Maximum Ozone level prediction in Athens with the aid of the CART system

A. Kaprara, K. Karatzas and N. Moussiopoulos
Laboratory of Heat Transfer and Environmental Engineering, Aristotle University Thessaloniki, Box 483, 54006 Thessaloniki, Greece

Keywords: Ozone forecast, prediction of air quality episodes, CART

1 Introduction

The Metropolitan area of Athens is frequently characterised by high photochemical pollution levels especially in the warm period. The highest levels of ozone concentrations are found in the urban background stations while the ozone concentrations are reduced in the traffic stations of the central and southern zone. For ozone prediction purposes, many methods are reported in the literature such as regression and ARIMA techniques (Chaloulakou et al, 1997; Robeson and Steyn, 1989), artificial neural networks (Spellman, 1999), 3-D air quality models (Moussiopoulos et al., 1995; Grell et al., 1994).
Within the APNEE project framework[1], a method for daily maximum ozone prediction has been developed. In order for the citizens to be informed about the levels of the air pollutants in time, CART is used for the daily prediction of the maximum ozone concentration in the Athens basin. Its results are compared with the corresponding ones of a simple regression equation according to some common model evaluation parameters and their ability to predict ozone episodes. CART has been recommended for ozone forecasting (EPA, 1999) and has been used for this purpose in several studies (Ryan, 1995; Burrows et al., 1995; Gardner and Dorling, 2000).

2 The CART tool

CART is an advanced tool for tree-structured data analysis and its structure is similar to the rule architecture in a neural network (Atlas et al. 1990). It is the acronym for Classification And Regression Trees, a statistical procedure introduced by Leo Breiman, Jerome Friedman, Richard Olsen and Charles Stone in 1984 (Breiman et al. 1984). CART can be used to analyze both continuous (regression) and categorical data (classification). The methodology used by CART is known as binary recursive partitioning. CART splits binary the data into two groups-nodes and this binary splitting is repeated until some conditions are satisfied. The result of this methodology is a binary tree, whose terminal nodes represent distinct classes or categories of data. CART is non-parametric, in the sense that the number of parameters used is not specified beforehand and there is no assumption about their distribution.
The CART analysis is based on a learning sample that consists of the variable which is going to be analyzed-predicted (predictand) and the variables used for the classification or regression of the predictand (predictors). Each binary splitting is based on a simple rule involving one or more predictor values. This rule is a simple comparison between a value of a selected predictor variable (or a linear combination of predictors) and a threshold. CART looks at all possible splits for all variables and selects the one that results in the most dissimilar nodes. Once the best split is found, the process is repeated until further splitting is impossible or stopped. In this way, a maximal tree is grown which then pruned until an optimum number of terminal nodes is found. The criterion for the pruning is the minimisation "cost" of miscalculation.
An important advantage of the CART method is that the resulting tree facilitates inference according to relationships between predictors and predictand and has a physical meaning. It can manipulate missing values, since for each split it develops alternative ones. Unlike regression equations, it can use both continuous and categorical variables.

[1] The APNEE project aims at increasing the knowledge of the citizen on air quality. (http://apnee.faw.uni-ulm.de/)

3 Methodology approach

For the daily maximum ozone prediction, meteorological and air quality data are used. In order for the prediction to have a practical value, it must be realized in time, e.g. every morning. Therefore, the data intended to be used for the forecast must be available every day. This fact limits the number of variables that can be utilized. For Athens basin, the daily dispensable data are:

- Air pollutants
  The maximum and minimum daily concentrations of ozone, nitrogen dioxide, sulfur dioxide, carbon monoxide and smoke. In the Athens area, nine air quality monitoring stations are operating by the Hellenic Ministry of Environment but only the total maximum and minimum concentrations of all stations are available on line. The site is updated everyday at 14:00 and the values that are given are the daily maximum and minimum concentrations of the previous day and of the current day until 13:00.
- Meteorological data
  Air temperature, relative humidity, pressure, maximum and mean wind velocity, mean and standard deviation of wind direction. The values of these variables are updated every 10 minutes and represent conditions over the entire urban area and can be used, therefore, to predict the overall maximum ozone levels.

Additionally, the number of the month is used as a categorical predictor. For the formation of the trees a 9-year long period record (1990-1998) has been used as the learning sample. The resulting trees have been tested with the use of the data taken from 1999 records. In the optimal tree, the important variables for the prediction of the maximum ozone concentration are the ozone concentration of the previous day, the maximum temperature, the mean wind velocity and the number of the month. For comparison purposes, a regression equation has been used for the ozone prediction with the same predictors (with the exception of the number of the month, which is a categorical variable).

4 Application results

For the evaluation of the CART analysis performance, several statistical measures have been used which are the following: Mean values and standard deviations of observations and predictions, mean absolute error, mean percentage error, normalized mean difference, root mean square error (RMSE), systematic and unsystematic RMSE, correlation coefficient and index of agreement. Of particular interest are the cases where the ozone concentration exceeds the threshold (180 $\mu g/m^3$) that is defined by the European Union as a critical value for the notification of the citizens. This case is considered as an alarm episode. For the evaluation of the ozone episodes forecasting ability, the percentage of correct (POD) and false alarms and the critical success index (CSI) are used (Ryan, 1995). The values of all these measures are presented in table 1 for both CART analysis and regression equation.

In figure 1, the optimal tree is shown. It is clearly illustrated that the most important predictors are the ozone concentration of the previous day and the maximum daily temperature. The first split uses the maximum temperature, which is expected taking into consideration that the ozone levels are higher when the temperature is increased. Also, the temperature is highly correlated to the solar radiation that can't be used in this study but has great effects on ozone levels. It is interesting to note that in most splits the ozone concentration of the previous day is used.

Table 1 Evaluation parameters of the performance.

1999 - test data	CART	Regression
Observed mean	136.126	136.126
Predicted mean	138.184	144.472
Observed standard deviation	49.781	49.781
Predicted standard deviation	40.421	47.883
Mean absolute error (MAE)	25.935	31.297
Mean percentage error (MPE)	-6.62%	-0.111
Normalized mean difference (NMD)	-0.015	-0.061
Root mean square error (RMSE)	34.295	38.517
Root mean square error, systematic ($RMSE_s$)	20.365	18.128
Root mean square error, unsystematic ($RMSE_u$)	27.594	33.984
Correlation coefficient	0.729	0.703
Index of agreement	0.838	0.826
% correct alarms (POD)	86.96%	66.67%
% false alarms	27.60%	17.49%
Critical success index (CSI)	0.353	0.346

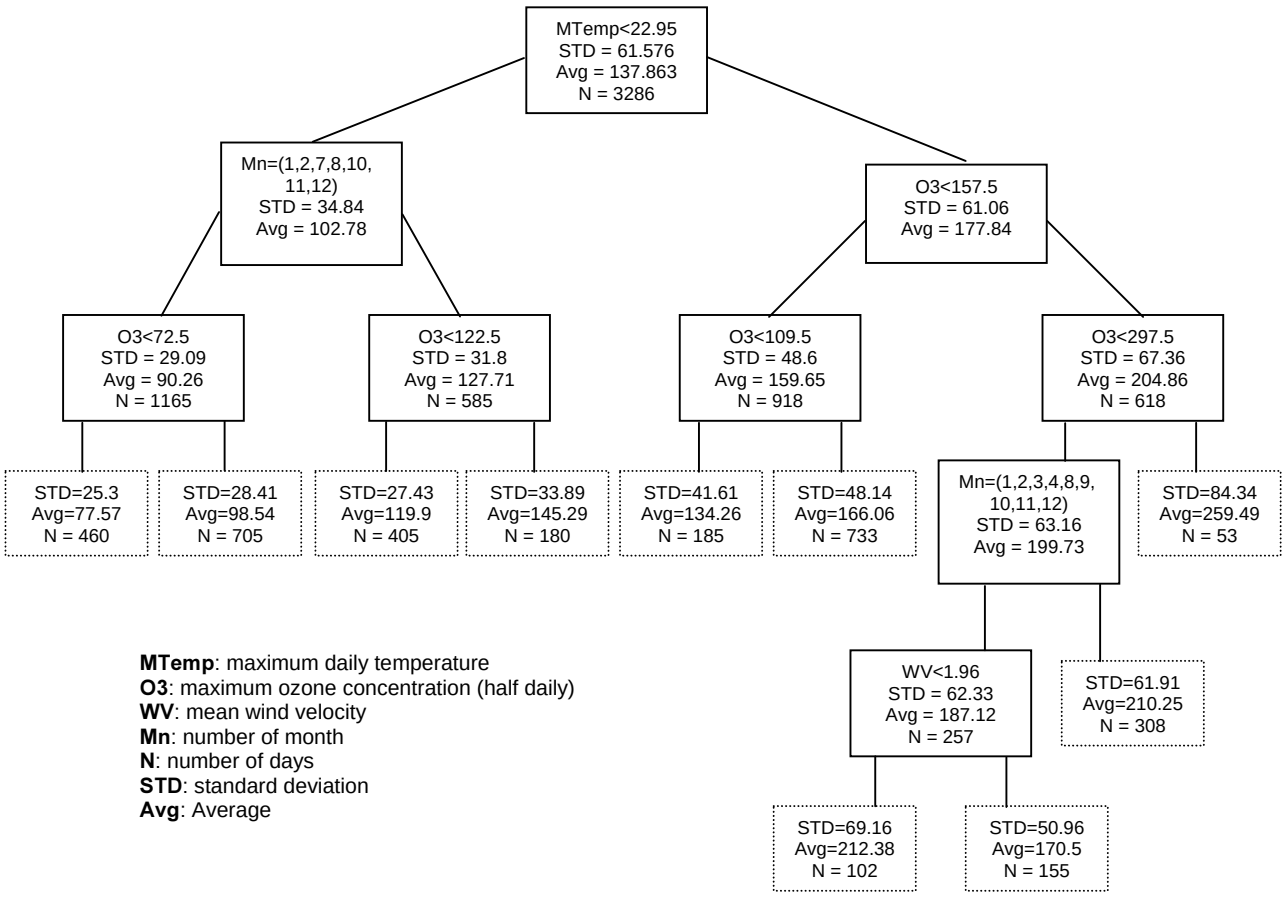

Figure 1 Regression tree model.

5 Discussion-Conclusions

The application of both the CART and the regression analysis methodologies result in a rather good index of agreement (>80%), while the additional measures calculated show a generally good performance. The results obtained are directly comparable (usually better) to what is mentioned in relevant literature (Chaloulakou et al, 1997; Spellman, 1999). These results assume other numerous tests with many variable sets, and represent the best outcome that can be expected taking into account a very important limitation: the on-line (on the Internet) availability of data, which is limited to general (and not detailed) air quality information, and only few meteorological information. Taking into account that the latter were made available to similar works reported in the literature (Gardner and Dorling, 2000; Prybutok et al. 2000), the results of the work presented here become more important, and underline a significant factor of success: the exhaustive tests that one has to perform in order to come up with results that are satisfactory and can support the operational use of the method. Future work will exploit more this direction.

References

1. Atlas L., Cole R. Muthusamy Y., Lippman A., Connor J., Park D., El-Sharkawi M., Marks R.J., (1990), 'A performance comparison of trained multilayer perceptons and trained classification trees', Proc. IEEE, Vol. 78, pp. 1614-1619.
2. Breiman, L., Friedman, J.H., Olshen, R., Stone, C.J., (1984), 'Classification And Regression Trees', Wadsworth, Belmont CA.
3. Burrows, W.R., Benjamin, M., Beauchamp, S., Lord, E.R., McCollor, D. & Thomson, B., (1995), 'CART decision-tree statistical analysis and prediction of Summer season Maximum surface Ozone for the Vancouver, Montreal and Atlantic regions of Canada', *Journal of Applied Meteorology*, Vol. 34, pp. 1848-1862.
4. Chaloulakou A., Assimakopoulos D., Lekkas T., (1999), 'Forecasting daily maximum ozone concentrations in the Athens basin', *Environmental Monitoring and Assessment*, Vol. 56, pp. 97-112.
5. Dudhia J., (1993), 'A non-hydrostatic version of the Penn State/NCAR mesoscale model: validation tests and simulation of an Atlantic cyclone and cold front', *Mon. Wea. Rev.* Vol. 121, pp. 1493-1513.
6. EPA-454-R-99-009, (1999), 'Guideline For Developing An Ozone Forecasting Program', US Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, North Carolina 27711, July 1999.
7. Gardner M.W., Dorling S.R., (2000), 'Statistical surface ozone models: an improved methodology to account for non-linear behaviour', *Atmos. Environ.*, Vol. 34, pp. 21-34, 2000.
8. Grell G.A., Dudhia J., and Stauffer D.R. (1994), 'A description of the fifth-generation Penn State/NCAR mesocale model (MM5)', Prepared by National Center for Atmospheric Research, Boulder, CO, NCAR Technical Note-398.
9. Moussiopoulos N., Sahm P. and Kessler Ch. (1995), 'Numerical simulation of photochemical smog formation in Athens, Greece-a case study', *Atmos. Environ.*, Vol. 29, pp. 3619-3632.
10. Prybutok V., R., Yi J., Mitchell D., (2000), 'Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations', *European Journal of Operational Research*, Vol. 122, pp. 31-40.
11. Robeson S., Steyn D., (1989), 'Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations', *Atmos. Environ.*, Vol. 24B, pp. 303-312.
12. Ryan W.F., (1995), 'Forecasting severe ozone episodes in the Baltimore metropolitan area', *Atmos. Environ.*, Vol. 29, pp. 2387-2398.
13. Spellman G.,(1999), 'An application of artificial neural networks to the prediction of surface ozone concentrations in the United Kingdom', *Applied Geography*, Vol. 19, pp. 123-136.