Q^2 Prediction of Ozone Concentrations

Jure Žabkar [a,*] , Rahela Žabkar [b] , Daniel Vladušič [a] ,
Danijel Čemas [c] , Dorian Šuc [a] and Ivan Bratko [a]

[a] *Faculty of Computer and Information Science, Tržaška 25, 1000 Ljubljana, Slovenia*
[b] *Faculty of Mathematics and Physics, Jadranska 19, 1000 Ljubljana, Slovenia*
[c] *Environmental Agency of the Republic of Slovenia, Vojkova 1b, 1000 Ljubljana, Slovenia*

Abstract

We describe a case study in which we applied Q^2 learning (*qualitatively faithful quantitative learning*) to the analysis and prediction of ozone concentrations in the cities of Ljubljana and Nova Gorica, Slovenia. We used program QUIN to induce a qualitative model from numerical data that include the measurements of several meteorological and chemical variables. The resulting qualitative model consists of tree-structured monotonic qualitative constraints. We show how this model for Nova Gorica enables a nice interpretation of complex meteorological and chemical processes that affect the level of ozone concentration. In addition to inducing a qualitative model from data, we extended the qualitative model to also enable numerical prediction for both cities. In this case, we used in addition to measured data also data from the European meteorological prognostic model ALADIN which itself does not model pollutants. The results suggest that the qualitatively constrained numerical model tends to improve numerical prediction in comparison with some standard numerical learning methods.

Key words: Ozone concentration prediction model, Air pollution, Qualitative modelling

* Corresponding author - Phone: +386 1 4768 813; Fax: +386 1 4768 386
Email addresses: jure.zabkar@fri.uni-lj.si (Jure Žabkar),
rahela.zabkar@fmf.uni-lj.si (Rahela Žabkar),
daniel.vladusic@fri.uni-lj.si (Daniel Vladušič),
danijel.cemas@gov.si (Danijel Čemas), dorian.suc@fri.uni-lj.si
(Dorian Šuc), ivan.bratko@fri.uni-lj.si (Ivan Bratko).

1 Introduction

In this paper we present an application of Q^2 learning (*qualitatively faithful quantitative learning*, [Šuc et al., 2004]) to the analysis and prediction of ozone concentrations in the cities of Ljubljana and Nova Gorica, Slovenia. For Nova Gorica, we induced a qualitative model from numerical meteorological data (temperature, relative humidity, wind speed and direction, solar radiation, precipitation) and air quality measurements (O_3, NO, NO_2, CO). The purpose of such a model is to provide the experts with a relatively simple, interpretable model of the complex dynamics. Although the principles of ozone formation are in general known, our goal is to induce a qualitative model that would outline the prevailing local precursors. The induced qualitative model shows local influences that were previously unknown and are found interesting by experts. For both cities, in addition to inducing a qualitative model from data, we extended the qualitative model to also enable numerical prediction. In this case, available data included aforementioned measurements as well as predictions of the European meteorological prognostic model ALADIN [Bubnova et al., 1995] for the period from 2001 to 2003. The measurements and ALADIN data were provided by Environmental Agency of the Republic of Slovenia (ARSO). We compared the numerical accuracy of our Q^2 model to the accuracy of two other, standard numerical learning methods: linear regression (LR) and regression trees (M5), both implemented in Weka [Witten and Frank, 2000]. In addition to superior explanatory power, the Q^2 model also had better numerical accuracy, although the differences were not statistically significant. Numerical predictions are, by expert opinion, good enough to be used operationally. The main obstacle that prevents better numerical predictions is limited available data.

The processes that are involved in ozone formation are numerous and complex. Analytical models, such as CAMx [Environ, 2004], consist of systems of differential equations to capture the physics of the system, and include over a hundred chemical reaction equations to describe the chemical processes. The overall understanding of such complex models is difficult. But even if that is achieved, such models are usually not useful in practice for prediction because we can only use equations that include the independent variables that we can measure.

In order to protect human health and the environment, European Union legislation (Council Directive on ambient air quality assessment and management, 96/62/EC and Directive 2002/3/EC related to ozone in ambient air) requires that citizens are properly informed about concentrations of ozone and other air pollutants which, besides measuring, also includes forecasting pollutant concentrations. That enables concerned citizens, industrial organizations and local authorities to take action to reduce harmful emissions of ozone precursors. The forecast of daily maximum ozone concentration is required at 8 a.m. local time.

In section 2 we give an overview of the Q^2 learning method. In section 3, we de-

scribe the ozone domain, some background facts and motivation. The available data is described in detail in section 4. We present the results in section 5, assess what has been achieved, and discuss related and future work in section 6. In the appendix we give the details of several induced models.

2 Q^2 learning method

The Q^2 learning (*qualitatively faithful quantitative learning*, [Šuc et al., 2004]) is a machine learning method that, given a set of numerical examples S, induces a function f for numerical prediction. Each example in S consists of the values of a non-empty set of independent variables and a dependent (i.e. class) variable. Q^2 learning solves the learning problem in two stages, as shown in Fig. 1.

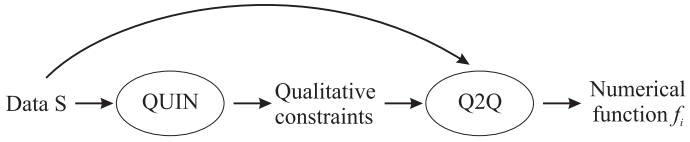

Fig. 1. A schema of Q^2 learning method.

In the first stage, it constructs qualitative constraints QC from the set of examples S. The qualitative constraints QC form a qualitative model that is by itself useful for the understanding and interpretation of the modelled domain. It is also used in the second stage of Q^2 learning, Q2Q (qualitative-to-quantitative transformation), in which a numerical function f is induced so that it respects the constraints QC and fits well the data S numerically. The resulting numerical function forms a quantitative model of the domain.

The two stages above can be carried out in various ways. In the application described in this paper we used program QUIN [Šuc, 2003] that induces qualitative constraints in the form of qualitative trees, and program QCGrid that for each leaf of the qualitative tree builds a piece-wise linear regression function that respects the qualitative constraints in the leaf.

In the following subsections we briefly describe both algorithms. More detailed explanation and a simple example are given in [Vladušič et al., 2005] (this volume).

2.1 Qualitative induction

QUIN (QUalitative INduction) is a learning program that looks for qualitative dependencies in numerical data. QUIN induces qualitative trees to express such de-

pendencies. The induction process is similar to the induction of decision trees [Breiman et al., 1984; Quinlan, 1992]. In choosing among attributes, QUIN uses a cost measure based on the minimum description length [Rissanen, 1978] principle. Another important difference between qualitative and decision trees is in the leaves. In a decision tree, the leaves are labeled with the values of the class variable, whereas in a qualitative tree the leaves are labeled with MQCs (monotonic qualitative constraints). MQCs are a kind of monotonicity constraints that are widely used in the field of qualitative reasoning [Forbus, 1984; Kuipers, 1994].

A monotonic qualitative constraint (MQC) $M^{s_1,\ldots,s_m}(x_1,\ldots,x_m)$ where $s_i \in \{+,-\}$ and $x_1,\ldots,x_m$ are independent variables, stands for an arbitrary function of m continuous variables that respects the qualitative constraints given by signs s_i. The qualitative constraint given by sign $s_i = +$ ($s_i = -$) requires that the function is strictly increasing (decreasing) in its dependence on the i-th variable while keeping the other variables constant. A MQC may be qualitatively ambiguous. Qualitative ambiguity occurs when the qualitative value of the function cannot be predicted (e.g. the qualitative change in $f = M^{+,-}(x_1, x_2)$ cannot be determined in the case of x_1 and x_2 both increasing). The degree of fit between the data and a MQC is evaluated by two measures: *qualitative consistency and* qualitative ambiguity. Qualitative consistency of a MQC is the percentage of the learning examples that are qualitatively consistent with the MQC. Qualitative ambiguity is the percentage of examples for which the MQC allows ambiguous predictions.

The QUIN algorithm has quite a high complexity. Finding the most consistent MQC in a leaf of a qualitative tree in a kind of greedy manner, requires the number of error-cost computations that is quadratic in the number of attributes. When building a qualitative tree, QUIN forms a MQC for both corresponding subtrees, for each possible split in every attribute. Therefore the time complexity is cubic in the number of attributes. Another factor that affects the time complexity of QUIN is the number of qualitative change vectors. The number of change vectors is quadratic in the number of examples. In the worst case, when many attributes are used, all the change vectors are processed. In cases with fewer attributes, a relatively small subset of change vectors suffices. Therefore QUIN's worst case time complexity is quadratic in the number of examples.

Empirical results [Šuc, 2003; Šuc et al., 2004] show that QUIN can handle noisy data and, at least in simple domains, produces qualitative trees that correspond to human intuition.

2.2 Qualitative-to-quantitative transformation with QCGrid

The QCGrid (Qualitatively Constrained Grid) algorithm is a regression algorithm that performs the Q2Q transformation shown in Fig. 1. For each leaf of the qualita-

tive tree, QCGrid takes the learning data and the MQC from the leaf as input, and induces a piece-wise linear function that respects given MQC. When constructing the grid, it first searches for a suitable grid spanned over the training data, using binary split search, commonly used in regression tree learning algorithms. The optimization problem of fitting the grid to the learning data while respecting the qualitative constraints is solved by quadratic programming. Numerical function f that is the output of Q2Q is defined by the functions induced in the leaves.

3 Ozone domain

It is well known that the amount of stratospheric ozone is decreasing in the last years. This phenomenon is known as the ozone hole. The protective layer of the stratospheric ozone makes the life on earth possible as it protects us from the harmful part of UV radiation. The problem of increasing tropospheric ozone is generaly less known. Ozone in the lower atmosphere (troposphere) has harmful effects on human health, vegetation and some materials. Even low ozone concentrations may cause chest pain, coughing, vomiting and throat irritation. Repeating exposure to higher concentrations may cause permanent damage to the lungs and influence bronchitis, asthma and heart diseases.

Tropospheric ozone is a secondary pollutant, formed by solar radiation in a series of photochemical reactions from nitrogen oxides (NO_x) and volatile organic compounds (VOC). They are released into the troposphere from a variety of biogenic and anthropogenic sources. Most of anthropogenic sources are emitted as results of the combustion of fossil fuels. The level of ozone concentration also depends on several physical and chemical processes and meteorological conditions. It is known that highest ozone concentrations appear in summer, at high temperatures and high level of solar radiation. It is also known that ozone concentrations in rural and elevated areas are typically twice as high as in urban areas. Concentrations of tropospheric ozone are known to have a daily cycle, similar to temperature's, with a maximum in the afternoon and a minimum in the early morning.

In Slovenia typical maximum ozone concentration during summer is between 200 and 230 mg/m^3. The information and alert thresholds that affect human health are 180 and 240 mg/m^3 per hour, respectively. In the capital, Ljubljana (LJ), they are only exceeded a few times per year. The small city of Nova Gorica (GO) in the western part of the country has on average higher levels of ozone concentration that also appear more often. High tropospheric ozone episodes in Slovenia are mainly due to the local sources (in LJ) and the long-range transport of ozone and its precursors (in GO), generally originating from Western Europe.

4 The learning data

Available data (Table 1) are meteorological and air quality measurements as well as predictions from the European meteorological prognostic model ALADIN. Because of the changes in the structure of the ALADIN model, we can only use its predictions for the years 2001 to 2003. Its values are available at 210 grid points over Slovenia with resolution of 11 km and time resolution of three hours. The measurements were taken from the same period with a lot of missing values. Meteorological and air quality measurements are measured half-hourly. The measuring tolerance is high and this prevents a potential improvement of prediction accuracy which has its error approximately in the order of the measurement error.

Table 1
Available data for analysis of ozone concentrations in Ljubljana and Nova Gorica. Meteorological and air quality measurements have time resolution of half an hour. Predictions of ALADIN model are available at 210 grid points with resolution of 11 km and time resolution of 3 hours. Ozone (O_3) is a dependent (i.e. class) variable.

Meteorological measurements	temperature	T
	relative humidity	RH
	wind speed	WS
	wind direction	WD
	solar radiation	S
	precipitation	P
Air quality measurements	nitrogen oxide	NO
	nitrogen dioxide	NO_2
	carbon oxide	CO
	ozone	O_3 (class variable)
Predictions of ALADIN model	temperature	T_{AL}
	solar radiation	S_{AL}
	precipitation	P_{AL}
	relative humidity	RH_{AL}

By expert opinion, the available data is deficient in many respects, namely the time period, the number of measured variables and the number of measurement stations. Various important measurements, such as $VOCs$, are not currently available and the process of acquiring them is underway.

The data was split at the very beginning into a learning set and a test set. The learning set was taken to include the data from 2001–2002 while data from 2003 was held out for testing.

The output data from ALADIN model are three dimensional fields of meteorological parameters. To the current stage of the project, only ground-level data was used. The values in model grid points present the average over the whole model grid cell and it is not possible to assess, within the model framework, a subgrid cell variation. For our use, we had to determine, from ALADIN's data, predictions at particular precise locations. When interpolating meteorological parameters in a selected point, for instance a meteorological station, from model output fields it is incorrect to assume that model outputs are the values in the centers of model grid cells. Instead of interpolation we decided to use stepwise linear regression method to build a linear regression model for each of meteorological parameters separately. With the stepwise method, regression model is built progressively. At each step, the independent variable which has the smallest probability of F (using F-test), is entered, but only if that probability is smaller than 0.05. Variables already in the regression equation are removed if their probability of F becomes larger than 0.1. The method terminates when no more variables are eligible for inclusion or removal.

No additional preprocessing is needed for meteorological and air quality measurements, since the predictions are made for the location of meteorological station.

The data from the ALADIN model were used with meteorological and air quality measurements to induce prediction models for Ljubljana and Nova Gorica. On the other hand, only half-hourly spaced meteorological and air quality measurements were used alone to induce a qualitative model for Nova Gorica, which was evaluated by a meteorologist and a chemist. The qualitative model was analyzed taking into account the principles used in the CAMx model.

5 Results

5.1 Qualitative model

The available data for qualitative model building was a set of meteorological and air quality half-hourly measurements (see Table 1) in Nova Gorica. Nova Gorica was chosen because the measurements showed higher levels of ozone concentrations and more interesting dynamics. Namely, the experts expected that the model would highly depend on wind direction because wind is known to be the reason for high level concentrations. To enable a reference to the time of the day, the attribute t was included as an index of the begining of each half-hourly period, i.e. $t \in [0, 47]$. QUIN cannot efficiently handle large learning sets, neither in terms of examples nor attributes. The complexity of QUIN algorithm is explained in section 2.1. The learning set was therefore sampled taking every fourth example.

Attribute selection was performed manually by experts. The reasons for not taking other possible feature selection methods are the following. Choosing the attributes from the model tree or a feature selection method could be inappropriate for QUIN which selects only the attributes that are important for its purpose. Model trees, for example, do not look for qualitative patterns in data and choose attributes by other criteria. On the other hand, the experts know the nature of the data well and have a lot of background knowledge which gives better chances of selecting good attributes.

Among the available attributes, wind speed and precipitation were omitted due to missing data. Scatter plot of wind direction was analysed but surprisingly, no clear dependence on wind can be found, which can, by one interpretation, indicate that local sources of ozone precursors in the city have an important role in ozone formation. It also turns out that the wind direction measurements themselves cannot indicate the information that the human expert can conclude from other sources, such as Italian air pollution cadastral registers etc., that were not at our disposal.

Temperature and solar radiation which are among most important attributes are highly correlated. So we decided to omit one of them and the solar radiation was chosen to stay because it directly participates in chemical reaction of ozone formation and is therefore more important. Relative humidity was accepted to substitute for precipitation. Among air quality measurements, CO was not selected because it does not directly participate in ozone formation but influences it through $VOCs$, which were not among available data. Nitrogen oxides, NO and NO_2, are highly correlated but since NO_2 is better correlated to ozone, NO_2 was selected. To capture the daily dynamics, time of the day t was also selected. That makes a set of four attributes (see Table 2), altogether, passed to QUIN.

Table 2
The set of selected attributes for qualitative induction. The process of attribute selection is described in text.

relative humidity	H
solar radiation	S
index of half-hour interval	t
nitrogen dioxide concentration	NO_2

The output models were evaluated by coverage and qualitative ambiguity which QUIN calculates. The resulting qualitative tree is shown in Fig. 2. It is a qualitative model of the dynamics in the ozone formation process. This model was presented to an expert chemist and a meteorologist for evaluation. Their interpretation is given later in this section.

The model is read as follows: a left branch from the node is chosen when the node

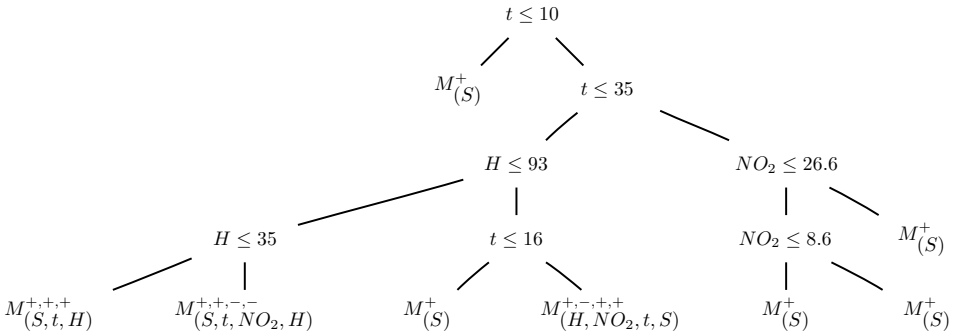

Fig. 2. The qualitative model of the dynamics in the ozone formation process, induced by program QUIN for the city of Nova Gorica.. Attributes: relative humidity (H), solar radiation (S), index of half-hour interval (t), nitrogen dioxide concentration (NO_2). The meaning of the formulae in the leaves are explained in text.

condition is satisfied, otherwise the right branch is chosen. This process continues from the root of the tree to one of the leaves. In a leaf, MQC defines the qualitative dependency between the class variable, O_3, and the attributes mentioned in the MQC. For example, the leaf $M^{(+,+,-,-)}_{S,t,NO_2,H}$ means that the ozone concentration is increasing with increasing solar radiation and time. On the other hand, ozone is decreasing while nitrogen dioxide and relative humidity are increasing. Of course, this is only limited to $t \in [10, 35)$ and $H \in [35, 93)$, as the path from the root to the leaf implies.

The following is experts' interpretation of the induced qualitative model. The split in the root of the tree is made on $t \leq 10$ which means 5 a.m. and clearly separates the dynamics at night and day. The monotonic qualitative constraint $M^{+}_{(S)}$ may seem disturbing since there is no solar radiation during the night, but the analysis of the examples in the leaf shows slightly increasing dependence, presumably in summer days.

In the right subtree, there is a split on $t \leq 35$, (5.30 p.m.). This break point separates the periods of increasing/decreasing dependence regarding t. The ozone concentration grows with t, i.e. $O_3 = M^{+}_{(t)}$, until 5.30 p.m. The production of ozone is higher than consumption. Although it has been generally known to happen in the late afternoon, there are two possible explanations, not excluding each other. Since we are modeling the system in the city it is very likely that the amount of traffic, which is increased in the afternoon when people go home from work, influences this by increasing NO_x emissions. These are known to cause the reactions with O_3, decreasing the level of O_3 concentration. The second explanation says that solar radiation is decreasing, resulting in O_3 decreasing. The right subtree of the $t \leq 35$ node includes two splits on NO_2. The value of the upper split separates the space to higher and lower NO_2 concentrations, while the split on $NO_2 \leq 8.6$ further separates low and average concentrations. Obviously, MQCs are the same but the regression functions in the leaves differ by slope which is the reason for three leaves instead of one. This is a concequence of highly non-linear chemical processes. Finlayson-

Pitts Finlayson-Pitts and Pitts [2000] discusses the non-linearity of the dependence of $O_3(NO_x, VOC)$ while the qualitative constraints are the same as in our model.

The left subtree of $t \leq 35$ demands a meteorological explanation. The space is nicely separated by relative humidity (H) to dry ($H \leq 35$), average wet ($35 < H \leq 93$) and precipitation ($H > 93$). The qualitative dependance in MQCs is in accordance with the experts' previous knowledge. Only the $M^{+}_{(H)}$ in the left leaf of the subtree is not very logical but the analysis again shows that the regression slope is very low, almost 0. In fact, this dependence could easily be removed from the tree in the pruning process, if necessary.

Linear regression models and model trees were also built for comparison with the qualitative model. Both methods were used to build two models – one on all the attributes and one on the same subset that was passed to QUIN. The models are shown in appendix (A.1.1 and A.1.2).

The experts' interpretation of LR models is that the signs of the coefficients give reasonable qualitative dependencies, but the daily dynamics is not captured as it is in QUIN's model.

Since model trees induce a non-linear model, we expected better results. Daily dynamics of ozone formation is badly captured compared to QUIN's model. It does not outline the expected maximum in the afternoon nor does it separate the night time processes and the minimum in the early morning. The nodes with relative humidity H split on values that cannot be so clearly explained as in the QUIN's case which clearly separates the examples in three groups, as shown above. QUIN's model is more intuitive and in general easier to explain.

5.2 Numerical predictions

The attributes used in the learning process were built from the ALADIN predictions at the model grid points, neighbouring the meteorological station point in both cities, as described in section 4. At that point, the meteorological measurements were performed. Because of different time resolution in measurements and ALADIN's predictions, experts prepared a set of seven attributes that could be used together in the learning process. These are:

- $MAXCO$ (max. concentration of CO in the last 36 hours before the prediction is made)
- $MAXNO$ (max. concentration of NO in the last 36 hours before the prediction is made)
- $MAXNO2$ (max. concentration of NO_2 in the last 36 hours before the prediction is made)
- $Tavg915GO/LJ$ (avg. of the ALADIN's predictions of temperature from 9 a.m.

to 3 p.m.) in Nova Gorica (GO) and Ljubljana (LJ)
- $Ssum015GO/LJ$ (the sum of ALADIN's predictions of solar radiation from midnight to 3 p.m. for each city)
- $Pavg015GO/LJ$ (avg. of the ALADIN's predictions of precipitation from midnight to 3 p.m.) in Nova Gorica (GO) and Ljubljana (LJ)
- $RHavg015GO/LJ$ (avg. of ALADIN's predictions of relative humidity from midnight to 3 p.m. for each city)

A subset of three ($MAXNO$, $Tavg915GO/LJ$, $Ssum015GO/LJ$) were selected for comparison between Q^2 learning, linear regression and model trees. The last two methods were also run on all seven attributes. These models are shown in appendix (B). Experts' explanation of them is presented here.

The qualitative trees for both cities are shown in Fig. 3. The structure of both trees is very similar and schematically shows that the ozone concentration is positively correlated to the temperature and solar radiation while negatively correlated to the concentration of NO. The concentration of NO in the roots of the qualitative trees evaluates the dominating mechanisms of the ozone cycle. Higher NO concentrations occur during night time hours with low ozone concentration (right branch). On the contrary, high ozone concentration as a result of photochemical formation prevents high NO concentration (left branch).

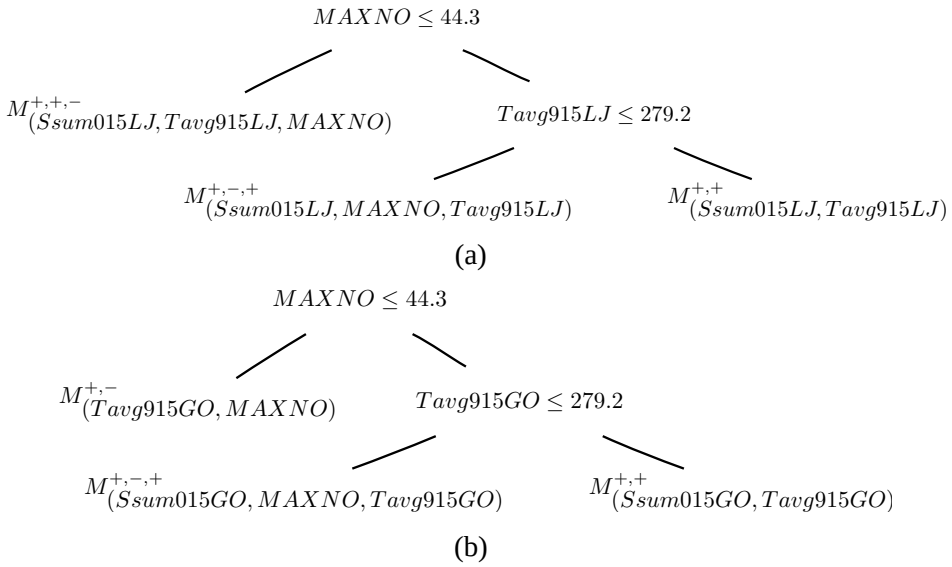

Fig. 3. Qualitative models for Ljubljana (a) and Nova Gorica (b), used for numerical prediction of ozone concentration.

Temperature and solar radiation are statistically highly correlated so model can choose any of the two variables to describe the presence and intensity of photochemical reactions in the atmosphere. During night time and cloudy days without solar radiation, temperatures are usually lower. In our case temperature showed bet-

ter statistical correlation to ozone concentration which resulted in the second split node. We must stress here the fact that unlike in section 5.1, ALADIN's predictions of temperature and solar radiation, not the measurements, are used. ALADIN is known for inaccurate predictions of solar radiation, which makes the temperature a better selection in this case. The trees prove that highest ozone concentrations occur during daytime in summer in hot sunny and dry weather.

Numerical accuracy of induced models is compared to LR and M5. Table 3 shows the RMSE achieved on the test set. Q^2 turns out to be superior to LR and M5 although not significantly. The prunning parameter of model trees was set to the best value according to internal 4-fold cross validation performed on the learning set. The same value was used for prediction on the test set.

Table 3
Comparison of the numerical accuracy of the competing methods according to root mean squared error (RMSE). The same set of attributes are used to build the models.

RMSE on test set	Linear regression	Model tree	Q^2
Ljubljana	21.63	22.94	19.9
Nova Gorica	20.98	19.92	19.81

As with the qualitative model in section 5.1, explanation of linear regression models cannot go any further beyond arguing that the signs of the coefficients are as expected. Model trees for Nova Gorica are quite complex, having 13 leaves (3 attributes used) and 10 leaves (all attributes used). Both model trees for Nova Gorica have a root that splits on $Ssum015GO$ at the same value. It is interesting that QUIN does not split on this attribute at all. On the smaller set of attributes 6 equations (out of 13) have a positive coefficient for $MAXNO$, which is strange as it is expected the ozone to be decreasing while NO is increasing. The model tree using all the attributes is smaller, but has a larger error on the test set. Relative humidity comes very near the root but its splitting value is not important since it is in the average wet and does not indicate precipitation. As expected $Pavg015GO$ only comes out in the leaves' equations because ALADIN's predictions of precipitation are not very good. Similar explanations are valid for the models in Ljubljana since the models are very similar. The only exception is a model tree for Ljubljana using all atributes, which has only the root node with solar radiation attribute in it. The performance of each model is worse than in the case of Q^2 learning.

6 Discussion and related work

A qualitative model was induced from available measurement data of meteorological and air quality variables. The purpose of this model is to describe the complex process of ozone formation. The qualitative model was evaluated from several perspectives - by expert meteorologist, expert chemist and compared to models in the

literature ([Finlayson-Pitts and Pitts, 2000]). The experts found the models explanatory and consistent with their understanding of the relevant processes.

Separate qualitative models were induced for ozone process analysis and prediction of ozone concentration in the city of Ljubljana. ALADIN model forecasts were used as attributes for this purpose. The accuracy of numerical predictions was compared to linear regression and regression trees. Q^2 learning gave (not significantly) better results. The experts found prediction models operationally useful and conclude that the prediction error is in the order of the measurement error.

Till now in the project, only ground-level data has been used. Further work should include the data from higher levels of the atmosphere. We expect improvement from that. ALADIN's model forecasts of wind speed and direction are much better at the higher levels of the atmosphere. By expert opinion, the information of the processes at higher levels could improve the predictions of ozone concentrations at ground-level.

Finally we here mention some of the related work on ozone modeling, although none of it involves qualitative models. Several statistical models [Jenkin and Clemitshaw, 2000; Athanasiadis et al., 2003; Canu and Rakotomamonjy, 2001; Hubbard and Cobourn, 1998; Cobourn and Hubbard, 1999] have been built in order to predict the ozone (O_3) concentration. On the other hand, Eulerian photochemical dispersion models, such as CAMx (Comprehensive Air quality Model with extensions) [Environ, 2004], are being developed. CAMx simulates the emission, dispersion, chemical reaction and removal of pollutants in the troposphere. The Eulerian continuity equation describes the time dependency of the concentration within each grid cell volume where specific physical and chemical processes are operating. Details on chemical processes can be found in [Finlayson-Pitts and Pitts, 2000]. The CAMx model has not been in operational use so far and no numerical prediction from this model is available for comparative study.

Acknowledgements

We are grateful to dr. Matevž Pompe from the Faculty of Chemistry and Chemical Engineering, University of Ljubljana, for evaluation of our qualitative model from the chemical point of view.

References

Athanasiadis, I. N., Kaburlasos, V. G., Mitkas, P. A., Petridis, V., 2003. Applying machine learning techniques on air quality data for real-time decision support.

In: First International NAISO Symposium on Information Technologies in Environmental Engineering, ITEE 2003.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees.

Bubnova, R., Hello, G., Benard, P., Geleyn, J.-F., 1995. Integration of the fully-elastic equations cast in the hydrostatic pressure terrain-following coordinate in the framework of the arpege/aladin nwp system. Monthly Weather Review 123 (2), 515 – 535.

Canu, S., Rakotomamonjy, A., 2001. Ozone peak and pollution forecasting using support vectors. In: International Federation of Ambulatory Care, IFAC 2001.

Cobourn, W. G., Hubbard, M. C., 1999. An enhanced ozone forecasting model using air mass trajectory analysis. Atmospheric Environment 33 (4), 4663–4676.

Environ, 2004. Comprehensive air quality model with extensions, version 4, user's guide. Tech. rep., Environ International Corporation, www.camx.com, Novato, California.

Finlayson-Pitts, B. J., Pitts, J. N., 2000. Chemistry of the Upper and Lower Atmosphere. Academic Press.

Forbus, K., 1984. Qualitative process theory. Artificial Intelligence 24, 85–168.

Hubbard, M. C., Cobourn, W. G., 1998. Development of a regression model to forecast ground level ozone in louisville, kentucky. Atmospheric Environment 32 (4), 2637–2647.

Jenkin, M. E., Clemitshaw, K. C., 2000. Ozone and other secondary photochemical pollutants: Chemical processes governing their formation in the planetary boundary layer. Atmospheric Environment.

Kuipers, B., 1994. Qualitative Reasoning: Modeling and Simulation with Incomplete Knowledge. MIT Press, Massachusetts.

Quinlan, J., 1992. Learning with continuous classes. In: Proceedings of the 5th Australian Joint Conference on Artificial Intelligence. pp. 343–348.

Rissanen, J., 1978. Modelling by shortest data description. Automatica 14, 465–471.

Vladušič, D., Kompare, B., Bratko, I., 2005. Modelling Lake Glumsø with Q^2 Learning. Ecological Modelling 1 (1), 1–2.

Šuc, D., 2003. Machine Reconstruction of Human Control Strategies. Vol. 99 of Frontiers in Artificial Intelligence and Applications. IOS Press, Amsterdam, The Netherlands.

Šuc, D., Vladušič, D., Bratko, I., 2004. Qualitatively faithful quantitative prediction. Artificial Intelligence Journal 158, 189–214.

Witten, I., Frank, E., 2000. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco.

A Linear regression and M5 models for analysis of qualitative dependencies

A.1 Models for analysis of qualitative dependencies

A.1.1 Linear regression models

LR models for analysis of qualitative dependencies are built on two sets of attributes for comparison to QUIN's model. In one LR model, all atributes are used while the other uses the attributes that were selected by experts and used in the learning process by QUIN. The attributes are normalized so that coefficients can be compared. The sign of the coefficient implies the qualitative dependency.

$$\begin{aligned} O_3 = & +0.0462\ t + \\ & +0.4908\ T + \\ & -0.2744\ H + \\ & +0.0821\ S + \\ & -0.1147\ CO + \\ & +0.0361\ NO + \\ & -0.3924\ NO_2 + \\ & +0.3099 \end{aligned}$$

$$\begin{aligned} O_3 = & +0.0879\ t + \\ & -0.2692\ H + \\ & +0.3222\ S + \\ & -0.5546\ NO_2 + \\ & +0.5307 \end{aligned}$$

A.1.2 Model trees

Below is a model tree for analysis of qualitative dependencies, built by M5 using the set of attributes that were selected by experts and used in the learning process by QUIN. The attributes are normalized so that coefficients can be compared. The sign of the coefficient implies the qualitative dependency.

$H \leq 0.636$:
| $S \leq 0.478$:
| | $t \leq 0.543 : LM1$
| | $t > 0.543$:

$$
\begin{array}{l}
| \quad | \quad | \quad S \leq 0.128 : LM2 \\
| \quad | \quad | \quad S > 0.128 : LM3 \\
| \quad S > 0.478 : LM4 \\
H > 0.636 : \\
| \quad NO_2 \leq 0.283 : \\
| \quad | \quad NO_2 \leq 0.115 : LM5 \\
| \quad | \quad NO_2 > 0.115 : \\
| \quad | \quad | \quad H \leq 0.833 : \\
| \quad | \quad | \quad | \quad t \leq 0.436 : LM6 \\
| \quad | \quad | \quad | \quad t > 0.436 : LM7 \\
| \quad | \quad | \quad H > 0.833 : \\
| \quad | \quad | \quad | \quad H \leq 0.981 : LM8 \\
| \quad | \quad | \quad | \quad H > 0.981 : LM9 \\
| \quad NO_2 > 0.283 : LM10
\end{array}
$$

$$
\begin{aligned}
LM1 : O_3 &= 0.353 + 0.0103t - 0.00781H + 0.239S - 0.395NO_2 \\
LM2 : O_3 &= -0.042 + 0.568t - 0.00579H + 1.95S - 0.442NO_2 \\
LM3 : O_3 &= -0.772 + 1.71t - 0.00579H + 0.794S - 0.0379NO_2 \\
LM4 : O_3 &= -0.0785 + 0.798t - 0.00224H + 0.432S - 0.0198NO_2 \\
LM5 : O_3 &= 0.462 + 0.107t - 0.00792H + 0.009S - 1.86NO_2 \\
LM6 : O_3 &= 0.261 + 0.00332t - 0.0115H + 0.011S - 0.0295NO_2 \\
LM7 : O_3 &= 0.719 + 0.00332t - 0.553H + 0.011S - 0.0295NO_2 \\
LM8 : O_3 &= 0.722 + 0.0541t - 0.499H + 0.00837S - 0.569NO_2 \\
LM9 : O_3 &= -8.4 + 0.00635t + 8.7H + 0.00837S - 0.0593NO_2 \\
LM10 : O_3 &= 0.484 + 0.054t - 0.301H + 0.378S - 0.359NO_2
\end{aligned}
$$

Model tree (all attributes, normalized to enable coefficient comparison)

$$
\begin{array}{l}
T \leq 0.552 : \\
| \quad NO \leq 0.0462 : \\
| \quad | \quad NO_2 \leq 0.158 : LM1 \\
| \quad | \quad NO_2 > 0.158 : LM2 \\
| \quad NO > 0.0462 : \\
| \quad | \quad H \leq 0.759 : \\
| \quad | \quad H > 0.759 : \\
| \quad | \quad | \quad NO \leq 0.107 : LM4 \\
| \quad | \quad | \quad NO > 0.107 : LM5 \\
T > 0.552 :
\end{array}
$$

$|\quad T \leq 0.751:$
$|\quad|\quad NO \leq 0.0173:$
$|\quad|\quad|\quad H \leq 0.685 : LM6$
$|\quad|\quad|\quad H > 0.685:$
$|\quad|\quad|\quad|\quad S \leq 0.00192 : LM7$
$|\quad|\quad|\quad|\quad S > 0.00192 : LM8$
$|\quad|\quad NO > 0.0173 : LM9$
$|\quad T > 0.751:$
$|\quad|\quad NO \leq 0.014 : LM10$
$|\quad|\quad NO > 0.014 : LM11$

$$
\begin{aligned}
LM1 : O_3 = & 0.31 + 8.4e-4t + 0.351T - 0.196H + 0.181S - 0.0094CO \\
& -6.67NO - 0.00371NO_2 \\
LM2 : O_3 = & 0.418 + 0.0466t + 0.115T - 0.245H + 0.295S - 0.426CO \\
& -2.67NO - 0.00371NO_2 \\
LM3 : O_3 = & 0.369 + 4.3e-4t + 0.004T - 0.273H + 0.00925S - 0.0018CO \\
& -0.00479NO - 0.183NO_2 \\
LM4 : O_3 = & 0.111 + 4.3e-4t + 0.004T - 0.00826H + 0.315S - 0.233CO \\
& -0.491NO + 0.0959NO_2 \\
LM5 : O_3 = & 0.0955 + 4.3e-4t + 0.004T - 0.0751H + 0.129S - 0.0314CO \\
& -0.00129NO + 0.03NO_2 \\
LM6 : O_3 = & 0.0219 + 0.1t + 0.914T - 0.234H + 0.00626S + 0.0075CO \\
& -8.3NO - 0.0105NO_2 \\
LM7 : O_3 = & -0.0295 + 0.00474t + 0.642T - 0.00793H + 0.0358S + 0.0075CO \\
& -10.2NO - 0.0105NO_2 \\
LM8 : O_3 = & 0.432 + 0.0697t + 0.0297T - 0.00793H + 0.465S + 0.0075CO \\
& -16.9NO - 0.0105NO_2 \\
LM9 : O_3 = & 0.016 + 0.0719t + 0.528T - 0.154H + 0.131S + 0.0142CO \\
& -0.941NO - 0.0178NO_2 \\
LM10 : O_3 = & -0.0517 + 0.252t + 0.618T - 0.00543H + 0.25S + 0.00876CO \\
& -24.4NO + 1.03NO_2 \\
LM11 : O_3 = & -0.363 + 0.253t + 0.887T - 0.00543H + 0.182S + 0.00876CO \\
& -2.37NO - 0.0117NO_2
\end{aligned}
$$

B Linear regression and M5 models for numerical prediction

B.1 The models for prediction in Ljubljana

B.1.1 Linear regression equation using all attributes

$$\begin{aligned} O_3 = & -2.5454\ MAXCO + \\ & -0.1189\ MAXNO + \\ & 0.1335\ MAXNO2 + \\ & 0.0155\ Ssum015LJ + \\ & 2.1441\ Tavg915LJ + \\ & -0.0047\ Pavg015LJ + \\ & -78.7417\ RHavg015LJ + \\ & -20.8075 \end{aligned}$$

B.1.2 Linear regression equation using the same attributes as used by QUIN

$$\begin{aligned} O_3 = & -0.1009\ MAXNO + \\ & 0.0301\ Ssum015LJ + \\ & 1.7136\ Tavg915LJ + \\ & -438.953 \end{aligned}$$

B.1.3 Model tree using all attributes

$$\begin{aligned} & Ssum015LJ \leq 809 : LM1 \\ & Ssum015LJ > 809 : LM2 \end{aligned}$$

$$\begin{aligned} LM1 : O_3 = & -550 - 0.239MAXCO - 0.0872MAXNO + 0.012MAXNO2 \\ & +0.00156Ssum015LJ + 2.63Tavg915LJ - 4.74e-4Pavg015LJ \\ & -113RHavg015LJ \\ LM2 : O_3 = & 367 - 0.163MAXCO - 0.00742MAXNO + 0.00817MAXNO2 \\ & +0.0264Ssum015LJ + 1.43Tavg915LJ - 0.00722Pavg015LJ \\ & -4.98RHavg015LJ \end{aligned}$$

B.1.4 Model tree using the same attributes as used by QUIN

$$
\begin{aligned}
&Ssum015LJ \leq 809 : LM1\\
&Ssum015LJ > 809 :\\
&\quad Tavg915LJ \leq 295 :\\
&\quad\quad Ssum015LJ \leq 1140 :\\
&\quad\quad\quad MAXNO \leq 13 : LM2\\
&\quad\quad\quad MAXNO > 13 : LM3\\
&\quad\quad Ssum015LJ > 1140 : LM4\\
&\quad Tavg915LJ > 295 :\\
&\quad\quad Ssum015LJ \leq 1940 : LM5\\
&\quad\quad Ssum015LJ > 1940 :\\
&\quad\quad\quad Ssum015LJ \leq 2130 : LM6\\
&\quad\quad\quad Ssum015LJ > 2130 : LM7
\end{aligned}
$$

$$
\begin{aligned}
LM1 &: O_3 = -333 - 0.0921 MAXNO + 0.0333 Ssum015LJ + 1.31 Tavg915LJ\\
LM2 &: O_3 = -89.4 - 0.0751 MAXNO + 0.00891 Ssum015LJ + 0.599 Tavg915LJ\\
LM3 &: O_3 = -52.2 - 0.0472 MAXNO + 0.00891 Ssum015LJ + 0.434 Tavg915LJ\\
LM4 &: O_3 = -3.68 - 0.0148 MAXNO + 0.0248 Ssum015LJ + 0.243 Tavg915LJ\\
LM5 &: O_3 = -91.7 - 0.02 MAXNO + 0.00538 Ssum015LJ + 0.677 Tavg915LJ\\
LM6 &: O_3 = -116 - 0.02 MAXNO - 0.0167 Ssum015LJ + 0.955 Tavg915LJ\\
LM7 &: O_3 = -122 - 0.02 MAXNO - 0.0167 Ssum015LJ + 0.955 Tavg915LJ
\end{aligned}
\tag{B.1}
$$

B.2 The models for prediction in Nova Gorica

B.2.1 Linear regression equation using all attributes

$$
\begin{aligned}
O3 = &-5.8999\ MAXCO +\\
&-0.0525\ MAXNO +\\
&0.2651\ MAXNO2 +\\
&0.0222\ Ssum015GO +\\
&1.7101\ Tavg915GO +\\
&-0.0039\ Pavg015GO +\\
&-44.3625\ RHavg015GO +\\
&-8.3602
\end{aligned}
$$

B.2.2 Linear regression equation using the same attributes as used by QUIN

$$
\begin{aligned}
O3 = & -0.0531\ MAXNO + \\
& 0.0328\ Ssum015GO + \\
& 1.2703\ Tavg915GO + \\
& -309.631
\end{aligned}
$$

B.2.3 Model tree using the same attributes as used by QUIN

$Ssum015GO \leq 825$:
| $MAXNO \leq 25.9$:
| | $Tavg915GO \leq 290$:
| | | $Ssum015GO \leq 518 : LM1$
| | | $Ssum015GO > 518 : LM2$
| | $Tavg915GO > 290 : LM3$
| $MAXNO > 25.9$:
| | $Tavg915GO \leq 285$:
| | | $Ssum015GO \leq 472$:
| | | | $MAXNO \leq 167 : LM4$
| | | | $MAXNO > 167 : LM5$
| | | $Ssum015GO > 472 : LM6$
| | $Tavg915GO > 285$:
| | | $MAXNO \leq 47.8$:
| | | | $Tavg915GO \leq 292 : LM7$
| | | | $Tavg915GO > 292 : LM8$
| | | $MAXNO > 47.8$:
| | | | $Ssum015GO \leq 702 : LM9$
| | | | $Ssum015GO > 702 : LM10$
$Ssum015GO > 825$:
| $Ssum015GO \leq 1560$:
| | $Ssum015GO \leq 1100 : LM11$
| | $Ssum015GO > 1100 : LM12$
| $Ssum015GO > 1560 : LM13$

$$
\begin{aligned}
LM1 &: O_3 = -222 - 0.0242 MAXNO + 0.00304 Ssum015GO + 1.04 Tavg915GO \\
LM2 &: O_3 = -223 + 0.142 MAXNO + 0.00304 Ssum015GO + 1.04 Tavg915GO \\
LM3 &: O_3 = -328 + 0.261 MAXNO + 0.00304 Ssum015GO + 1.43 Tavg915GO \\
LM4 &: O_3 = -27.3 - 0.0275 MAXNO + 0.00776 Ssum015GO + 0.274 Tavg915GO
\end{aligned}
$$

$$LM5 : O_3 = -33.4 - 0.0257MAXNO + 0.00776Ssum015GO + 0.274Tavg915GO$$
$$LM6 : O_3 = 0.899 - 0.0117MAXNO - 0.0412Ssum015GO + 0.274Tavg915GO$$
$$LM7 : O_3 = -50.3 + 0.0537MAXNO + 0.0403Ssum015GO + 0.274Tavg915GO$$
$$LM8 : O_3 = -50.8 + 0.0537MAXNO + 0.0403Ssum015GO + 0.274Tavg915GO$$
$$LM9 : O_3 = -230 + 0.0782MAXNO + 0.0413Ssum015GO + 0.893Tavg915GO$$
$$LM10 : O_3 = -189 + 0.0453MAXNO + 0.0393Ssum015GO + 0.778Tavg915GO$$
$$LM11 : O_3 = -44.2 - 0.0037MAXNO + 0.0101Ssum015GO + 0.432Tavg915GO$$
$$LM12 : O_3 = -308 - 0.0037MAXNO + 0.00857Ssum015GO + 1.38Tavg915GO$$
$$LM13 : O_3 = -277 - 0.0037MAXNO + 0.00753Ssum015GO + 1.32Tavg915GO$$

B.2.4 Model tree using all attributes

$Ssum015GO \leq 825$:
| $RHavg015GO \leq 0.691 : LM1$
| $RHavg015GO > 0.691$:
| | $Tavg915GO \leq 285$:
| | | $MAXNO2 \leq 51.4 : LM2$
| | | $MAXNO2 > 51.4 : LM3$
| | $Tavg915GO > 285$:
| | | $MAXNO \leq 29 : LM4$
| | | $MAXNO > 29$:
| | | | $MAXNO \leq 47.8$:
| | | | | $Tavg915GO \leq 292 : LM5$
| | | | | $Tavg915GO > 292 : LM6$
| | | | $MAXNO > 47.8$:
| | | | | $Ssum015GO \leq 702 : LM7$
| | | | | $Ssum015GO > 702 : LM8$
$Ssum015GO > 825$:
| $Ssum015GO \leq 1560 : LM9$
| $Ssum015GO > 1560 : LM10$

$$LM1 : O_3 = 7.11 - 1.38MAXCO - 0.0647MAXNO + 0.15MAXNO2 + 0.0488Ssum015GO + 0.603Tavg915GO - 0.00123Pavg015GO - 21.7RHavg015GO$$
$$LM2 : O_3 = 339 - 1.21MAXCO - 0.00505MAXNO - 0.239MAXNO2 + 0.00201Ssum015GO + 1.01Tavg915GO - 0.0054Pavg015GO - 18.6RHavg015GO$$
$$LM3 : O_3 = 858 - 1.21MAXCO - 0.00505MAXNO - 0.155MAXNO2 + 0.00201Ssum015GO + 1.01Tavg915GO - 0.0103Pavg015GO$$

$$
\begin{aligned}
& -77RHavg015GO \\
LM4 : O_3 = & -272 - 1.21MAXCO + 0.306MAXNO - 0.0817MAXNO2 \\
& +0.00201Ssum015GO + 2.2Tavg915GO - 0.00266Pavg015GO \\
& -18.6RHavg015GO \\
LM5 : O_3 = & 1.15 - 1.21MAXCO + 0.0375MAXNO - 0.0817MAXNO2 \\
& +0.0211Ssum015GO + 1.17Tavg915GO - 0.00266Pavg015GO \\
& -18.6RHavg015GO \\
LM6 : O_3 = & 0.829 - 1.21MAXCO + 0.0375MAXNO - 0.0817MAXNO2 \\
& +0.0211Ssum015GO + 1.17Tavg915GO - 0.00266Pavg015GO \\
& -18.6RHavg015GO \\
LM7 : O_3 = & -202 - 1.21MAXCO + 0.0371MAXNO - 0.0817MAXNO2 \\
& +0.0251Ssum015GO + 1.87Tavg915GO - 0.00266Pavg015GO \\
& -18.6RHavg015GO \\
LM8 : O_3 = & -177 - 1.21MAXCO + 0.041MAXNO - 0.0817MAXNO2 \\
& +0.023Ssum015GO + 1.79Tavg915GO - 0.00266Pavg015GO \\
& -18.6RHavg015GO \\
LM9 : O_3 = & -324 - 0.373MAXCO - 0.0948MAXNO + 0.337MAXNO2 \\
& +0.0268Ssum015GO + 1.42Tavg915GO - 2.87e-4Pavg015GO \\
& -3.25RHavg015GO \\
LM10 : O_3 = & -988 - 7.47MAXCO - 0.464MAXNO + 0.683MAXNO2 \\
& +0.00662Ssum015GO + 1.17Tavg915GO + 0.00706Pavg015GO \\
& +67.8RHavg015GO
\end{aligned}
$$