Application of Decision Trees in Problem of Air Quality Modelling in the Czech Republic Locality

KAŠPAROVÁ MILOSLAVA, KŘUPKA JIŘÍ, JIRAVA PAVEL
Institute of System Engineering and Informatics
Faculty of Economics and Administration, University of Pardubice
Studentská 84, 532 10 Pardubice
CZECH REPUBLIC
Miloslava.Kasparova@upce.cz, Jiri.Krupka@upce.cz, Pavel.Jirava@upce.cz

Abstract: Air pollution belongs to environmental threats. This paper deals with an application of selected algorithms used in decision trees in order to an air quality modelling. We focused on daily observations of air polluting substances concentrations in one of the cities in the Pardubice region in the Czech Republic. After data collection, data description, and data preprocessing, we worked on the creation of classification models and the analysis of the achieved results. As modeling algorithms we selected C5.0 algorithm, boosting, CHAID method and CR&T. Differences of results were minimal.

Key-words: classification model, air pollution, air quality, algorithm, daily observations, Czech Republic

1 Introduction

Pollution belongs to environmental aspects of sustainable development. The earth can be considered to consist of a number of ecosystems, which have a certain capacity for regeneration. For instance the atmosphere's composition is able to regenerate itself, but the rate of regeneration depends on the current state of the atmosphere. However problems can occur when the stresses on ecosystems increase, for instance through high levels of pollution. Many environmental problems involve the degradation of renewable natural sources or their use at rates greater that they can be renewed. Many of the impacts of pollution are long term and wide ranging [7].

Air pollution belongs to environmental threats. There are a number of different types of air pollutants, including suspended particulate matter, lead, sulphur dioxide, carbon monoxide and nitrogen oxides. Sources of suspended particulate matter include incomplete fuel combustion and vehicle exhaust gases, particulary form diesel engines. Health effects include increased incidence of respiratory diseases, such as asthma, bronchitis and emphysema, and increased mortality of people [7].

On the basis of knowledge of system definitions, and an experience how to this knowledge can be applied in defining a steerable object such as a model of sustainability development, and [3, 19] we can depict possibilities for steering in a sustainable development model (Fig.1).

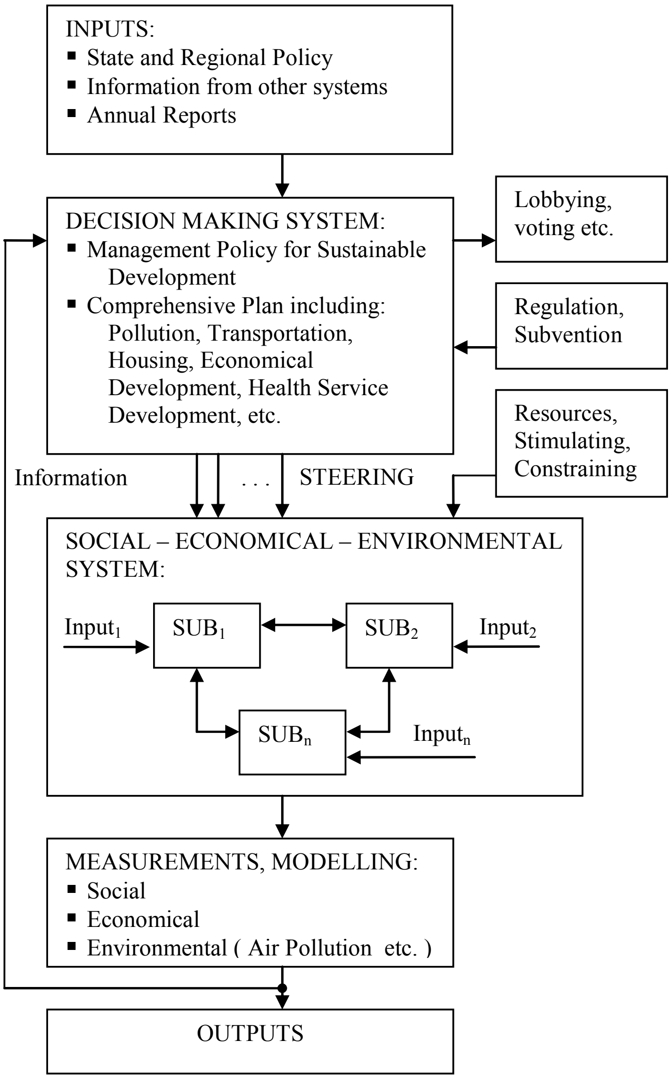

INPUTS:
State and Regional Policy
Information from other systems
Annual Reports
DECISION MAKING SYSTEM:
Management Policy for Sustainable Development
Comprehensive Plan including: Pollution, Transportation, Housing, Economical Development, Health Service Development, etc.
Lobbying, voting etc.
Regulation, Subvention
Resources, Stimulating, Constraining
Information
. . .
STEERING
SOCIAL – ECONOMICAL – ENVIRONMENTAL SYSTEM:
Input$_1$
SUB$_1$
SUB$_2$
Input$_2$
SUB$_n$
Input$_n$
MEASUREMENTS, MODELLING:
Social
Economical
Environmental (Air Pollution etc.)
OUTPUTS

Fig.1 Possibilities for steering - sustainable development

As shown in Fig.1, the model is assumed to consists of a set of related subsystems SUB_1, SUB_2, … SUB_n, and process which use inputs and deliver output to other subsystems, and to the outside. We can take advance of the output of an air pollution model as feedback (information) about quality of environment in the decision making system.

An environment is our surroundings. It includes living and non-living things around us. It is a system compounded of natural, artificial, and social components that are in interaction with one another. It is all what forms natural conditions to an existence of organisms, including human, and the preconditions of their evolution.

Firstly, air, water, rocks, land, organisms, ecosystems, and energy are components of this. The weakening of components results in an imbalance and degradation of the environment.

The State environmental policy of the Czech Republic (SEP CR) [17] belongs to documents that deal with protection and quality assurance of the environment in the Czech Republic.

It is a fundamental reference document for other sectors and regional policies, from the standpoint of the environment. Although SEP CR is a governmental document its implementation requires an active participation of the general public, partners in the business sector, science and research, and others.

The SEP CR is a policy that should be followed by Czech Corporations, as well as other organizations, as an instrument that will assist them in their strategic and every-day operative decision-making, so as to lead not only to the creation of new economic, social, and cultural values, but also to an improvement in the quality of life and quality of the environment.

The state of the environment is regularly monitored and evaluated (annual reports of Ministry of the Environment submitted by the Government to the Chamber of Deputies of the Parliament of CR and the public) and consequently SEP CR reacts to all the important changes (negative trends) in the state of the environment.

In accordance with the state of the environment transposition and implementation of European law, and the basic principles of the protection of the environment and its sustainable development, the updated SEP CR concentrates on the following four priority areas [17]:

1. Nature conservation, protection of the landscape, and biological diversity.
2. Sustainable use of natural resources, material flows, and waste management.
3. Environment and the quality of life.
4. Protection of the climate system of the Earth and prevention of long-range transport of air pollution.

This classification emphasizes not only protection of the basic components of the environment (air, water, lithosphere), but primarily integrated protection of ecosystems and the landscape (conservation of biodiversity), sustainable development, and an improvement in the quality of life. The fourth area reflects the responsibility of CR for the European and global environment (climate system, ozone layer) and the international cooperation entailed therein.

On the basis of these areas many partial goals are defined. One of the goals is to uplift air quality through defined steps and provisions.

In relation to protection of human health, it is necessary to monitor the quality of drinking water and to reduce the burden on the human population resulting from the pollution of the air and foodstuffs.

The Czech Hydrometeorological Institute (CHI) achieves, with the aid of various laws, the establishment and operation of a national network of monitoring stations that measure the amount of air pollution in the Czech Republic. Some of the stations in this network are designed for automated air polluting monitoring (A2PM).

Measuring stations work in continuous operation and give measured values in real time to CHI centers. In the Czech Republic, 97 measuring station's A2PM work is run by CHI. Except for the results from other measuring stations outside of these 97 stations, the results are submitted in the information system. Most of the stations have analyzers to measure sulfur dioxide concentrations [SO_2], nitrogen monoxide [NO], nitrogen dioxide [NO_2], and suspended particles [PM_{10}]. Concentrations of ozone [O_3] and carbon monoxide [CO] are only measured in few measuring stations. A selected amount of A2PM stations also measure concentrations of some volatile organic matter (benzene, toluene, xylene).

Pardubice, the seat of the Pardubice region, is situated at the confluence of the Labe and Chrudimka rivers and is one of the most beautiful towns in East Bohemia. The area of this city is practically 78 km^2 and approximately 90 thousand inhabitants live there. It lies in an altitude of 215 to 237 meters above sea-level. With regard to an industrial enterprises existence, heavy traffic, and other factors, Pardubice belongs to air pollution areas.

Data used in this paper is from daily observations of air polluting substances concentrations in part of Pardubice-Dukla (Dukla) in 2007. An automated monitoring system is located in a park (in the campus of a primary school). The target of the measurement program is to evaluate the total level of concentrations and an evaluation of the effect on the population's health. Basic information about this measure is in the Table 1.

Table 1 Basic information about locality of measure

Basic Information	Value
Locality code	EPAU
Name	Pardubice - Dukla
State	Czech Republic
Owner	CHI
Basic admin. unit	Pardubice
Coordinates	50° 1' 26,54 " North latitude; 15° 45' 48,78 " East longitude
Altitude	239 m
EOI - zone type	Urban
EOI - zone characteristic	Residential
Terrain	Plane, not much (sparsely) undulating terrain
Landscape	Multi-storey building (housing estates of the recent decades)
Measuring programme	Automated measuring programme

The air quality evaluation is based on the result of the weight concentrations measures of substance in the air. The evaluation of air quality by [1] is in the Table 2.

Table 2 The air quality evaluation

Air Quality	Index	SO_2	NO_2	CO	O_3	PM_{10}
		1h	1h	8h	1h	1h
		[in µg/m^3]				
Very good	1	0-25	0-25	0-1000	0-33	0-15
Good	2	25-50	25-50	1000-2000	33-65	15-30
Favorable	3	50-120	50-100	2000-4000	65-120	30-50
Satisfactory	4	120-250	100-200	4000-10000	120-180	50-70
Bad	5	250-500	200-400	10000-30000	180-240	70-150
Very bad	6	500-	400-	30000-	240-	150-

This evaluation takes the possible influence of human health into account [17]. New limits of monitoring and air quality evaluation are specified in the regulation of the Czech Republic government No: 597/2006 Coll. These limits are set separately for health protection and vegetation and ecosystems protection.

2 Problem Formulation

The goal of this paper is to create a model of air quality in a given locality through the use of selected methods. It means to design and verify a classification model through the usage of decision trees. The following are the steps of realization:

- data description and data preprocessing
- classification model creation by decision trees
- testing of classifiers and comparison of results.

2.1 Data Description and Data Preprocessing

Original data was obtained from the daily observation of air polluting substances concentrations in 2007 in Dukla. In this first step we realized data cleaning, standardization, and correlation.

Data cleaning techniques [11] fill in missing values, smooth noisy data, identify outliers, and correct inconsistencies in the data. Methods used for dealing with missing values include: ignoring the objects, filling in the missing value manually, using the attribute mean to fill in the missing value, etc. [6, 11, 22]. In our case we ignored objects with missing values. The attribute means that using the most probable value or most frequent value is a convenient method in this data. Original data matrix included 365 observations. After an elimination of missing values, 330 daily observations (objects, data) described by 11 attributes (variables) were achieved. It means, we achieved data matrix O in dimension 330 × 11. Every observation o_i for i = 1, 2, …, 330 can be described by the following vector $\mathbf{o_i}$ = (x_{i1}, x_{i2}, ..., x_{i11}). Basic descriptive characteristics of attributes are in the Table 3. Although the air pollution rate is the result of many factors, the classification model is created on the basis of this available data.

In the determination of air quality in Dukla locality, (output attribute y_k) on the basis of the achieved data, the techniques for the air quality evaluation in the Table 2 were used. It means we work with the index (class) of air quality evaluation y_k for k = 1, 2, 3, 4, 5, 6 and the final vector is the following: $\mathbf{o_i}$ = (x_{i1}, x_{i2}, ..., x_{i11}, y_k).

The mean monthly values (Table 3) of variables x_1, x_2, x_3, x_4 and x_5 from the Table 2 measured in Dukla locality in 2007 are in the Fig.2 and Fig.3. Other variables x_6, x_7 and x_8 are in the Fig.4.

Representation of data by the index (class) of air quality evaluation y_k is in the Fig.5. We can see that this locality belongs to areas with good (48.79 %) and favorable (44.55 %) air quality.

In the measure of the relationship between variables we used the correlation [6, 22]. The most widely-used type of a correlation coefficient is Pearson correlation coefficient ρ_{ij}. In the data matrix for classification model with 330 observations described by 11 inputs variable and 1 output variable, the top correlation was found between variables x_5 and x_7 (ρ_{ij} = 0.973) and x_6 and x_8 (ρ_{ij} = 0.975). On the basis of variables in Table 2 that are used for air quality evaluation, we eliminated attributes x_7 and x_8.

Table 3 Basic descriptive characteristic of input attributes

At.	Name of Attribute	Min	Max	Mean	Std. Dev
Air polluting substances					
x_1	Sulfur dioxide (SO_2)	1.40	113.30	7.95	7.88
x_2	Nitrogen dioxide (NO_2)	6.00	50.20	19.90	8.03
x_3	Carbon monoxide (CO)	128.60	1490.40	532.09	321.50
x_4	Ozone (O_3)	9.00	105.20	51.23	22.71
x_5	Suspended particles (PM_{10})	6.00	91.40	26.37	15.18
x_6	Nitrogen monoxine (NO)	0.70	66.40	7.53	9.68
x_7	Suspended particles ($PM_{2,5}$)	3.50	61.40	18.16	10.88
x_8	Nitrogen oxides (NO_x)	7.70	168.80	32.36	24.27
Meteorological attributes					
x_9	Solar radiation	7.10	423.80	156.42	113.17
x_{10}	Temperature two meters above the surface of the Earth	266.10	297.70	283.16	7.48
x_{11}	Relative air humidity	63.40	82.40	76.80	3.49

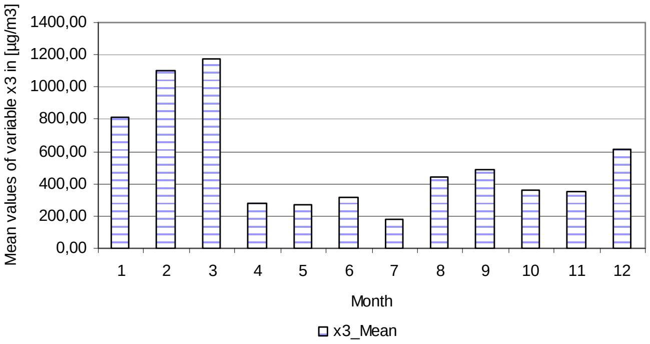


Fig.3 Mean monthly values of x_3 variable

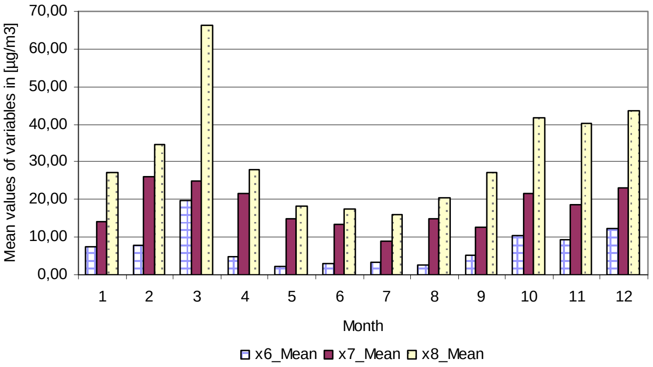


Fig.4 Mean monthly values of x_6, x_7 and x_8

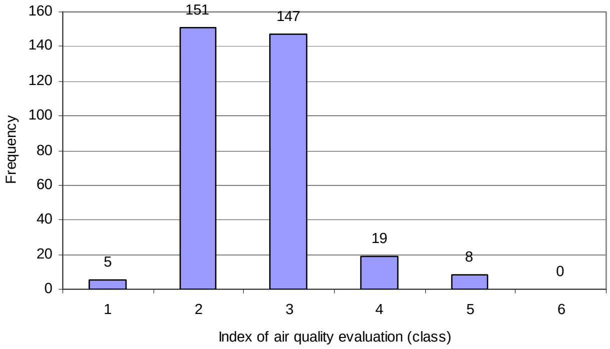


Fig.5 Representation of data by the index (class) of air quality evaluation

A model of system [2] is an idealized representation – an abstract and simplified description – of a real world situation that is to be studied and/or analyzed.

Model formulation [2, 9] is the task of converting a precise problem description into a mathematical model. It is a complex task requiring diverse types of knowledge. The appropriateness of a model depends on a variety of factors such as accuracy, tractability, availability of relevant data, and understandability.

In [2] the basic model of air pollution is described. It should be noted that this model presumes no chemical reaction of pollutants but only gradual dissipation of the pollutant. This model is known as Acid Rain Management Model and it is solved as an optimization problem.

In the next part we design the model of classification of air quality evaluation to five classes on the basis decision trees.

3 Classification Model Creation

The content of this paper is to describe the designed classification model (classifier) and the achieved results of classification.

For the modeling of air quality we used a data set that contains 1 dependent variable y_k and 9 independent variables x_1, x_2, x_3, x_4, x_5,x_6, x_9, x_{10} and x_{11}.

We randomly partitioned the dataset into two parts. In regards to the classification model creation, two thirds of the original dataset was allocated to the training set and the remaining objects were allocated to the testing set. Using the same objects to train and estimate their accuracy may result in misleading estimates due to overfitting. In this case if we used training set for testing we can only determine the resubstitution error R_c [6, 22]. It is the error rate in the training data set. It is calculated by resubstituting the training instances into a classifier that was constructed from them. Although it is not a reliable predictor of the true error rate on new data, it is nevertheless often useful to know.

We dealt with decision trees. The C5.0 algorithm and boosting were used in this example and focused to CHAID algorithm as well. For modeling purposes we used software Clementine Desktop 10.01. The classification model design is in the Fig.6.

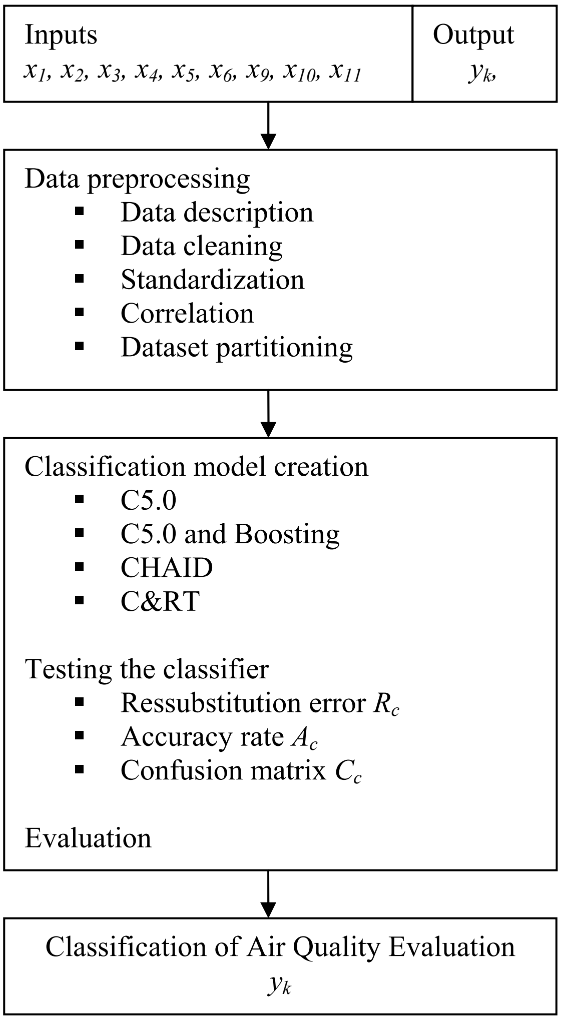


Fig.6 The classification model design

3.1 Decision Trees

A decision tree [6, 14, 15, 16, 18] is a predictive model which can be used to represent both classifiers and regression models. In operations research, on the other hand, decision trees refer to a hierarchical model of decisions and their consequences. When decision tree is used for classification tasks, it is more appropriately referred to as a classification tree. Classification trees are used to classify an object or an instance to a predefined set of classes based on their attributes values. These trees are frequently used in applied fields such as finance, marketing, engineering and medicine. They are useful as an exploratory technique [14].

3.1.1 Classification trees

A decision tree is a classifier expressed as a recursive partition of the instance space. This tree consists of root, nodes that form a rooted tree, and branches and leafs (terminals). A node with outgoing edges is referred to as an internal or test node. All other nodes are called terminals. In the decision tree, each internal node splits the instance space into two or more subspaces according to a certain discrete function of the input attribute values.

Each leaf is assigned to one class representing the most appropriate target value. Instances are classified from the root of the tree down to a leaf, according to the outcome of the test along the path. In term of the size of the decision tree, decision makers prefer a decision tree that is not complex. Usually the tree complexity is measured by one of the following metrics: the total number of nodes, total number of terminals, tree depth and number of attributes. The tree complexity is explicitly controlled by the stopping criteria and the pruning method that are used. Advantages and disadvantages of these trees are for example in [14].

Algorithms as ID3, C4.5, QUEST, etc. are used for the building of decision trees (more examples in [4, 6, 13, 15, 16, 18]).

3.1.2 The C5.0 Method and Boosting

A C5.0 method [16] works by splitting the sample based on the attribute that provides the maximum information gain [5, 12, 22]. Each subsample defined by the first split is then split again, usually based on a different field, and the process repeats until the subsamples cannot be split any further. Finally, the lowest-level splits are reexamined, and those that do not contribute significantly to the value of the model are removed or pruned.

Boosting (also Adaptive Resampling and Combinating) is a general method for improving the performance of any learning algorithm [11]. It works by building multiple models in a sequence. The first model is built in the usual way. Then, a second model is built in such a way that it focuses especially on the records that were misclassified by the first model. Then a third model is built to focus on the second model's errors, and so on. Finally, cases are classified by applying the whole set of models to them, using a weighted voting procedure to combine the separate predictions into one overall prediction. Boosting can significantly improve the accuracy of a C5.0 model, but it also requires longer training (more examples in [6, 11, 22]).

3.1.3 The CHAID Method

CHAID (Chi-squared Automatic Interaction Detection) is a classification method for building decision trees by using chi-square statistics to identify optimal splits. It was originally designed to handle nominal attributes only [11, 14].

For each input attribute a_i CHAID finds the pair of values in V_i that is least significantly different with respect to the target attribute. The significant difference is measured by the p value obtained from a statistical test. The statistical test used depends on the type of target attribute. If the target attribute is continuous a F test is used. If it is nominal, it is a Pearson chi-squared test and for ordinal target attribute it is a likelihood ratio test. For each selected pair of values, the CHAID checks if the obtained p value is greater then a certain merge threshold. If the answer is positive, it merges the values and searches for an additional potential pair to be merged the process is repeated until no significant pairs are found. The best input attribute to be used for splitting the current node is then selected, such that each child node is made of a group of homogenous values of the selected attribute. No split is performed if the adjusted *p* value of the best input attribute is not less than a certain split threshold. This procedure is also stopped when one of the following conditions is fulfilled [rokach]: maximum tree depth is reached; minimum number of cases in a node for being a parent is reached, so it can not be split any further; minimum number of cases in a node for being a child node is reached.

The CHAID does not perform pruning [14].

3.1.4 Classification and Regression Tree

The Classification and Regression Tree (C&RT) is a tree-based classification and prediction method. Similar to C5.0, this method uses recursive partitioning to split the training records into segments with similar output field values. The C&RT starts by examining the input fields to find the best split, measured by the reduction in an impurity index that results from the split. The split defines two subgroups, each of which is subsequently split into two more subgroups, and so on, until one of the stopping criteria is triggered. All splits are binary - only two subgroups [16].

Classification and Regression Trees give the option to first grow the tree, and then prune based on a cost-complexity algorithm that adjusts the risk estimate based on the number of terminal nodes. This method, which allows the tree to grow large before pruning using more complex criterion, may result in smaller trees with better cross-validation properties. Increasing the number of terminal nodes generally reduces the risk for the current - training data, but the actual risk may be higher when the model is generalized to unseen data. In an extreme case it can be a separate terminal node for each record in the training set. The risk estimate would be 0 % since every record falls into its own node, but the risk of misclassification for unseen - testing data would almost certainly be greater than 0. The cost-complexity measure attempts to compensate for this [16].

Impurity measures: C&RT works by choosing a split at each node such that each child node created by the split is more pure than its parent node. Here purity refers to similarity of values of the target field. In a completely pure node, all of the records have the same value for the target field. C&RT measures the impurity of a split at a node by defining an impurity measure.

There are three different impurity measures used to find splits for C&RT models, depending on the type of the target field. For symbolic target fields, you can choose Gini or towing, for continuous targets it is the least-squared deviation (LSD) method. Because we work with symbolic target (output y_k) we choose Gini index [6, 11, 14].

4 Classifiers and Results of Classification

The resulting classifiers were tested on the train and test sets, and many tests were realized. We used the resubstitution error R_c, the accuracy rate A_c, and confusion matrix C_c [6, 11, 22], a convenient tool for analyzing the performance of a classifier. It is a square matrix that specifies the accuracy of the classifier to the classification problem. A good classifier should have a diagonal confusion matrix (all off-diagonal values are zero) [11].

The accuracy of a classifier A_c on a given test set is the percentage of test set objects that are correctly classified by the classifier. It refers to the ability of a classifier to correctly predict the class label of new or previously unseen data. The associated class label of each test object is compared with the learned classifier's class prediction for that object [6].

Achieved results by these methods are in Table 4 and in the Fig.8., Fig.9, Fig.10 and Fig.11. The mean results of tests can be seen in the Table 4.

Firstly we applied C 5.0 algorithm, C5.0 with boosting, and the CHAID. We can say that the best method of classification under the term of this problem is C5.0 with $A_{c(c5.0)}$ = 94.2 % of correct classification, with boosting it is $A_{c(c5.0\ boost)}$ = 94,43 % ($A_{c(C5.0)} < A_{c(C5.0\ boost)}$). The best result was achieved by C5.0 with boosting (99.06 % of correct classification) and the worst was CHAID method with result 88.29 % of correct classification.

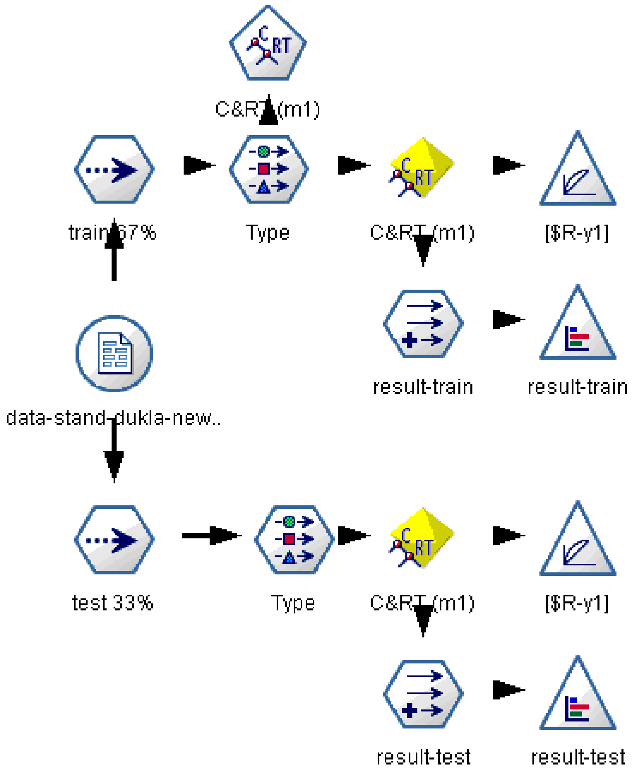


Fig.7 Model C&RT(m1)

After the testing these models the new model based on C&RT was created. Firstly maximum tree depth was set to 10 levels below root (a model C&RT(m1)) and secondly to 5 levels below root (a model C&RT(m2)).

By created model (m1) it was achieved the best results (the mean accuracy rate $A_{c(C\&RT(m1))}$ = 94.81 % and the mean resubstitution error $R_{c(C\&RT(m1))}$ = 94.56 %). It means, the mean accuracy rate $A_{c(C\&RT(m1))}$ is better then the mean accuracy rate $A_{c(C5.0\ boost)}$ by the application of the algorithm C 5.0 with boosting. A difference is 0.38 %.

By the model C&RT(m2) the mean accuracy rate $A_{c(C\&RT(m2))}$ is 93.59 % and the mean resubstitution error $R_{c(C\&RT(m2))}$ is 93.58 %. It means, the mean accuracy rate $A_{c(C\&RT(m2))}$ is better then the mean accuracy rate $A_{c(CHAID)}$ by the application of the algorithm CHAID (a difference is 2.18 %) and less than the mean accuracy rate $A_{c(C5.0)}$.

The differences all used algorithms with regard to the best mean result are in the Fig.8. The best and the worst results by use of all algorithms are in the Table 5 and Table 6. The model based on the C5.0 algorithm with boosting created in Clementine 10.1 is in the Fig.7

Generally, on the basis of realized tests it means:

$$A_{c(CHAID)} < A_{c(C\&RT(m2))} < A_{c(C5.0)} < A_{c(C5.0\ boost)} < A_{c(C\&RT(m2))}. \quad (1)$$

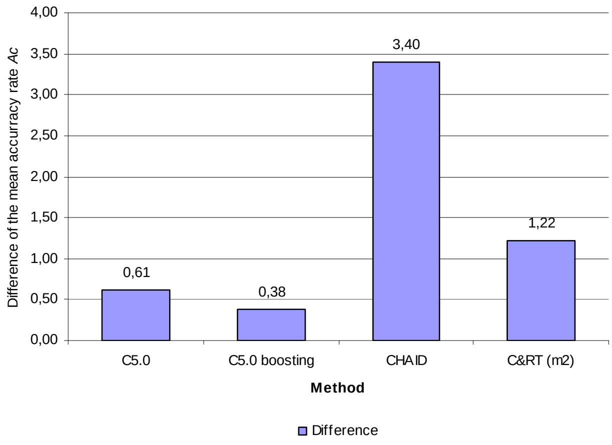


Fig.8. The differences among results of models

Table 4 Mean values of tests in test data

Method	Accuracy rate A_c [in %]	Resubstitution error R_c [in %]
C5.0	94.20	94.48
C5.0 boosting	94.43	93.36
CHAID	91.41	91.67
C&RT (m1)	**94.81**	94.56
C&RT (m2)	93.59	93.58

Table 5 The best values of tests in test data

Method	Accuracy rate A_c [in %]
C5.0	96.23
C5.0 boosting	99.06
CHAID	94.44
C&RT (m1)	**99.07**
C&RT (m2)	96.67

Table 6 The worst values of tests in test data

Method	Accuracy rate A_c [in %]
C5.0	91.75
C5.0 boosting	90.4
CHAID	88.29
C&RT (m1)	91.2
C&RT (m2)	89.22

An example of the confusion matrix $C_{c(C5.0)}$ for classifier based on C5.0 algorithm is in the Table 7. The accuracy rate is 95.12 %. The rows represent actual observed values, and the columns represent predicted values. The cell in the table indicates the number of records for each combination of final and actual values.

Table 7 Example of confusion matrix

			Final values of classifier $C_{c(C5.0)}$				
			Class				
			1	**2**	**3**	**4**	**5**
Actual observed values	**Class**	**1**	1	1	0	0	0
		2	0	55	4	0	0
		3	0	1	51	0	0
		4	0	0	0	8	0
		5	0	0	0	0	2

The result comparison of methods is in the Fig.9. We can see, every method achieves approximately similar results of classification.

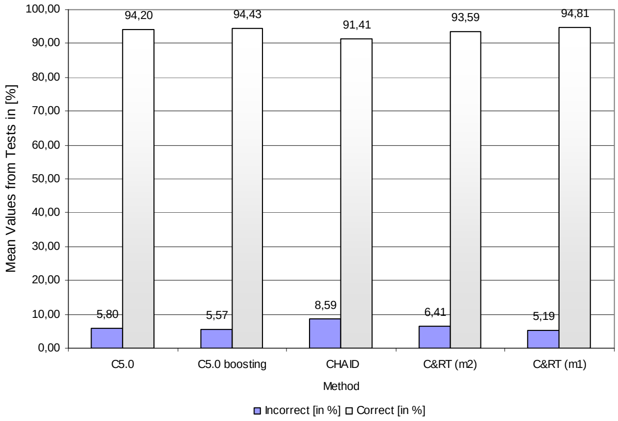


Fig.9 Comparison of methods: the mean values

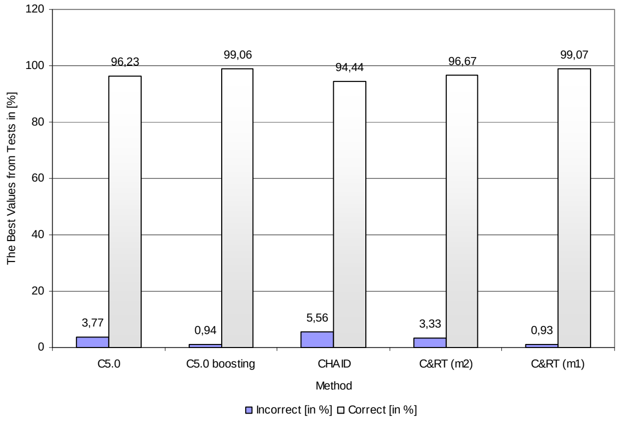


Fig.10 Comparison of methods: the best values

In terms of structure, these created models have these tree depths: the depth of trees created by the application of C 5.0 ana C 5.0 with boosting is 8 levels and the depth of trees by the CHAID is 4 levels, by the C&RT (m1) is 6 levels and by the C&RT (m2) is 4 levels.

The output of created decision trees is also decision rules. Example of rules extracted from the decision tree based on the C&RT algorithm (m1) is in the Fig.12 and the part of decision tree is in the Fig.13.

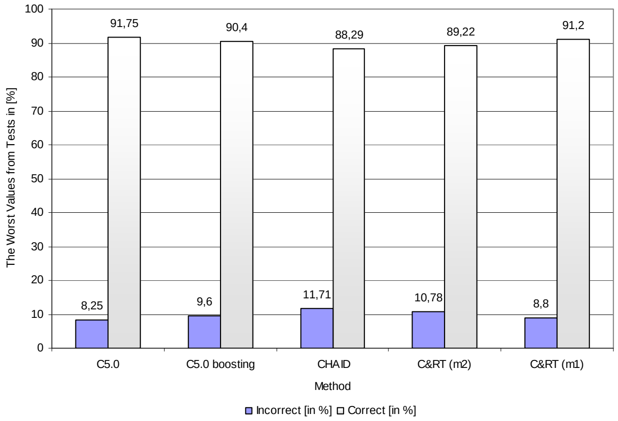
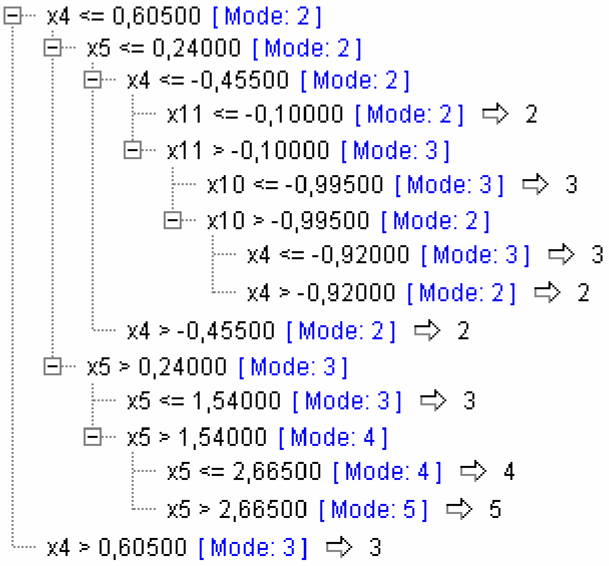


Fig.11 Results comparison of methods: the worst values



```
⊟ x4 <= 0,60500 [Mode: 2]
   ⊟ x5 <= 0,24000 [Mode: 2]
      ⊟ x4 <= -0,45500 [Mode: 2]
           x11 <= -0,10000 [Mode: 2] ⇨ 2
         ⊟ x11 > -0,10000 [Mode: 3]
              x10 <= -0,99500 [Mode: 3] ⇨ 3
            ⊟ x10 > -0,99500 [Mode: 2]
                 x4 <= -0,92000 [Mode: 3] ⇨ 3
                 x4 > -0,92000 [Mode: 2] ⇨ 2
        x4 > -0,45500 [Mode: 2] ⇨ 2
   ⊟ x5 > 0,24000 [Mode: 3]
        x5 <= 1,54000 [Mode: 3] ⇨ 3
      ⊟ x5 > 1,54000 [Mode: 4]
           x5 <= 2,66500 [Mode: 4] ⇨ 4
           x5 > 2,66500 [Mode: 5] ⇨ 5
  x4 > 0,60500 [Mode: 3] ⇨ 3
```

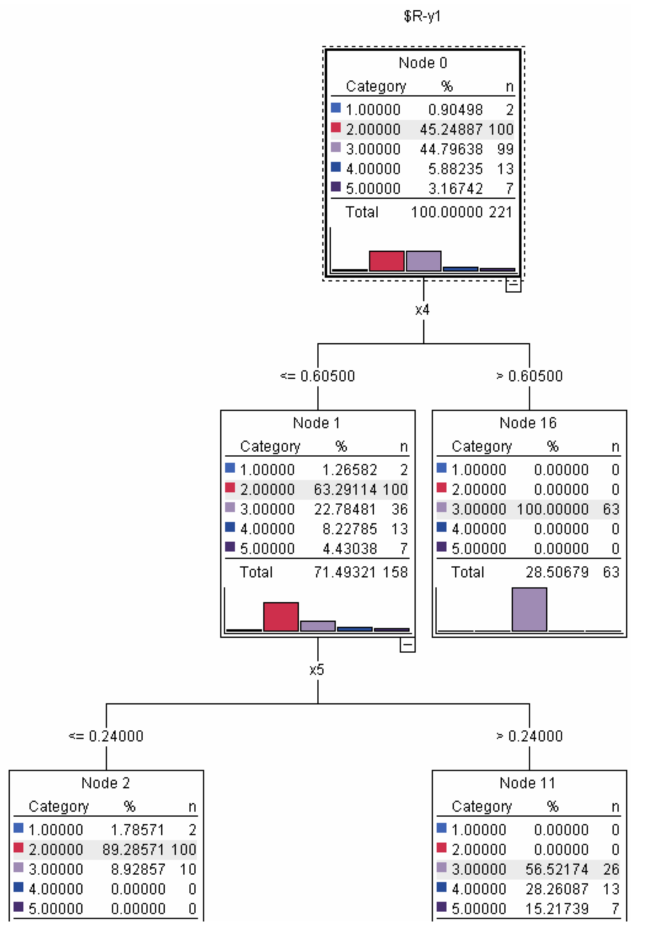
Fig.12 Decision Rules: the C&RT (m1) in Clementine 10.1

5 Conclusions

Air pollution belongs to basic environmental problems in the world. Improvement of air quality and quality other parts of the environment is a part of a sustainable development of regions and countries. For example in [3, 5, 7, 10, 12, 20, 21] we can see approaches to a solving of these problems.

In the Czech Republic air quality belongs also to very important and actual questions. We focused on the air quality modeling in Pardubice-Dukla locality. We collected daily observations of air polluting substances concentrations described by eleven attributes and we analyzed them. In the data preprocessing step we standardized data and used correlation. On the basis of result of correlation we eliminated two attributes x_7 and x_8.

We defined output variable y_k on the basis of air quality evaluation (Table 2). In this step it is possible to use other ways of the output definition, for example to use cluster analysis, neural networks, etc. These approaches are solved for examples in [8, 12].



Node 0

Category	%	n
1.00000	0.90498	2
2.00000	45.24887	100
3.00000	44.79638	99
4.00000	5.88235	13
5.00000	3.16742	7
Total	100.00000	221

Node 1

Category	%	n
1.00000	1.26582	2
2.00000	63.29114	100
3.00000	22.78481	36
4.00000	8.22785	13
5.00000	4.43038	7
Total	71.49321	158

Node 16

Category	%	n
1.00000	0.00000	0
2.00000	0.00000	0
3.00000	100.00000	63
4.00000	0.00000	0
5.00000	0.00000	0
Total	28.50679	63

Node 2

Category	%	n
1.00000	1.78571	2
2.00000	89.28571	100
3.00000	8.92857	10
4.00000	0.00000	0
5.00000	0.00000	0

Node 11

Category	%	n
1.00000	0.00000	0
2.00000	0.00000	0
3.00000	56.52174	26
4.00000	28.26087	13
5.00000	15.21739	7

Fig.13 Part of the decision tree

For the classification model creation we used decision trees. We focused on the C5.0 algorithm, boosting, CHAID and C&RT. Afterwards we analyzed and compared achieved results of classification. We state that the used methods give very similar results.

However, the best is the C&RT. Achieved mean accuracy rate is 94.81 %. Generally class 2 (good) and class 3 (favourable) belong to the most frequent classis (class 2 (43.93 %); class 3(49.53 %)). Proportion in the class 4 (satisfactory) is 3.37 % and in the class 5 (bad) is 2.8 %).

For the improvement of models it seems appropriate to use more variables that cause air pollution and work with more daily observations.

6 Future work

The areas where future investigations could be directed can be divided into three groups.

Firstly it is the investigation of theoretical context and the possibilities to use the proposed models. For example aplication of Kohonen maps, Rough sets theory or Fuzzy sets theory is in our opinion possible. Secondly it is further data analysis tools development and thirdly to gain data sets based on daily observations from other Czech republic regions and realize their analysis and comparison.

7 Acknowledgement

This work was supported by the scientific research project of Ministry of Environment, Czech Republic under Grant No: SP/4i2/60/07 with title Indicators for Valuation and Modeling of Interactions among Environment, Economics and Social Relations.

References:

[1] Czech Hydrometeorological Institute [online], URL http://www.chmi.cz, (in Czech).

[2] Gass SI, Harris CM, *Encyclopedia of operations research and management science* (Kluwer Academic Publishers, Boston, 2004).

[3] Graf, H. J., Musters, C.J.M., Keurs, W. J., *Regional Opportunities for Sustainable Development: Theory, Methods and Applications,* Kluwer Academic Publisher, 1999.

[4] Guidici, P. *Applied Data Mining: Statistical Methods for Business and Industry,* West Sussex: Wiley, 2003.

[5] Hajek, P., Olej, V. Air Quality Modelling by Kohonen's Self –organizing Maps and LVQ Neural Networks. *WSEAS Transactions on Environment and Development,* WSEAS Press, Isue 1, Vol. 4. January 2008, pp. 45-55.

[6] Han, J., Kamber, M. *Data Mining: Concepts and Techniques*, Morgan Kaufmann Press, 2001.

[7] Hersh, M., *Mathematical Modelling for Sustainable Development*, Springer, 2006.

[8] Jirava, P., Křupka, J., Classification Model based on Rough and Fuzzy Sets Theory, *WSEAS Computational Intelligence, Man-Machine Systems and Cybernetic,* 2007, pp. 199-203.

[9] Krishnan, R., Model Management: Survey, Future Research Directions and a Bibliography. *ORSA CSTS Newsletter,* Vol.14, No.1.

[10] Křupka, J., Olej. V, Obršálová, I., Multiple Criteria Decision Making in Environmental

System. *WSEAS Transaction of Systems,* Vol.5, No.1, 2006, pp.148-155.

[11] Maimon, O., Rokach, L., *Decomposition Metodology for Knowledge Discovery and Data Mining,* World Scientific Publishing, 2005.

[12] Olej, V., Hájek, P., Křupka, J., Obršálová, I., Air Quallity Modelling by Kohonen's Neural Networks, *WSEAS Environmental Science, Ecosystems & Development,* 2007, pp. 221-226.

[13] Pyle, D. *Business Modeling and Data Mining*, Morgan Kaufmann Publishers, 2003.

[14] Rokach, L., Maimon, O., *Data Mining with Decision Trees: Theory and Applications,* World Scientific Publishing, 2008.

[15] Rusell, S. J., Norvig, P., *Artificial Intelligence: A Modern Approach,* Prentice Hall, 2002.

[16] SPSS Inc. *Clementine® 7.0 User's Guide*, 2002.

[17] State Policy of Environment in the Czech Republic 2004 – 2010, Praha: Ministry of Environment, 2004, (in Czech).

[18] Turban, E. et al., *Decision Support Systems and Inteligent Systems,* Prentice Hall, 2005.

[19] Villareal, L.V., Kelleher, V. (ed.), Tietze, U. (ed.), Guidelines on the collection of demographic and socio-economic information on fishing communities for use in coastal and aquatic resources management. *FAO Fisheries Technical Paper*. No.439, 2004.

[20] Vongmahadlek, Ch., Satayopas, B., Applicability of RAMS for a Simulation to Provide Inputs to an Air Quality Model: Modeling Evaluation and Sensitivity Test, *WSEAS Transactions on Environment and Development,* Vol.3, No.8, 2007, pp. 129-138.

[21] Vonkeman, G. H., *Sustainable Development of European Cities and Regions,* Kluwer Academic Publishers, 2000.

[22] Witten, I. H., Frank, E., *Data Mining: Practical Machina Learning Tools and Techniques,* Morgan Kaufman, 2005.