Using Qualitative Constraints In Ozone Prediction

Jure Žabkar[1], Daniel Vladušič[1], Rahela Žabkar[2], Danijel Čemas[3], Dorian Šuc[1*], and Ivan Bratko[1]
[1]Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia
{jure.zabkar,daniel.vladusic,dorian.suc,ivan.bratko}@fri.uni-lj.si
[2]Faculty of Mathematics and Physics, Jadranska 19, 1000 Ljubljana, Slovenia
[3]Environmental Agency of the Republic of Slovenia, Vojkova 1b, 1000 Ljubljana, Slovenia
[*]current address: National ICT Australia, University of New South Wales, Sydney, NSW Australia

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. For Ljubljana, in addition to inducing a qualitative model from data, we extended the qualitative model to also enable numerical prediction. 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. Program QCGrid was used to induce a numerical prediction model which respects the constraints in the qualitative model and fits the data well. We show that qualitatively constrained numerical model improves numerical prediction in comparison with some standard numerical learning methods.

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. 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. For Ljubljana, 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 [Aladin, 1997] for the period from 2001 to 2003. The measurements and ALADIN data were provided by Environmental Agency of the Republic of Slovenia (ARSO). For the Ljubljana qualitative model, we performed a qualitative-to-quantitative transformation, discussed later, to induce a numerical prediction model. The advantage of Q^2 learning, used here, is in its paying attention to the qualitative correctness of induced numerical models. 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, although the differences in accuracy were not statistically significant. Numerical predictions are, by expert opinion, good enough to be used operationally.

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 100 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 section 2, we describe the ozone domain, some background facts and motivation. We give an overview of Q^2 learning method in section 3. The available data is described in detail in section 4. We present the results in section 5, assess what has been achieved, and discuss future work in section 6.

2 The Ozone problem

Meteorological and chemical processes that affect the level of ozone concentration are very complex. Ozone (O_3) in the lower atmosphere (troposphere) has harmful effects on vegetation and human health. 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 O_3 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. The highest ozone concentrations occur in summer, with a maximum in the afternoon and a minimum in the early morning [A. Planinšek, 2000]. Nitrogen oxides and VOCs (volatile organic compounds) 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. Ozone concentration in rural and elevated areas is typically twice as high as in urban areas.

Ozone concentrations are regularly monitored, and, according to European regulations, environmental agencies have to provide short term predictions. In this paper we apply the Q^2 approach to machine learning to induce a qualitative and quantitative model for such predictions.

3 Q^2 learning method

The learning problem addressed by the Q^2 method is as follows. Given is a set of numerical examples S, (observations, measurements), where each example consists of the values of a set of independent variables and a set of dependent variables. The problem is to find a numerical function f_i for each dependent variable, for predicting the value of the i-th dependent variable given the values of the independent variables. Q^2 learning solves this problem in two stages (Fig. 1): (1) Construct qualitative constraints QC that hold in the modeled domain, and (2) Construct numerical functions f_i so that these functions (a) respect the constraints QC, and (b) fit the data S numerically . The resulting numerical functions constitute a numerical model of the domain. The intermediate qualitative constraints QC are also part of the overall model, the part that is not useful directly for making numerical predictions, but useful for the understanding and interpretation.

The two stages above can be carried out in various ways. In this paper we used program QUIN [Šuc, 2003] that induces qualitative constraints in the form of qualitative trees, and program QCGrid [Vladušič *et al.*, 2003] that performs piecewise linear regression respecting a given qualitative tree. In the following paragraphs we present in more detail the building blocks of the Q^2 learning approach, and a simple illustrative example (subsection 3.3).

3.1 Qualitative induction

QUIN (QUalitative INduction) is a learning program that looks for *qualitative patterns* in numerical data. QUIN expresses such qualitative patterns by *qualitative trees*. In this section we only give a brief introduction to QUIN; its detailed description and evaluation is given in [Šuc, 2003].

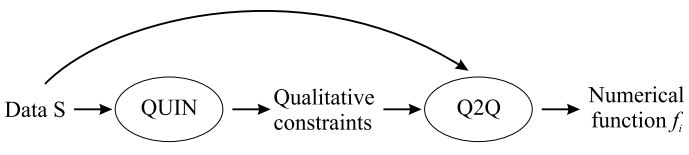

Figure 1: Q^2 Learning Schema

QUIN induces qualitative trees in a top-down greedy fashion, similarly to induction of decision trees [Breiman *et al.*, 1984; Quinlan, 1992]. The first difference between two approaches is that in induction of qualitative trees, a different error measure is used (based on minimum description length [Rissanen, 1978]). The second difference is in labels, assigned to leaves. In decision trees, the leaves are labelled with values of the dependent variable, whereas in qualitative trees the leaves are labelled with what we call *monotonic qualitative constraints*, which are a kind of monotonicity constraints that are widely used in the field of qualitative reasoning [Forbus, 1984; Kuipers, 1994].

A *monotonic qualitative constraint* $M^{s_1,\ldots,s_m}$ where $s_i \in \{+,-\}$, stands for an arbitrary relation between the class variable and m attributes, so that such a relation respects the qualitative constraints given by signs s_i. The class and any of the attributes can be either continuous or ordinal. For example, consider the constraint $Y = M^{s_1,\ldots,s_m}(X_1,...,X_m)$. A relation $(Y, X_1, ..., X_m)$ between class Y and m attributes $X_1, \ldots, X_m$ respects this constraint if for all $i = 1, \ldots, m$, class Y is s_i-related to attribute X_i. We say that Y is "+"-related (*positively related*) to an attribute X if for all pairs (y_1, x_1) and (y_2, x_2) of values of X and Y in the projection of the relation on (Y, X): $x_1 < x_2 \Rightarrow y_1 < y_2$. "Negatively related" is defined analogously.

Note that this definition does not require that the class variable is a function of the attributes mentioned in an MQC. As defined above, MQCs can have more than one argument. For example, $Z = M^{+,-}(X, Y)$ says that Z monotonically increases in X and monotonically decreases in Y. If Z=$M^{+,-}(X, Y)$ and both X and Y increase, then according to this constraint, Z may increase, decrease or stay unchanged. In such a case, a MQC cannot make an unambiguous prediction of the qualitative change in Z. This is called *qualitative ambiguity*. As qualitative ambiguity usually increases with the number of arguments in an MQC, QUIN prefers MQCs with less arguments.

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 the human intuition.

3.2 QCGrid

The QCGrid algorithm (QCGrid stands for *Qualitatively Constrained Grid*) is a regression algorithm that performs the Q2Q transformation shown in Fig. 1. Inputs to the QCGrid algorithm are (Fig. 2 - lines 8 and 9): learning data and qualitative constraints of one leaf of the qualitative tree and parameter k that defines the maximum density of the grid. Results of the QCGrid algorithm are piece-wise linear functions that respect given qualitative constraints.

When constructing the grid we search for a suitable grid spanned over the training data, using binary split search, commonly used in regression tree learning algorithms. Let S denote the learning data in the current leaf and consist of one attribute (x) and the function value (y). First, we perform linear regression and obtain the error, denoted as $MSE_{\text{undivided}}$, where MSE stands for *Mean Squared Error*. In the next step we try to find the value x_g of the independent variable that

1: **function** NTree = Q^2(S, k)
2: QTree = QUIN(S)
3: NTree = CopyStructure(QTree) {*Copy structure of the QTree to NTree*}
{*Induce quantitative models in leaves of the QTree*}
4: **for** Leaf ∈ QTree **do**
5: Data = ExtractData(S, QTree, Leaf) {*Use qualitative tree (QTree) to partition S over leaves*}
6: QConstr = GetQualitativeConstraints(QTree, Leaf) {*Store qualitative constraints of the current Leaf*}
7: Grid = GridSearch(Data, k) {*Find grid points with respect to parameter k using Data*}
8: QGrid = QRegress(Data, Grid, QConstr) {*Learn (qualitatively consistent) function values in the points of the Grid*}
9: NTree(Leaf) = QGrid {*Copy regression result into current Leaf of NTree*}
10: **end for**

Figure 2: Outline of the Q^2 learning approach. The two stages of the algorithm are divided as: (1) Induction of the qualitative model (line 2) and (2) induction of qualitatively consistent piece-wise linear functions (lines 4 — 11).

splits S into two subsets S_l and S_r "best". The subset S_l contains the learning examples that satisfy $x \leq x_g$ and the S_r contains examples where $x > x_g$ holds. Let X denote the values of the independent variable x in data set S. Candidates for the grid point x_g are all values in the X. Using each value as a candidate for x_g value, we divide S into S_l and S_r and perform linear regression in both subsets to obtain the error $MSE_{\text{divided}}(x_g)$. In order for some x_g to be found as the "best" grid point, all of the following criteria must hold:

$$\forall x \in X : MSE_{\text{divided}}(x_g) \leq MSE_{\text{divided}}(x)$$
$$MSE_{\text{divided}}(x_g) < MSE_{\text{undivided}}$$
$$|S_l| >= k \;\&\; |S_r| >= k$$

In the above equations, k denotes the minimal number of examples in each subset and is usually given as percentage of the examples in the leaf. When the best grid point in the S has been found, the algorithm recursively proceeds to both subsets S_l and S_r. The output is a set of grid points (G). If dataset had n attributes, the above procedure would be employed for each attribute thus obtaining $G_1, G_2, \ldots, G_n$ grids. The resulting grid G would be obtained with Carthesian product of the onedimensional grids.

In the next step of the QCGrid algorithm we learn qualitatively consistent function values in the previously found grid points G. To this end we use quadratic programming algorithm [Coleman and Li, 1996; Gill *et al.*, 1981], mathematically formulated as:

$$\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^\top \mathbf{H}\mathbf{x} + \mathbf{f}^\top \mathbf{x}$$
$$\mathbf{A}\mathbf{x} \leq \mathbf{b}$$

The (quadratic) criterion function is given by matrix $\mathbf{H}$ and vector $\mathbf{f}$. It must be minimized over all vectors $\mathbf{x}$. The remaining three equations give additional constraints: linear inequalities are defined with matrix $\mathbf{A}$ and vector $\mathbf{b}$; similarly $\mathbf{A}_{\text{eq}}$ and $\mathbf{b}_{\text{eq}}$ define linear equalities. Lower and upper bounds of $\mathbf{x}$ are defined by $\mathbf{lb}$ and $\mathbf{ub}$ respectively.

When approximating function values in points of the grid we maximize the fit to the data between the grid points using linear models. Thus, we transform the general MSE equation:

$$MSE = \frac{1}{|S|}\sum_{y_i \in S}(y_{\text{int},i} - y_i)^2 =$$
$$\frac{1}{|S_r|}\sum_{(x_i,y_i)\in L_s}(ax_i + b - y_i)^2$$

where y_i and $y_{\text{int},i}$ denote the original and predicted function values. As we use linear models of the original data, we can replace the term $y_{\text{int},i}$ with $ax_i + b$. Further transformation follows with incorporation of grid points $g_i \in G$ - division of the approximation function into piece-wise linear function:

$$MSE = \frac{1}{|S|}\left[\sum_{(g_1 \leq x_i \leq g_2, y_i)\in S}(a_1x_i + b_1 - y_i)^2 + \ldots + \right.$$
$$\left. \sum_{(g_{n-1} \leq x_i \leq g_b, y_i)\in S}(a_nx_i + b_n - y_i)^2\right]$$

We rewrite the above equation in order to express the slope coefficients explicitly: for i-th region $a_i = \frac{b_{i+1}-b_i}{g_{i+1}-g_i}$). The translation of the x_i point into $x_i - g_i$ gives the intercept coefficients (b_i) a new meaning - they become the function values at the grid points. The rewritten equation for our example is given below.

$$MSE = \frac{1}{|S|}\left[\sum_{(x_i,y_i)\in S}\left(\frac{x_i - g_1}{g_2 - g_1}(b_2 - b_1) + b_1 - y_i\right)^2 + \right.$$
$$\left. \ldots + \sum_{(x_i,y_i)\in S}\left(\frac{x_i - g_{n-1}}{g_n - g_{n-1}}(b_n - b_{n-1}) + b_{n-1} - y_i\right)^2\right],$$

where $g_{j-1} \leq x_i \leq g_j$ for $j = 2, \ldots, n$.

In the above formulation of the MSE equation the intercept coefficients b_i denote the function values at the grid points. In orded to minimize MSE on the dataset S, the matrix $\mathbf{H}$ and $\mathbf{f}$ (the criterion function) must be filled with coefficients at b_i. The matrix $\mathbf{H}$ is square and symmetric, and its fields contain the coefficients that are located next to the mixed terms after we have simplified the above expression. We mark the coefficient next to the mixed term $b_i \cdot b_j$ with k_{ij}. It then follows: $\mathbf{H}_{ij} = k_{ij}$. Similarly, $\mathbf{f}$ contains coefficients that are placed next to the term b_i.

Qualitative constraints are taken into account with matrix $\mathbf{A}$ and $\mathbf{b}$. The values are set according to the $y = M^+(x)$ which means that the function values (b_i) at split points $G_1 < G_2 < \ldots < G_{end}$ must satisfy the inequalities: $b_1 < b_2 < \ldots < b_{end}$)) (see [Šuc and Bratko, 2003] for detailed description).

The outline of the Q^2 algorithm in Fig. 2 shows construction of the numerical model. In every leaf of the qualitative tree, qualitative constraints are replaced with consistent piece-wise linear functions, based on the data contained in that particular leaf. When glueing these functions together, the problem of discontinuities in the class variable at the borders between leaves is not addressed. We have considered several possible approaches towards this issue, but found them to be overall unsatisfactory, as they cannot guarantee both continuity and qualitative faithfulness of the model.

3.3 Q^2: Simple example

Here we show an example of the Q^2 learning approach. We sampled a simple function and used the Q^2 learning approach in order to reconstruct the function qualitatively and quantitatively.

For the purpose of the example we sampled the function $y = x^2$. We randomly chose 20 values of the independent variable x from the interval $x \in [-2, 2]$. In each of the sampled points, we computed the value of the dependent variable y. This way, we obtained a dataset with 20 examples, each example consisting of the value of the independent variable x and the corresponding function value $y = x^2$.

The first step of the Q^2 approach is the construction of a qualitative model. A hand crafted qualitative model for our simple example is shown in Fig. 3*(a)*. The induced qualitative model is shown in Fig. 3*(b)*. We can see that the only difference between the models is the value of the internal (splitting) node - when constructing the hand crafted model we knew the correct splitting value, whereas QUIN was given only the sampled learning data and no background knowledge regarding the function to be modeled. Both qualitative trees are fully consistent with the learning data.

To perform qualitative-to-quantitative transformation we used the QCGrid algorithm. Fig. 4 shows the result of the Q2Q transformation, when using QUIN-induced qualitative tree. QCGrid induced piecewise linear functions in both leaves of the qualitative tree - both functions consist of three segments, as the value of parameter k was set to 0.2. Hence the minimum number of examples in segments of both leaves was 2. It can also be observed that extreme data values are end points of the underlying grid.

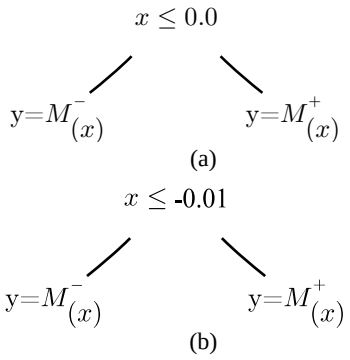

Figure 3: *(a)* Hand-crafted qualitative tree for the $y = x^2$ domain. *(b)* Qualitative tree induced by QUIN.

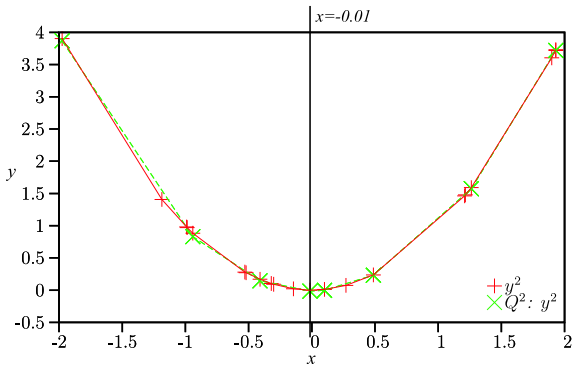

Figure 4: QCGrid model for $y = x^2$. The root split from QUIN's qualitative tree is indicated by a vertical line at $x = -0.01$.

We extend this simple example with comparison of the numerical accuracy between the described Q^2 approach and *locally weighted regression - LWR* [Atkeson *et al.*, 1997], implemented in WEKA [Witten and Frank, 2000]. We obtained learning data with additional resampling of the $y = x^2$ function thus obtaining 10 datasets, each consisting of 20 examples, with no added noise. The comparison was performed as follows: Both methods used internal 4-fold cross-validation to determine the best prediction parameters. LWR used Gaussian weighting function, with possible values for number of nearest neighbours taken from: $[1, 2, \ldots, 20]$. Possible values for parameter k of the QCGrid algorithm were taken from $[0.1, 0.2, \ldots, 0.5]$. Each of the 10 datasets was once used as the learning set for both methods. The induced models were then tested on the remaining 9 datasets, thus we performed 90 experiments. The error was measured with *root mean squared error* (RMSE).

To evaluate the obtained results, we first performed a paired comparison of all 90 obtained numerical errors. The Q^2 approach had lower RMSE in 74% of the experiments. Mean RMSE over all 90 experiments of the Q^2 approach was 0.12, with standard deviation 0.1, while LWR achieved only 0.32, with standard deviation 0.29. We then grouped numerical results, obtained with the same learning set and averaged

them. So, we had 10 average prediction errors, each resulting from one of the learning sets. Such comparison between learning algorithms shows that Q^2 is on average better in 9 out of 10 cases. To determine whether these differences are significant, we performed t-test, which showed the significance at level 0.04.

The obtained results were further analysed in order to determine the reason for poor performance of the LWR approach. Fig. 5 shows prediction of both methods on one of the test sets, if learning set shown in Fig. 4 is used. We can see that LWR makes rather large quantitative and qualitative errors - a consequence of poor learning examples coverage in some regions of the learning dataset. Using qualitative models as guidance in such problematic areas alleviates prediction, hence better performance of the Q^2 approach.

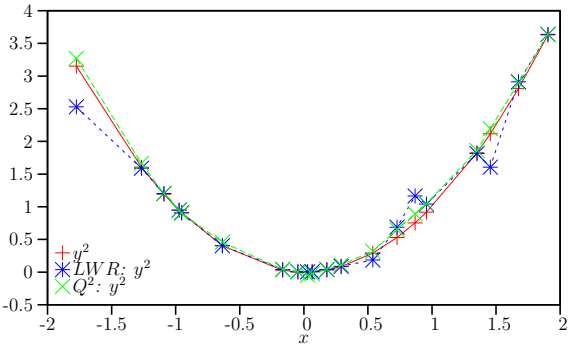

Figure 5: Prediction of the two competing approaches: LWR makes rather large quantitative and more qualitative errors, while Q^2 predictions are qualitatively consistent.

4 The learning data

Available data are meteorological (temperature, relative humidity, wind speed and direction, solar radiation, precipitation) and air quality measurements (O_3, NO, NO_2, CO) as well as predictions of the ALADIN model. Because of changes in the structure of the ALADIN model, we are limited to use its predictions only from 2001 to 2003. The measurement data was taken from the same period with a lot of missing values encountered. The measuring tolerances are also high and prevent potencial improvement of numerical accuracy which is in the same range as the tolerances.

By an expert opinion, the amount of data is too small in many aspects, namely the time period, the number of measured variables and the number of measurement stations. Various important measurements, such as VOC, are currently not available and the process of acquiring them is underway. Even so, it was possible to induce meaningful qualitative models, and a numerical model whose prediction accuracy suffices for operational use. To enable the evaluation of the accuracy of the induced model, the data was split at the very beginning into the learning set and the test set. The learning set was taken to include the data from 2001-2002, while data from 2003 was left for testing.

ALADIN output data are 3D fields of meteorological parameters with a finite resolution, in our case 11 km. Up 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 sub grid cell variation. When interpolating meteorological parameters in a selected point (for instance a meteorological station) from model output fields, it is erroneously assumed that model output represents values in the centers of model grid cells. Therefore we approximated the values at the required points through numerical regression. We decided to use stepwise linear regression method to build a regression model for each of the meteorological parameters separately. With the stepwise method, a regression model is built progressively. At each step, the independent variable which has the smallest P-value (using F-test), is added to the model, but only if that probability is smaller than 0.05. Variables already in the regression equation are removed if their P-value becomes larger than 0.1. The method terminates when no more variables are eligible for inclusion or removal. In our case, independent variables were meteorological measurements of temperature, solar radiation, relative humidity and pressure at the meteorological stations in Ljubljana and Nova Gorica. The dependent variables are prognostic model values at 210 grid points over Slovenia with resolution of 11 km. Time resolution of data is 3 hours. Final results were eight linear regression models (for four meteorological variables at two stations).

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

The data from the ALADIN model were used together with meteorological and air quality measurements to induce a prediction model for Ljubljana. 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 principals used in CAMx model were compared to the induced qualitative 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 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 the attributes. The learning set was therefore sampled taking every 4^{th} example and subsets of 4 attributes were passed to QUIN [Šuc, 2003]. The output models were evaluated by coverage and qualitative uncertainty which QUIN calculates. The following set of attributes came out to be the best on the given learning set: relative humidity (H), solar radiation (S), index of half-hour

interval (t), nitrogen dioxide concentration (NO_2). The resulting qualitative tree is shown in figure 6.

Experts' interpretation

The first look at the selected attributes shows that no irrelevant attributes were chosen. Surprisingly, no 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, from the analysis of data scatter plots, 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.

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 (MQC) $M^{+}_{(S)}$ 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 $t \leq 35$, i.e. 17:30. 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 17:30. The production of O_3 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 cities 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 nonlinear 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 MQCs are also easily explained since we always have one or more dependence from $O_3 = M^{+,+,+}_{(t,S,NO_2)}$. 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.

5.2 Numerical predictions

The attributes used in the learning process were built from the ALADIN predictions at the model grid points, neighboring the meteorological station point in both cities, as described in section 4. At that point, the meteorological measurements were performed. The attributes used are: $MAXNO$ (max. concentration of NO in the last 36 hours before the prediction is made), $Tavg915LJ$ (avg. of the ALADINs predictions of temperature from 09:00 to 15:00) in Ljubljana (LJ) and $Ssum015LJ$ (the sum of ALADINs predictions of solar radiation from 00:00 to 15:00). The qualitative tree for Ljubljana is shown in figure 7. It shows that the ozone concentration is positively correlated with the temperature and solar radiation while negatively correlated with the concentration of NO. The concentration of NO in the leaves of the qualitative trees reflects the dominating mechanisms of the ozone cycle. Higher NO concentrations occur during night time with low ozone concentration (right branch). On the contrary, high ozone concentration as a result of photochemical formation prevents high NO concentration (left branch).

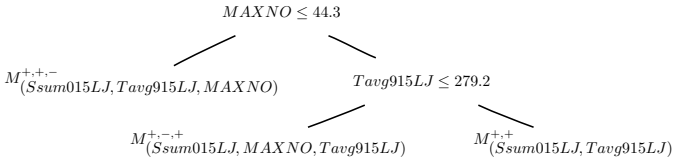

Figure 7: Qualitative model for Ljubljana

Temperature and solar radiation are statistically highly correlated, so a model can choose each of the 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 better statistical correlation with ozone concentration, which resulted in the second leaf. This proves that highest ozone concentrations occur during daytime in summer in hot, sunny and dry weather.

We used QCGrid to build a numerical model from the qualitative one. Numerical accuracy of induced model is compared to linear regression (LR) and model trees (M5) [Quinlan, 1992]. Table 1 shows the RMSE measured on the test set. Q2 turns out to be superior to LR and M5, although not significantly.

Table 1: Comparison of the numerical accuracy of the competing methods

RMSE on test set	LR	M5	Q2
Ljubljana	21.63	22.94	19.9

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.

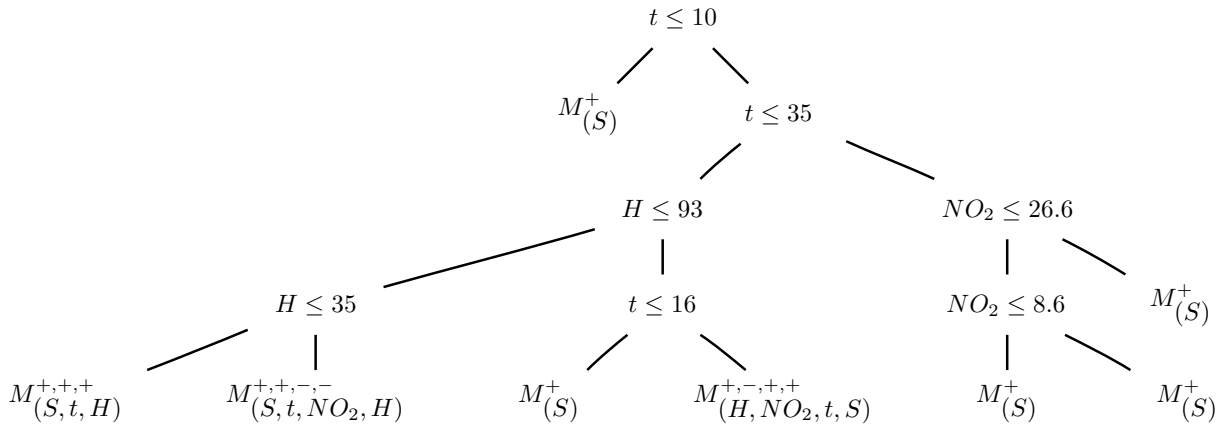

Figure 6: The qualitative model for Nova Gorica built from measurement data. Attributes: relative humidity (H), solar radiation (S), index of half-hour interval (t), nitrogen dioxide concentration (NO_2)

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 results that are slightly better, but the improvements are not significant. 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; I. N. Athanasiadis and Petridis, 2003; S. Canu, 2001; M. C. Hubbard, 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 would like to thank dr. Matevž Pompe from the Faculty of Chemistry and Chemical Engineering, University of Ljubljana, for evaluation of our qualitative model from the chemistry point of view. National ICT Australia is founded by the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council.

References

[A. Planinšek, 2000] et al. A. Planinšek. Air pollution in slovenia in 2001. Technical report, Environmental Agency of the Republic of Slovenia, 2000.

[Aladin, 1997] Aladin. Aladin international team, the aladin project: Mesoscale modelling seen as a basic tool for weather forecasting and atmospheric research. *WMO Bulletin*, 46:317–324, 1997.

[Atkeson *et al.*, 1997] C. Atkeson, A. Moore, and S. Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11:11–73, 1997.

[Breiman *et al.*, 1984] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. 1984.

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

[Coleman and Li, 1996] T.F. Coleman and Y. Li. A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables. *SIAM Journal on Optimization*, 6(3):1040–1058, 1996.

[Environ, 2004] Environ. Comprehensive air quality model with extensions, version 4, users guide. Technical report, Environ International Corporation, www.camx.com, Novato, California, 2004.

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

[Forbus, 1984] K. Forbus. Qualitative process theory. *Artificial Intelligence*, 24:85–168, 1984.

[Gill *et al.*, 1981] P.E. Gill, W. Murray, and M.H. Wright. *Practical Optimization*. Academic Press, London, 1981.

[I. N. Athanasiadis and Petridis, 2003] P. A. Mitkas I. N. Athanasiadis, V. G. Kaburlasos and V. Petridis. 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*, 2003.

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

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

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

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

[Rissanen, 1978] J. Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.

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

[Vladušič *et al.*, 2003] D. Vladušič, D. Šuc, and I. Bratko. Q2q software for inducing quantitative models from designers vague specifications. clockwork project report rest-57. Technical report, University of Ljubljana, 2003.

[Šuc and Bratko, 2003] D. Šuc and I. Bratko. Improving numerical accuracy with qualitative constraints. In N. Lavrač, D. Gamberger, H. Blockeel, and L. Todorovski, editors, *Proceedings of the 14th European Conference on Machine Learning*, pages 385–396. Springer, 2003. Dubrovnik, Croatia.

[Šuc *et al.*, 2004] D. Šuc, D. Vladušič, and I. Bratko. Qualitatively faithful quantitative prediction. 2004.

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

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