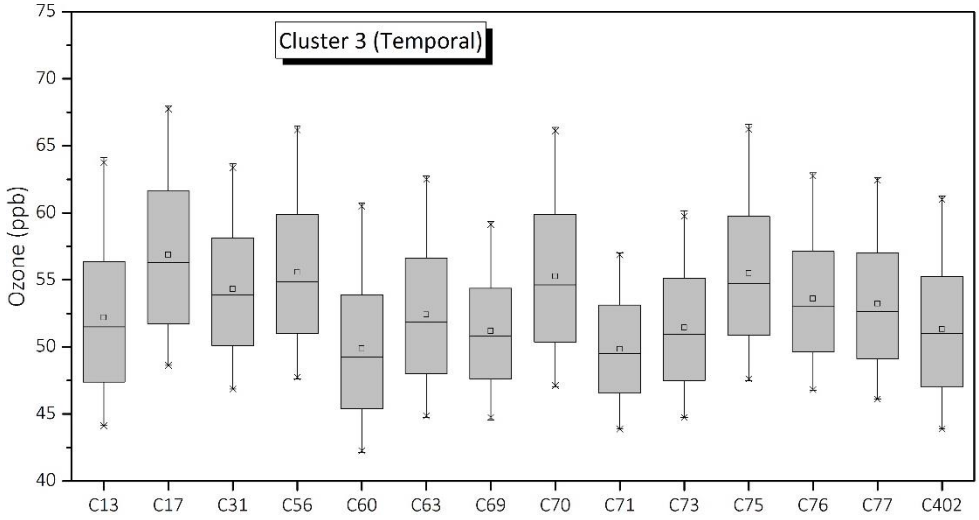


Figure 11. Box plot presentation of CAMS in temporal cluster #3

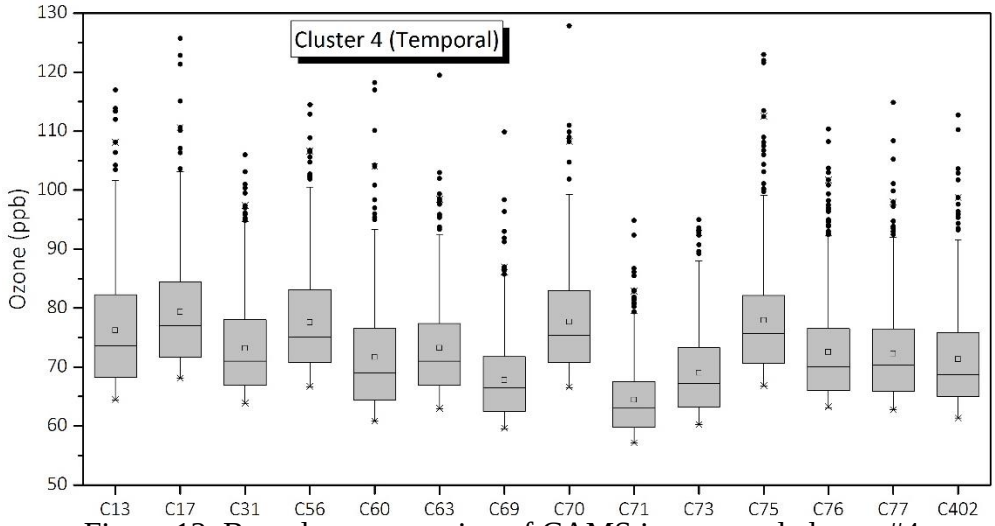


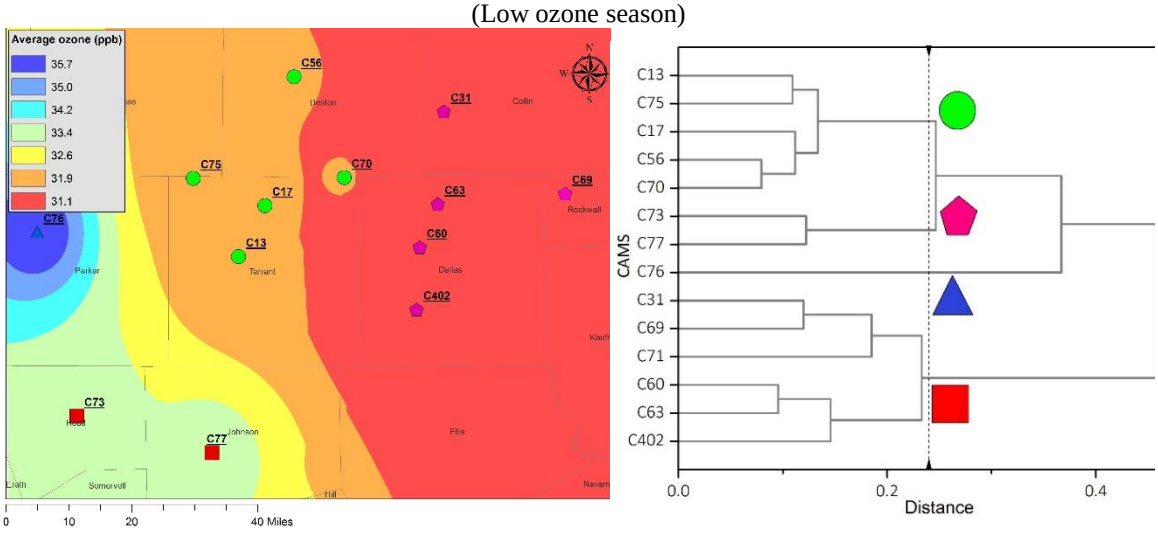
Figure 12. Box plot presentation of CAMS in temporal cluster #4

Cluster	Season	Months			
#1	Low	Jan	Feb	Nov	Dec
#2	-	-	-	-	-
#3	Moderate	Mar	Apr	May	Oct
#4	High	Jun	Jul	Aug	Sep

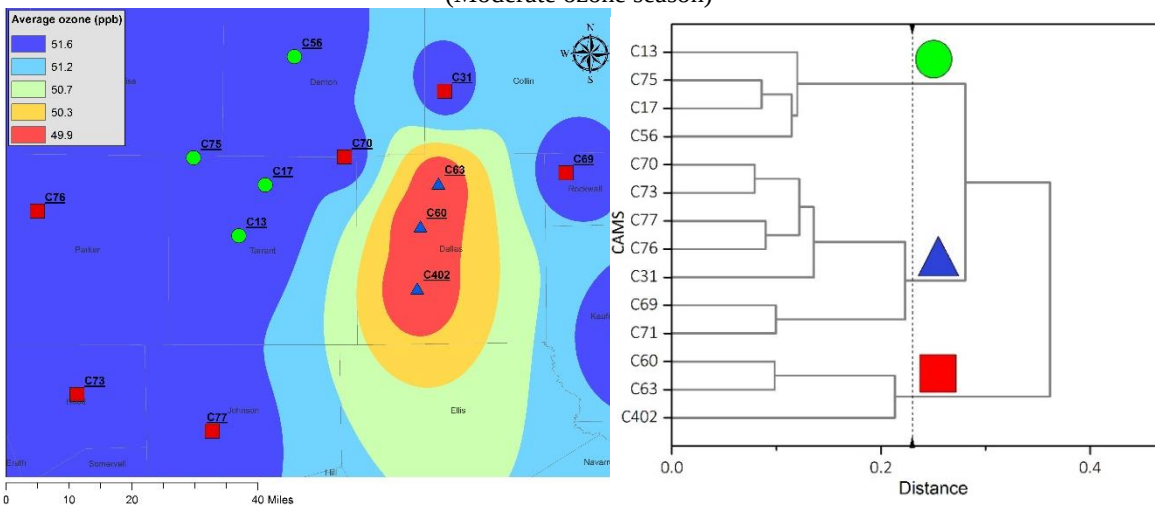
Table 1. Ozone seasons resulted from ozone time series clustering

In step V, CAMS in each season were clustered based on similarities so that spatial pattern of ozone behavior can be recognized. Agglomerative hierarchical cluster analysis was performed following the Ward's method where the increase in squared error when two clustered are merged is the criterion for making a new cluster. Fig. 13 shows the hierarchical trees and clusters on the map of the area. The average value of ozone in each zone (i.e. cluster) is used with inverse distance weighting function (Fortin and Dale, 2005) in ArcGIS© software to produce the maps.

(Low ozone season)



(Moderate ozone season)



(High ozone season)

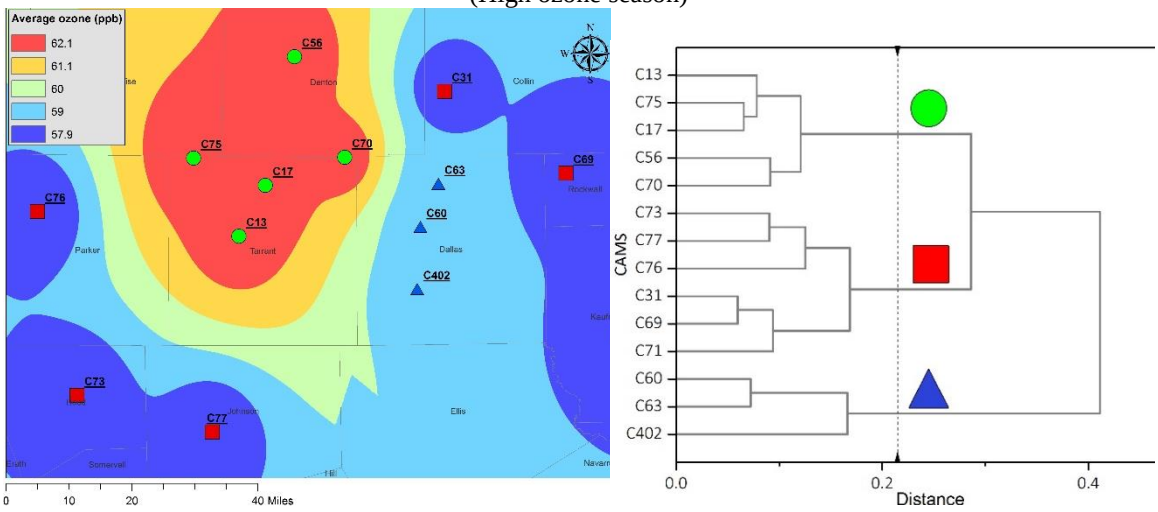


Figure 13. Hierarchical cluster trees and average ozone concentration for low, moderate, and high (top, middle, bottom) seasons

5. Ozone Forecasting

The last step is developing a multiple linear regression model. The general form of the multiple linear function for ozone forecasting ozone at time (t) is given by:

$$\text{Log}[O_3]_t = \alpha T_{t-i} + \beta \text{SR}_{t-j} + \chi W_{t-k} + \omega \text{ Log}[O_3]_{t-l} + \lambda \quad (1)$$

where $\alpha, \beta, \chi, \omega, and\ \lambda$ are linear regression coefficients for T (temperature) at time $t-i$, SR (solar radiation) at time $t-j$, W (wind speed) at time $t-k$, and logarithm of ozone at time $t-l$ respectively. The goal is to determine coefficients and time lags so that R^2 (coefficient of determination) and RMSE (root mean square error) of the fitting are optimum. The best time lags in were determined by varying i, j, k, and l independently and evaluating R^2 and RMSE. Results show the best time lag for meteorological factors is zero ($i = j = k = 0$) and for previous ozone is one hour ($l = 1$).

Time series of average 1-hr ozone in three spatial clusters in the high season (shown in fig. 13, bottom) were forecasted by three independent multiple linear regression models. The parameters of liner regression model are presented in Table 1. The scatter plot of predicted versus observed 1-hr ozone concentration for three spatial clusters during high ozone are shown in fig. 14. The linear model developed for each cluster is applied to predict ozone and the results are shown in fig. 15. High accuracy of prediction even without using time series of ozone precursors is shown.

Cluster	α	β	χ	ω	λ	R^2	RMSE
#1	0.002797	0.359812	0.021030	0.793794	0.176548	0.827	7.361
#2	0.000206	0.241734	-0.000247	0.871374	0.356038	0.890	5.142
#3	0.002699	9.652480	0.027754	0.687160	0.311890	0.827	7.872

Table 1. Summary of linear regression parameters for spatial clusters in high ozone season

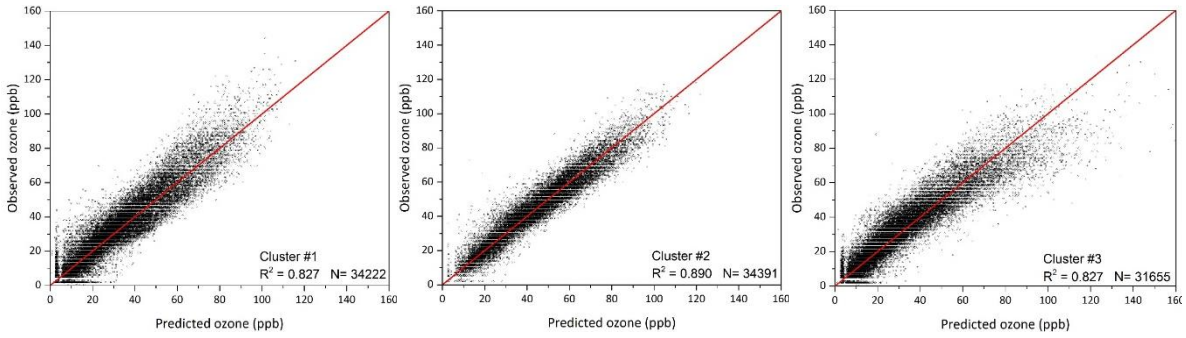


Figure 14. Scatter plots of predicted against observed ozone concentration for three clusters in high ozone season

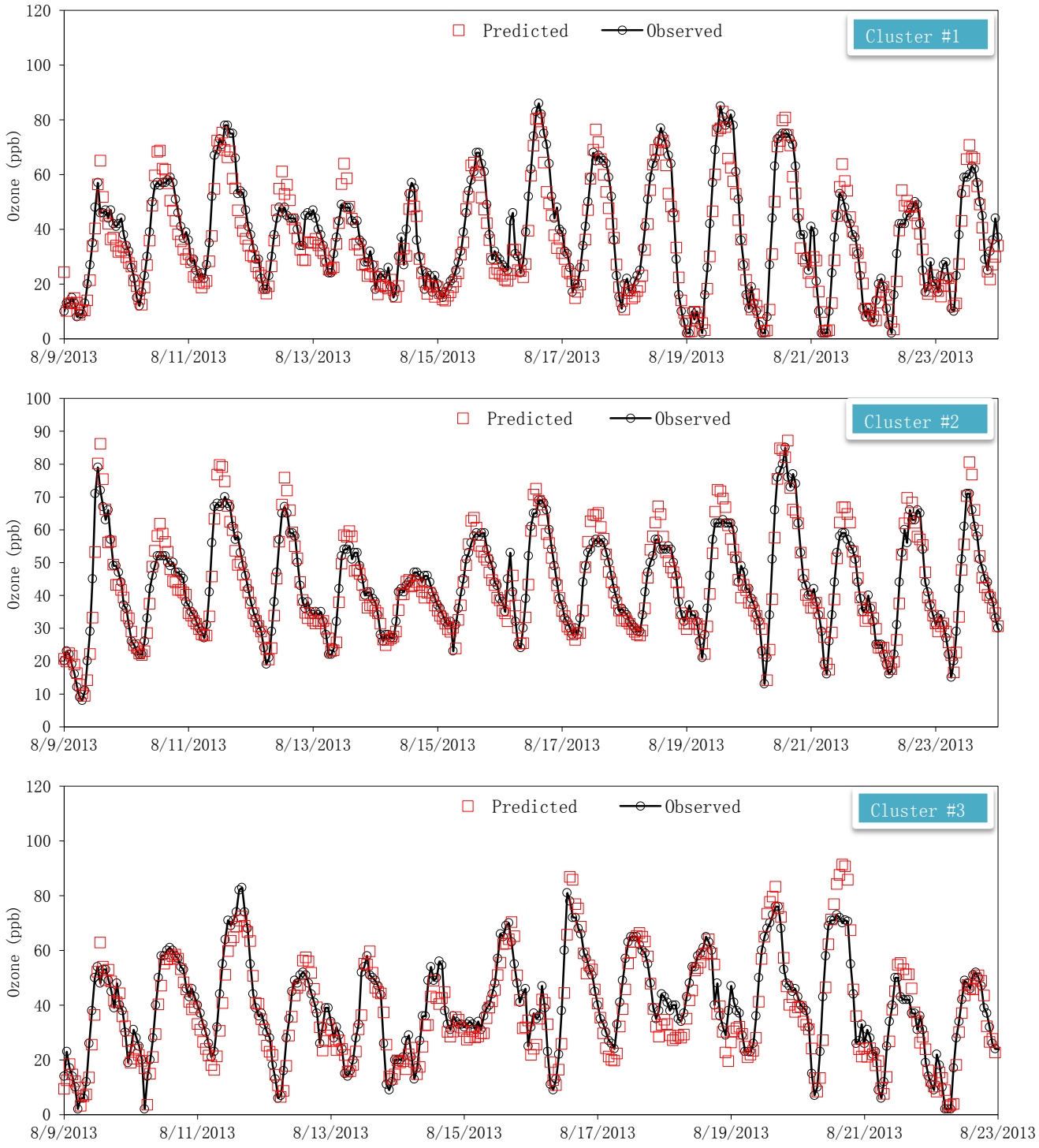


Figure 15. Comparison of observed and predicted 1-hr ozone concentration in spatial clusters in high ozone season

6. Conclusion

In this research multivariate data mining techniques were used to increase ozone forecasting accuracy and ease the interpretation of the model. Instead of categorizing time series to conventional seasons of the year, ozone seasons were driven from the measurement data. Temporal pattern recognition helps reducing the variability of ozone in each cluster. Also, hierarchical cluster analysis was performed on fourteen monitoring stations in the area to recognize spatial pattern. The method allows to forecast ozone for a zone even when all but one of the air monitoring stations are down. Measurement data is used to validate the accuracy of linear models in each cluster. Results show very high accuracy of ozone forecast using only meteorological variables. Therefore, it can be concluded that using data mining techniques in the proposed way can increase accuracy of the estimating ozone and flexibility of forecasting.

7. References

ABDUL-WAHAB, S., BOUHAMRA, W., ETTOUNEY, H., SOWERBY, B. & CRITTENDEN, B. D. 1996. Predicting ozone levels. *Environmental Science and Pollution Research,* 3**,** 195-204.

AL-ALAWI, S. M., ABDUL-WAHAB, S. A. & BAKHEIT, C. S. 2008. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. *Environmental Modelling & Software,* 23**,** 396-403.

AUSTIN, E., ZANOBETTI, A., COULL, B., SCHWARTZ, J., GOLD, D. R. & KOUTRAKIS, P. 2014. Ozone trends and their relationship to characteristic weather patterns. *Journal of Exposure Science and Environmental Epidemiology*.

BRUNO, F., COCCHI, D. & TRIVISANO, C. 2004. Forecasting daily high ozone concentrations by classification trees. *Environmetrics,* 15**,** 141-153.

CASSMASSI, J. & BASSETT, M. 1993. Air quality trends in the south coast air basin. *Southern CaliJbrnia Air Qualit3, Study Data Analysis: Proceedings of an International SpeciaIt3, Conjkrence***,** 1-6.

DUEÑAS, C., FERNÁNDEZ, M., CAÑETE, S., CARRETERO, J. & LIGER, E. 2002. Assessment of ozone variations and meteorological effects in an urban area in the Mediterranean Coast. *Science of the Total Environment,* 299**,** 97-113.

FEISTER, U. & BALZER, K. 1991. Surface ozone and meteorological predictors on a subregional scale. *Atmospheric Environment. Part A. General Topics,* 25**,** 1781-1790.

FIORE, A. M., JACOB, D. J., LOGAN, J. A. & YIN, J. H. 1998. Long-term trends in ground level ozone over the contiguous United States, 1980–1995. *Journal of Geophysical Research: Atmospheres (1984–2012),* 103**,** 1471-1480.

FORTIN, M. J. & DALE, M. R. T. 2005. *Spatial Analysis: A Guide for Ecologists*, Cambridge University Press.

GUARDANI, R., AGUIAR, J. L., NASCIMENTO, C. A., LACAVA, C. I. & YANAGI, Y. 2003. Ground-level ozone mapping in large urban areas using multivariate statistical analysis: Application to the Sao Paulo Metropolitan area. *Journal of the Air & Waste Management Association,* 53**,** 553-559.

KABURLASOS, V. G., ATHANASIADIS, I. N. & MITKAS, P. A. 2007. Fuzzy lattice reasoning (FLR) classifier and its application for ambient ozone estimation. *International Journal of Approximate Reasoning,* 45**,** 152-188.

KATSOULIS, B. D. 1996. The relationship between synoptic, mesoscale and microscale meteorological parameters during poor air quality events in Athens, Greece. *Science of the total environment,* 181**,** 13-24.

KORSOG, P. E. & WOLFF, G. T. 1991. An examination of urban ozone trends in the northeastern US (1973–1983) using a robust statistical method. *Atmospheric Environment. Part B. Urban Atmosphere,* 25**,** 47-57.

KOVAČ-ANDRIĆ, E., BRANA, J. & GVOZDIĆ, V. 2009. Impact of meteorological factors on ozone concentrations modelled by time series analysis and multivariate statistical methods. *Ecological Informatics,* 4**,** 117-122.

KUNTASAL, G. & CHANG, T. Y. 1987. Trends and relationships of O3, NOx and HC in the south coast air basin of California. *JAPCA,* 37**,** 1158-1163.

LENGYEL, A., HÉBERGER, K., PAKSY, L., BÁNHIDI, O. & RAJKÓ, R. 2004. Prediction of ozone concentration in ambient air using multivariate methods. *Chemosphere,* 57**,** 889-896.

LOU THOMPSON, M., REYNOLDS, J., COX, L. H., GUTTORP, P. & SAMPSON, P. D. 2001. A review of statistical methods for the meteorological adjustment of tropospheric ozone. *Atmospheric Environment,* 35**,** 617-630.

SAHU, S. K. & BAKAR, K. S. 2012. Hierarchical Bayesian autoregressive models for large space–time data with applications to ozone concentration modelling. *Applied Stochastic Models in Business and Industry,* 28**,** 395-415.

SAHU, S. K., GELFAND, A. E. & HOLLAND, D. M. 2007. High-resolution space–time ozone modeling for assessing trends. *Journal of the American Statistical Association,* 102**,** 1221-1234.

SCHLINK, U., DORLING, S., PELIKAN, E., NUNNARI, G., CAWLEY, G., JUNNINEN, H., GREIG, A., FOXALL, R., EBEN, K. & CHATTERTON, T. 2003. A rigorous inter-comparison of ground-level ozone predictions. *Atmospheric Environment,* 37**,** 3237-3253.

SEINFELD, J. H. & PANDIS, S. N. 2012. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, Wiley.

SOUSA, S., MARTINS, F., ALVIM-FERRAZ, M. & PEREIRA, M. C. 2007. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environmental Modelling & Software,* 22**,** 97-103.

WALKER, H. M. 1985. Ten-year ozone trends in California and Texas. *Journal of the Air Pollution Control Association,* 35**,** 903-912.

ZELDIN, M., CASSMASSI, J. & HOGGAN, M. Ozone trends in the South Coast Air Basin: an update. Tropospheric Ozone and the Environment: Papers from an International Conference, edited by RL Berglund, DR Lawson, and DJ McKee, 1990. 1-12.