Ground-level ozone forecast based on machine learning

R. Zabkar[1], J. Zabkar[2] & D. Cemas[3]
[1]*Faculty of Mathematics and Physics, Ljubljana, Slovenia*
[2]*Faculty of Computer and Information Science, Ljubljana, Slovenia*
[3]*Environmental Agency of the Republic of Slovenia, Ljubljana, Slovenia*

Abstract

In this paper we apply methods of machine learning to the problem of ground-level ozone forecasting, using measured data and data calculated by the numerical weather prediction model ALADIN (Aire Limitee Adaptation Dynamique developement InterNational). Our goal is to build a simple ozone-forecasting tool to predict the daily maximum of ground-level ozone concentration per day using meteorological and air quality data.

Tropospheric ozone episodes in Slovenia are mainly due to the local traffic sources and the long-range transport of ozone and its precursors, generally originating from Western Europe.

Ozone forecast model was developed for the purpose of issuing public alerts to avoid exposure to high ground-level ozone levels and so that concerned citizens, industrial organizations and local authorities could take action to reduce harmful emissions of ozone precursors.

The results enable an empirical approach to the short term ozone level forecasts, estimating the ozone trends and increasing the scientific understanding of the underlying mechanisms.

Keywords: ground-level ozone, air quality modelling, air pollution, machine learning.

1 Introduction

Ozone in the lower atmosphere has harmful effects on vegetation and human health. In Slovenia typical maximum ozone concentration during summer is between 200 and 230 μg/m^3. The information and alert thresholds for human health (180 and 240 μg/m^3 in 1 h respectively) are only exceeded a few times per

Air Pollution XII, C. A. Brebbia (Editor)

year in capital city – Ljubljana, but it is more problematic in coastal regions. High tropospheric ozone episodes in Slovenia are mainly due to the local sources and the long-range transport of ozone and its precursors, generally originating from Western Europe. The highest ozone concentrations occur in summer, showing a diurnal variation with a maximum in the afternoon and a minimum during the night [5]. Ozone concentration in rural and elevated areas is typically twice higher as in urban areas but during summer also in populated areas ozone concentration is increased.

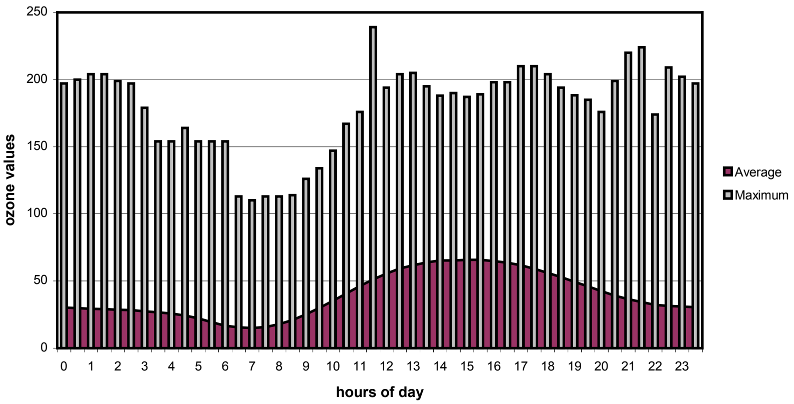

Figure 1: Average and maximum ozone concentrations in the period 1996 – 2002 in Ljubljana.

Ozone forecast model was developed for the purpose of issuing public alerts to avoid exposure to high ground-level ozone levels and so that concerned citizens, industrial organizations and local authorities could take action to reduce harmful emissions of ozone precursors.

At air quality forecasting different methods can be applied. Empirical models on one hand such as regression ([4] or [2]) as statistical approach and deterministic models based on weather predictions systems and emission inventories [3] are most common approaches to ozone forecast. In this paper we apply methods of machine learning to the problem of ground-level ozone prediction, using measured data and predictions by ALADIN. Machine learning methods that we used learn rules from training data and the learnt model can be used for forecasting. Thus we expect that a learning program may be able to figure out how ALADIN's prediction in similar situations in the past is correlated to ozone concentrations. In addition, the ALADIN model is only run at midnight, and we may improve the forecast by incorporating the meteorological and air quality data collected during the day.

We used an approach to machine learning called regression tree learning implemented in machine learning software Weka [7]. We chose regression trees because they are not black boxes unlike some other machine learning methods

Air Pollution XII, C. A. Brebbia (Editor)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-722-1

such as neural networks or linear weighted regression. A learnt regression tree can be easily understood and thus provide new insight to the expert.

Our goal is to build a simple ozone-forecasting tool to predict the maximum ozone concentration of the current day using meteorological and air quality data gathered up to 8 a.m. on the current day. In order to have more records and for practical reasons half-hourly data were used and predicted by machine learning model.

2 Machine learning using regression trees

The input for a machine learning algorithm is a dataset which contains the data available for analysis. Records (rows) of the dataset are called examples and fields (columns) are called attributes, except for one field – the one we are trying to predict – which is called a class. The algorithm tries to learn how the attributes determine the class value in the examples.

We do not go into details about the algorithm here. They can be found either in [7] or any classical machine learning literature. We just mention the basic idea. A regression tree is a binary tree having attributes in its nodes and making splits on their values. How attributes are chosen for each node and the way the splits are found can be found in [7]. The leafs of the tree contain a linear regression formula which fits the examples that reach each leaf.

3 Input data

Input data are based on meteorological and air quality measurements and data calculated by mesoscale meteorological model ALADIN [1]. Measurements and model data were provided by Environmental Agency of the Republic of Slovenia.

For our purposes we have determined the following attributes:

- ALADIN model predictions in three-hour increments in 4 points (named A, B, C and D - nearest calculating point SW, SE, NE and SW of Ljubljana) around Ljubljana at ground level: temperature, dew-point temperature, wind speed and wind direction, solar radiation and precipitation.
- Measurements in half-hour increments in 1 measuring points in Ljubljana:
- meteorological measurement: temperature, relative humidity, wind speed and wind direction, solar radiation, precipitation
- air quality measurements: ozone (O_3), nitrogen monoxide (NO), nitrogen monoxide (NO_2), carbon monoxide (CO)

We also included cosine of the day in the year, which roughly approximating yearly variations of ozone.

Class value was maximum ozone concentration of the current day.

The data was available for years 2002 and 2003. There were a lot of missing values among measured data which made us throw away some learning

Air Pollution XII, C. A. Brebbia (Editor)

examples. We considered the final set of examples small, especially because we had to make a division on train and test set at the very beginning. Out of 547 examples, 66% were used for training and the remaining was used for testing. Data, included in the model, were manually selected from the large quantity of data available from measurement and weather forecasting system (ALADIN). Only the most important air quality measurements and ground level ALADIN data were used, as described in the next chapter.

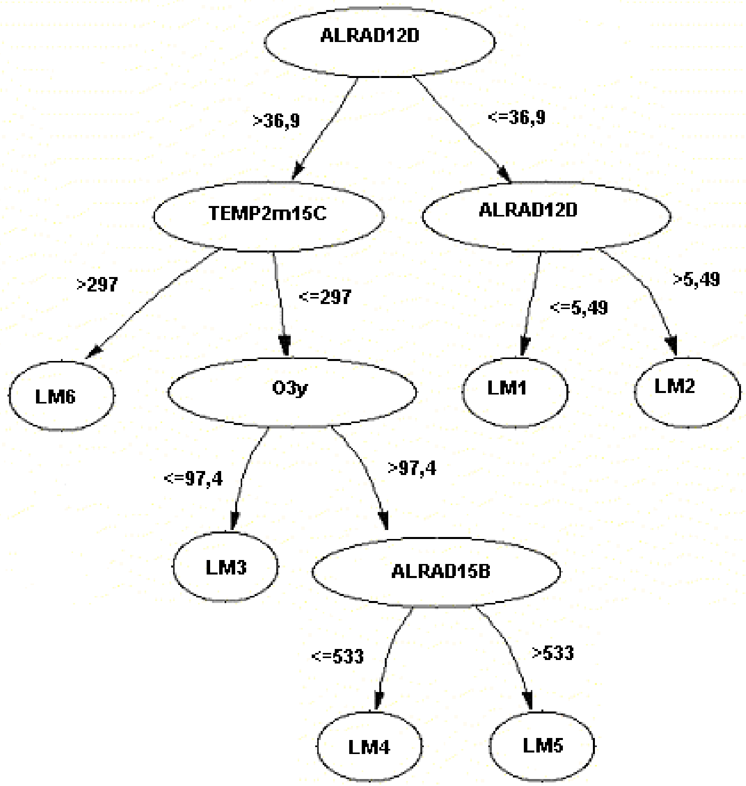

Figure 2: Regression tree constructed for ozone forecast.

4 Results and discussion

Only a small number of available attributes were used for learning. At the very first stage we made a selection based on expert knowledge. We were, for instance, sure that ALADIN's predictions of solar radiation and temperature for midday would have high impact on ozone concentration. The number of attributes selected this way (144) was still too high to be efficiently used on such a small set of instances. At the next stage we used machine learning algorithm

ReliefF [6] which is used for attribute selection (Appendix 1). This gave us the final set of attributes that we used for learning.

Table 1: Evaluation of ground-level ozone in Ljubljana forecasted by linear regression on test split.

	Regression tree
CORR	0,9099
MAE	12,3164
RMSE	15,3611

Abbreviations used:
CORR correlation coefficient
MAE mean absolute error
RMSE root mean squared error

Small changes in learning dataset can result in great differences in regression tree construction. This means that some trees are more accurate for certain examples than others.

Decision tree, constructed by model (Figure 1) and linear regression formulas (Appendix 2), shows that ozone concentration is mostly correlated to solar radiation, temperature and relative humidity. As already mentioned in the text, highest ozone concentrations occur during daytime in summer in hot sunny and dry weather. Also concentration of ozone one day before and cosine of day in the year plays important role, which shows some conservative characteristics in ozone episodes. Surprisingly no wind dependence can be found, which can indicate that local sources of ozone precursors in the city has important role to ozone formation in Ljubljana.

4.1 Results of other methods

In the course of research we also tried how plain linear regression perform on our problem. In linear regression model we used the same attributes as in machine learning algorithm. As we can see from final linear regression formula, only five of them show statistically significant influence on maximum daily ozone concentrations:

O3 = -434,639 + 0,361 O3y + 0,024 ALRAD15B + 0,603 COSY – 30,572 RH2m15C + 1,733 TEMP2m12B

Table 2: Evaluation of ground-level ozone forecasted by linear regression in Ljubljana on test split.

	Linear regression
CORR	0,9144
MAE	12,7345
RMSE	16,3005

Air Pollution XII, C. A. Brebbia (Editor)

The results of simple linear regression model were nearly the same compared to regression trees – a little better correlation and MAE and a little worse RMSE. From Table 2 one can see that maximum ozone forecast accuracy predicted by linear regression is nearly the same as predicted by regression tree.

5 Conclusion

Machine learning using regression trees is a simple ozone-forecasting tool for prediction of maximum ozone concentration of the current day using meteorological and air quality data gathered up to 8 a.m. on the current day.

Maximum ozone concentration forecast accuracy predicted by linear regression is nearly the same as predicted by regression tree. Small changes in learning dataset can result in great differences in regression tree construction. Results shows that ozone concentration is mostly correlated to solar radiation, temperature and relative humidity and no correlation to wind velocity and direction can be found.

Our experiments show that machine learning can be efficiently used to predict maximum ozone concentrations. Forecast with machine learning is fast and can easily be used in daily forecasts. It can also produce forecasts based on that common data such as temperature, solar radiation and air quality data without complex data that other deterministic tools would need.

The results enable the empirical approach to the short term ozone level forecasts, estimating the ozone trends and increasing the scientific understanding of the underlying mechanisms.

In this paper only ozone data in Ljubljana was investigated. We expect different attributes to be important in costal regions where long-range transport from Western Europe could have an important role on ozone concentration. In that case we expect that forecast would improve significantly by using trajectory analysis. We also intend to use more sophisticated machine learning methods for further analysis.

Appendix 1

Number of instances: 547
Test mode: split 66% train, remainder test
Number of attributes: 28
Selected attributes:
ALRAD12SW - solar radiation predicted by Aladin at 12 p.m. in point A
TEMP2m15A - temperature predicted by Aladin at 15 p.m. in point A
RH2m15A - relative humidity predicted by Aladin at 15 p.m. in point A
ALRAD15A - solar radiation predicted by Aladin at 15 p.m. in point A
TEMP2m12B - temperature predicted by Aladin at 12 p.m. in point B
TEMP2m15B - temperature predicted by Aladin at 15 p.m. in point B
RH2m15B - relative humidity predicted by Aladin at 15 p.m. in point A
ALRAD15B - solar radiation predicted by Aladin at 15 p.m. in point B
TEMP2m12C - temperature predicted by Aladin at 12 p.m. in point C

Air Pollution XII, C. A. Brebbia (Editor)

TEMP2m15C - temperature predicted by Aladin at 15 p.m. in point C
RH2m15C – relative humidity predicted by Aladin at 15 p.m. in point C
ALRAD15C - solar radiation predicted by Aladin at 15 p.m. in point C
TEMP2m12D - temperature predicted by Aladin at 12 p.m. in point D
ALRAD12D - solar radiation predicted by Aladin at 12 p.m. in point D
TEMP2m15D - temperature predicted by Aladin at 15 p.m. in point D
RH2m15D – relative humidity predicted by Aladin at 15 p.m. in point D
ALRAD15D - solar radiation predicted by Aladin at 15 p.m. in point D
COSY - cosine of the day in the year
RR - measured precipitation one day before
MAXCO - maximum CO concentration one day before
MAXNO - maximum NO concentration one day before
MAXNO2 - maximum NO_2 concentration one day before
M_O3 - trend of O_3 concentration one day before
M_CO - trend of CO concentration one day before
M_NO - trend of NO concentration one day before
M_NO2 - trend of NO_2 concentration one day before
O3y – O_3 concentration one day before
O3 – ozone concentration (predicted)

Appendix 2

Linear regression formulas:

LM1: O3 = -24.4 + 0.291ALRAD12A - 57RH2m15A + 0.0201ALRAD15A + 0.198TEMP2m12B + 0.37TEMP2m15B + 10.7RH2m15B + 0.0277ALRAD15B + 0.434TEMP2m12C - 1.51TEMP2m15C - 12.9RH2m15C - 0.00333ALRAD15C + 0.982TEMP2m12D - 0.0959ALRAD12D - 0.0958TEMP2m15D + 0.472RH2m15D + 3.21COSY - 0.124MAXCO - 0.00306MAXNO + 1.18M_NO2 + 0.0632O3y

LM2: O3 = -459 + 0.291ALRAD12A - 6.91RH2m15A + 0.0201ALRAD15A + 0.198TEMP2m12B + 0.37TEMP2m15B + 10.7RH2m15B + 0.0277ALRAD15B + 0.434TEMP2m12C - 1.51TEMP2m15C - 50.4RH2m15C - 0.00333ALRAD15C + 2.5TEMP2m12D - 0.0959ALRAD12D - 0.0958TEMP2m15D + 0.472RH2m15D + 3.21COSY - 0.124MAXCO - 0.00306MAXNO + 1.18M_NO2 + 0.263O3y

LM3: O3 = -560 - 0.937ALRAD12A - 0.242TEMP2m15A + 1.13RH2m15A + 0.969ALRAD15A + 2.17TEMP2m12B + 0.376TEMP2m15B - 45.1RH2m15B + 9.21e-4ALRAD15B - 0.133TEMP2m12C - 0.212TEMP2m15C + 4.27RH2m15C + 0.00551ALRAD15C + 0.898TEMP2m12D - 0.552TEMP2m15D - 3.67RH2m15D + 9.84COSY + 0.0622O3y

LM4: O3 = -525 - 0.0705ALRAD12A - 2.27TEMP2m15A + 1.13RH2m15A + 0.0642ALRAD15A + 3.78TEMP2m12B + 0.533TEMP2m15B - 3.38RH2m15B

- 0.0444ALRAD15B - 0.133TEMP2m12C - 0.159TEMP2m15C + 7.08RH2m15C + 0.0471ALRAD15C + 0.973TEMP2m12D - 0.64TEMP2m15D - 4.79RH2m15D + 13COSY + 0.077O3y

LM5: O3 = -397 - 0.0705ALRAD12A - 1.47TEMP2m15A + 1.13RH2m15A + 0.0642ALRAD15A + 0.613TEMP2m12B + 2.34TEMP2m15B - 3.38RH2m15B + 9.21e-4ALRAD15B - 0.133TEMP2m12C - 0.159TEMP2m15C + 10.1RH2m15C + 0.0107ALRAD15C + 1.15TEMP2m12D - 0.797TEMP2m15D - 6.33RH2m15D + 3.45COSY + 0.36O3y

LM6: O3 = -373 + 0.443ALRAD12A - 2.11TEMP2m15A + 1.53RH2m15A - 0.361ALRAD15A + 0.146TEMP2m12B + 3.17TEMP2m15B + 39.8RH2m15B + 9.21e-4ALRAD15B - 4.3TEMP2m12C + 4.22TEMP2m15C - 1.36RH2m15C + 0.401TEMP2m12D - 0.14TEMP2m15D + 0.889COSY + 0.354O3y

References

[1] Aladin International Team, *The ALADIN project: mesoscale modelling seen as a basic tool for weather forecasting and atmospheric research*, WMO Bull., **46**, pp. 317-324, 1997.

[2] Cobourn, W. G. and Hubbard, M. C., *An Enhanced Ozone Forecasting Model using Air Mass Trajectory Analysis*, Atmospheric Environment, 33, pp. 4663-4676, 1999.

[3] ENVIRON, *User's Guide for Comprehensive Air Quality Model with Extensions (CAMx), version 4.00*, ENVIRON International Corporation, Novato, CA, 2004.

[4] Hubbard, M. C. and Cobourn, W. G., *Development of a Regression Model to Forecast Ground Level Ozone in Louisville, Kentucky*, Atmospheric Environment, 32, pp. 2637-2647, 1998.

[5] Planinsek, A. et al., *Onesnazenost zraka v Sloveniji v letu 2001 (Air pollution in Slovenia in 2001)*, Environmental Agency of the Republic of Slovenia, Ljubljana, 2000.

[6] Robnik Sikonja, M., *Properties and application of ReliefF heuristic in machine learning,* PhD Thesis, Faculty of Computer and Information Science, University of Ljubljana, 2001.

[7] Witten, I.H., E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers, 2000.

[8] Zabkar, J., Indihar A., Vrabec B., Petek M., Suc D., Bratko I., Merse J., Improving ALADIN's temperature forecast with machine learning. *Proc. of the 11th International Electrotechnical and Computer Science Conference*, Volume B, ed. B. Zajc, pp. 459-460, 2002.

Air Pollution XII, C. A. Brebbia (Editor)