DATA MINING FOR FOG PREDICTION AND LOW CLOUDS DETECTION

Juraj Bartok

MicroStep-MIS, spol. s r. o.
Čavojského 1
841 08 Bratislava, Slovakia
e-mail: jurob@microstep-mis.com

František Babič

Department of Cybernetics and Artificial Intelligence
Faculty of Electrical Engineering and Informatics
Technical University of Košice
Letná 9, 042 00 Košice, Slovakia
e-mail: frantisek.babic@tuke.sk

Peter Bednár

Centre for Information Technologies
Faculty of Electrical Engineering and Informatics
Technical University of Košice
Letná 9, 042 00 Košice, Slovakia
e-mail: peter.bednar@tuke.sk

Ján Paralič

Department of Cybernetics and Artificial Intelligence
Faculty of Electrical Engineering and Informatics
Technical University of Košice
Letná 9, 042 00 Košice, Slovakia
e-mail: jan.paralic@tuke.sk

Computing and Informatics, Vol. 30, 2011, 1441–1464

Jozef Kováč

Centre for Information Technologies
Faculty of Electrical Engineering and Informatics
Technical University of Košice
Letná 9, 042 00 Košice, Slovakia
e-mail: jozo.kovac@gmail.com

Ivana Bartoková

MicroStep-MIS, spol. s r. o.
Čavojského 1
841 08 Bratislava, Slovakia
e-mail: ivana.bartokova@chello.sk

Ladislav Hluchý

Institute of Informatics
Slovak Academy of Sciences
Dúbravská cesta 9
845 07 Bratislava, Slovakia
e-mail: ladislav.hluchy@savba.sk

Martin Gera

Department of Astronomy, Physics of the Earth and Meteorology
Faculty of Mathematics, Physics and Informatics
Comenius University, Bratislava, Slovakia
e-mail: mgera@fmph.uniba.sk

Abstract. This paper describes our contribution to the research of parametrized models and methods for detection and prediction of significant meteorological phenomena, especially fog and low cloud cover. The project covered methods for integration of distributed meteorological data necessary for running the prediction models, training models and then mining the data in order to be able to efficiently and quickly predict even sparsely occurring phenomena. The detection and prediction methods are based on knowledge discovery – data mining of meteorological data

using neural networks and decision trees. The mined data were mainly METAR aerodrome messages, meteorological data from specialized stations and cloud data from special airport sensors – laser ceilometers.

Keywords: Meteorological prediction, aviation, fog, cloudiness, data mining

Mathematics Subject Classification 2010: 68T30, 68T05, 68T10

1 INTRODUCTION

The interest in short-term weather warnings with higher localization accuracy has increased recently. The significant and hazardous meteorological events influence many parts of the society like traffic, agriculture, tourism, power generation, construction industry and public safety.

Fog impacts transportation activities by air, road or sea. An improvement of fog prediction methods is of importance to the human society as a whole.

Low cloud cover over a significant portion of the sky has impact on aircraft operations, and can result even in a temporary closing of airport for landing or take-off (according to ICAO[1] rules [26]). Low cloud cover is a mandatory component of routine observations and forecasts issued for each and every international airport worldwide.

Data mining is emerging as a suitable method for extracting patterns from extensive sets of heterogeneous data related to prediction of meteorological phenomena [4, 5].

This paper starts with short description of studied phenomena and continues with presentation of related work. Next, we elaborate on the possibilities of enhancing short-term prediction using data mining. For this purpose we performed data mining process on real meteorological data and present the whole process and its results in detail here.

For the purpose of performing this research, we have created a consortium of business (MicroStep-MIS, spol. s r. o.) and scientific organizations (Institute of Informatics of the Slovak Academy of Sciences (II SAS), Faculty of Electrical Engineering and Informatics of the Technical University of Košice (FEI TUKE)) within the *Data Mining Meteo* (DMM) project.

1.1 The Data Mining Meteo Project

The project consortium consists of business partner with extensive experience in meteorology (both commercial and research projects [8]) and two scientific partners with experience in data integration and data mining [7, 9, 10, 11]. In the past, the

[1] International Civil Aviation Organization

partners already cooperated on a scientific project [12]. In more detail, MicroStep-MIS spol. s r.o. develops, deploys and markets monitoring and information systems in the fields of meteorology, seismology, radiation and emission monitoring and crisis information systems. The product range covers measurement, real-time data collection, data processing, archiving and tools for analysis and decision support. The products connected to this project include meteorological modeling and warning systems with requirements for fog prediction capability. The most frequently installed products include the Airport Weather System, where fog forecast and ceilometer cloud data analysis is of great interest to airport users. Institute of Informatics of the Slovak Academy of Sciences (II SAS) is one of the leading grid computing research institutions in Slovakia. Faculty of Electrical Engineering and Informatics of the Technical University of Košice (FEI TUKE) has long time experience with data mining in various domains including the mining on unstructured textual data. The main contribution of FEI TUKE is the design of data preprocessing tasks and selection and application of the data mining methods.

2 SIGNIFICANT AND HAZARDOUS METEOROLOGICAL PHENOMENA

2.1 Fog

We begin with an example of currently used approach to the prediction of fog (see [27]). It starts with a common 3D meteorological model executed for a limited region; its outputs are converted using empirical formula into visibility [3]. This approach by itself cannot achieve results of satisfactory quality and common meteorological models can fail to handle inversion weather conditions, which commonly produce fog. Therefore there are several experimental models in development worldwide, which further process the results of common meteorological model: 1D physical fog modeling methods or statistical post-processing of model outputs [1, 2]. The result is then interpreted by a meteorologist, who takes into account further factors – mainly his/her experience with meteorological situations and local conditions, satellite imagery, real-time data from meteorological stations suggesting that fog has started to form, or that conditions are favorable for the occurrence thereof, conditions of the soil in the target locations, snow cover, recent fog occurrences, etc. We used data mining to create short term forecasting model capable of autonomous operation. The model is usable also as a supporting tool for meteorological forecaster.

2.2 Low Clouds

Two characteristics of clouds are the most important in airport operations – the cloud base height and the extent of cloud cover. Laser ceilometers are sensors routinely deployed at airports for measurement of cloud base heights above the

points of their installation. The extent of cloud cover (cloud amount) is assessed by a meteorological observer and reported periodically (1 hour or 30 min period) and in cases of significant change. Sometimes the ceilometers data is also used to determine the cloud amount using the FAA (Federal Aviation Administration) method, where the result is a simple combination of laser reflection counts in different height categories. The original method uses data from only one ceilometer. We use data mining to obtain more comprehensive 2D information on cloud amount. We mine data from several deployed ceilometers installed to serve for different runways, making the cloud information continuous and available in fully automated operation also.

3 RELATED WORK

In this section we present some interesting applications of data mining methods in meteorological data analyses.

Fog prediction at airports has been solved in many localities around the world, e.g. the research group from Italian Aerospace Research Center developed several fog classifiers based on Bayes networks [13]. The same method was used in [14] both for fog and low cloud prediction. In this case the authors specified the methodology for creation of basic network structure based on information from observations in the examined locality, available specific existing guidance and expert opinions. This basic model was further adapted on data collected from major Australian airports in order to adapt and test it in various conditions. The achieved results represent more than 55 positive forecasted fogs in row instead of previous operational 7–8 cases. On the other hand, identification of important parameters with strong influence on fog formation at the International Airport of Rio de Janeiro was performed based on neural network classification with rules extraction algorithm [15]. The 20-years historical dataset was used to build a classification model with obtained error of about 6 %. Very important were the rules extracted from created model of neural network.

Similar approach was used in [19] to classify collected historical data from Melbourne airport in two basic categories (fog and no fog) and further extraction of interesting associative rules based on The Combinatorial Rule Model [20]. The extracted rules were evaluated within domain experts and the best ones were stored into knowledge base and will be used in expert system for decision support. The neural networks algorithms have broader audience in meteorological data mining due to their ability to learn important characteristics from past historical data and use them for prediction [16, 17]. This advance was applied for fog prediction at the Canberra International Airport based on 44-years database of standard meteorological observations [18]. The resulted neural network produced values for 3h prediction with 0.937 cross-validated mean value.

The interesting business case represents exploitation of several artificial intelligence methods for prediction of delays in aircraft management at the Frankfurt

airport according to relevant weather conditions. The travel time was used as target attribute for algorithm of linear regression, neural networks, decision trees and fuzzy clustering with up to 20 % higher prediction accuracy as in the case of the simple mean estimators [21].

The algorithm of K-nearest neighbor was used in [22] to develop a system for climate prediction. These experiments were performed on two datasets containing 40 thousand or 80 thousand records to predict around 17 climatic attributes, as e.g. binary attribute fog occurrence with 96.66 % accuracy for the smaller dataset and 98.33 % for the larger one.

4 MINING OF METEOROLOGICAL DATA

In this chapter we describe the process of knowledge discovery in databases (KDD) [6] applied to the studied problem. The goal of this process is to gain new and potentially useful knowledge from extensive databases. The knowledge is in the form of a set of generalized rules allowing us to better understand domain-specific relations, and to better predict future cases. This process is generally iterative as well as interactive (i.e. it cannot be fully automated).

The process of knowledge discovery in databases is detailed in the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology. CRISP-DM represents an industry initiative to develop a common and tool-neutral standard for whole data mining process. This methodology was designed based on past practical experiences of various companies gained when solving data mining tasks. The whole data mining process can be understood as a life cycle containing six main phases, see Figure 1.

1. Business (goal) Understanding

Understanding the Goal of the Fog Prediction. In its projects, the business partner of this DMM consortium has been running a fog prediction system composed of in site and remote monitoring parts and physical prediction models. (Under physical prediction models we mean numerical weather prediction models based on numerical solving of physical equations describing atmospheric processes.) The output of models, as well as real-time measured data, are left for operator's interpretation. Operator has advantage to fully automatic system, because typical meteorological situations may occur in which model overpredicts/underpredicts, and experienced operator can consider this experience in his/her decision making. The expected outputs of the data mining model were as follows:

- visibility prediction for three hours ahead (the usual short-term warning system range);
- categorical prediction – whether the visibility falls below 1 000 m or not (1 000 m is threshold in meteorological definition of fog [27]).

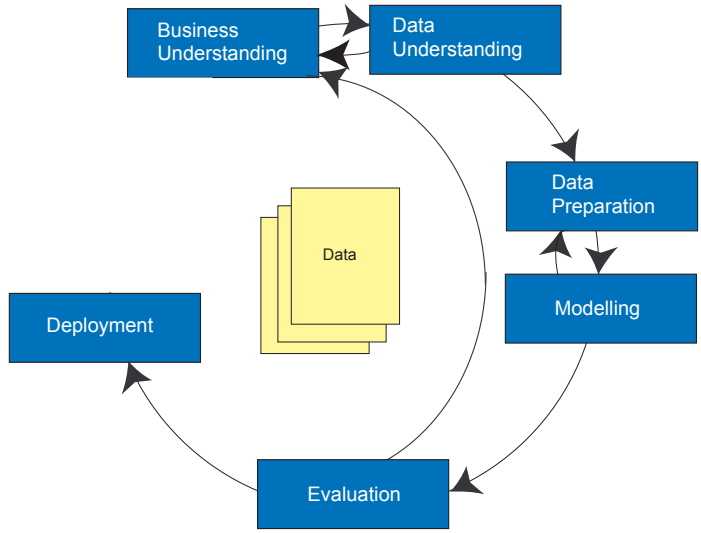


Fig. 1. Cross industry standard process for data mining

An important requirement was that the quality of the final fog prediction model should by measured be means of true skill score parameter, which is defined as recall minus false alarm; $Recall = TP/(TP+FN)$ and $False\ alarm = FP/(TP+FP)$ ($TP(FP)$ is the number of true (false) positive and $TN(FN)$ is the number of true (false) negative examples, respectively). True skill score larger than 20 is considered as good and suitable for the target application.

Understanding the Fog Prediction Goal Using Specialized AWS Data. According to CRIPS-DM methodology target event has to be defined and relevant data have to be selected. According to final user requirements, which were targeted to higher traffic safety on highways, fog had to be predicted with sufficient advance before commencement of morning rush hour. We decided to predict the target event at 2:00 AM. Formation of fog lasting longer than 8 minutes between 2:02 and 7:52 was considered as positive fog event.

Two locations were analysed – Nazwa and Um Al Moameneen, UAE. 365 observations per year from 2:00 AM were selected from each location. There were 7 days missing in Nazwa due to sensor maintenance in those days. An interesting question was if is possible to create single model with stable prediction on both locations or separate models have to be created for both locations in order to achieve sufficient prediction score.

Understanding the Low Clouds Data Mining Goal. As the original method for low clouds prediction uses data from only one ceilometer, one of

the important novelty is incorporation of three ceilometers (installed to serve for different runways) at once. Of course, standard meteorological information in form of METAR and SYNOP messages should be used. The goal was to achieve more than 90 % prediction accuracy for automatic detection of low clouds coverage. It is most important to correctly detect clouds situations leading to necessary constraints in air traffic at the airport. More discrete levels of low clouds coverage need to be taken into account (not just binary prediction as it was the case for fog).

2. Data Understanding This phase started with the selection of relevant input data from available data sources. In this work we utilized the measured data and also physically predicted information to extract fog prediction by the data mining model. The input data are grouped by source:

Meteorological messages with measured data. These data are available from standard network of meteorological stations (text codes SYNOP, METAR [28]). They are available in 1-hour and 3-hour intervals. Messages immediately preceding the forecasting interval are of importance, as well as recent history (last 72 hours).

Data from specialized automated weather stations (AWS) measuring elements significant for fog occurrence. Such stations comprise standard meteorological sensors plus soil humidity and temperature in different depths, visibility sensors, and additional air temperature sensors. The data are available each 2 minutes.

3D physical model weather prediction. Predictions for 48 hours are available each 6 hours, with 1.8 km horizontal resolution and 42 vertical levels.

1D physical model weather prediction. Predictions for 12 hours are available each hour above selected station sites, with finer vertical resolution.

Understanding the Data for Fog Prediction. The input data consist of the measured or observed values of physical measurements or meteorological phenomena. For the selected target geographical area, we have obtained data in two formats: in METAR format and in SYNOP format for time extent of 10 years. METAR data were measured with one hour period and SYNOP messages were measured each 3 hours. Besides of the measured or observed values, we have decided for this task to use also predicted values of the attributes computed by standard physical prediction model. We have executed 700 runs of the physical model for 48 hours prediction for each run. Data were computed with the time resolution of one hour (i.e. we have obtained 33 600 records).

Expected predictors among meteorological elements were *temperature*,*relative humidity*, *visibility*, *wind speed*, *wind direction*, *precipitation*, *cloud amount*, *soil temperature* and *moisture*, etc. We have had the mentioned elements available both measured and forecasted by physical models. Temperature drop is necessary to start fog, especially in radiation fog cases. Relative humidity is one of the

measures of water content in atmosphere and shows how near the air is to saturation and condensation of fog droplets. Visibility is connected to fog presence and density, although visibility can be lowered also by other means (rainfall, snow, dust). Wind is an important selector of radiation fogs – low wind speeds favor forming of radiation fog, while both zero and strong wind speeds do not. Wind direction is important in sea shore areas – wind can transport humidity form the sea. Soil moisture informs if there is a sufficient amount of source humidity in the soil. Soil temperature in several depths (a profile) influences how the soil will cool the air to create the necessary temperature drop.

Understanding the Data for Fog Prediction Using Specialized AWS Data. The data form specialized AWS were sampled with high frequency, each 2 minutes. In addition to standard meteorological elements available in METAR reports, the following has been provided:

- soil temperature in three depths of 5, 10 and 25 cm;
- soil humidity in three depths of 5, 10 and 25 cm.

A camera was also available, pointing towards the surrounding landscape, with 10 minutes imaging frequency. It provides additional visual information for users of sensor data and proved useful for verification of visibility conditions. Within this work, it was used to verify the target attribute – fog occurence.

Data from two different locations from period between 1st Jan 2010 and 31st Dec 2010 were available for this task. Derived attributes were added to final dataset:

- time to the next fog (minutes) – only for definition of target event;
- time from the last fog (minutes);
- dew point deficit (Temperature – Dew point).

Understanding the Low Clouds Data. Two data sources have been used (data from ceilometers and METAR data) and gradually integrated into one dataset after all necessary preprocessing operations. The first data source included the data measured from three ceilometers deployed at the Bratislava airport. These three ceilometers measure the height of a cloud base above the ceilometer in 15s intervals. We used records measured between 2007 and 2010. The whole dataset contained 93 CSV files with more than 11 million records. After making the identified data available, an initial data examination was performed, leading mainly to verification of the quality of the data.

The standard record consists of the following attributes:

- Time and date;
- Identifier (ID) of ceilometer which perform the measurement;

- Detection status – nominal attributes with several values as no low cloud cover, one level of cloud detected, two levels of cloud detected, three levels of cloud detected, insensible clouds, no measurement;
- Warning and Alarm info – attribute identifying actual system operational status of the ceilometer, e.g. self-test, alarm and warning;
- Level 1 – height of the first detected level of cloud and visibility in vertical course;
- Level 2 – height of the second detected level of cloud;
- Level 3 – height of the third detected level of cloud, in many cases this attribute has value ///// which means no level of cloud detected;
- Information about ceilometer – not important for our data mining goal.

The second data source included the above-mentioned METAR messages. We have extracted data from the same time period as for ceilometers measurements taking into account 30-minutes periodicity of the METAR messages broadcasting. The extracted data were available as one CSV file in standard METAR format within 45 attributes. We have selected only 15 from them as important for our data mining goal:

- time and date;
- *C1* – quantity of clouds in the first level of cloud cover;
- *BASE1* – height of cloud basis for first level;
- *TYPE1* – type of storm cloud for first level;
- *C2*, *BASE2* and *TYPE2* – the same as C1, BASE1 and TYPE1 but for the second level of cloud cover;
- *C3*, *BASE3* and *TYPE3* – the same as in the previous two cases, but for the third level of cloud cover;
- *C4*, *BASE4* and *TYPE4* – the same as in the previous cases, but for the fourth level of cloud cover, but this values appear only in unique situations;
- *CAVOK* (Ceiling And Visibility Okay) – indicates no cloud below 1 500 m, visibility of 10km and no cumulonimbus at any level;
- *CTOT* – target attribute describing the level of low cloud cover (see Table 1).

Value	Description – level of low cloud cover – x/8 part of sky
0.0	0/8
1.5	1/8, 2/8
3.5	3/8, 4/8
6.0	5/8, 6/8, 7/8
8.0	8/8
9.0	Other meteorological phenomena
Missing (-1)	Unsuccessful measurement

Table 1. Meanings of values for the target attribute *CTOT*

The data examination started with computation of basic statistical characteristics for key attributes, identification of hidden dependencies within correlations and visualizations of values distribution. The most important dependency represents relations between *CAVOK* attribute value equal to 1 and all other attributes. This value represents good weather conditions without significant clouds.

3. **Data Preparation** Data preprocessing is usually the most complex and also most time consuming phase of the whole data mining process (usually taking 60 to 70 percent of the overall time). The goal of relevant data operations is to prepare final version of the dataset based on requirements of selected algorithm that will be applied in the next phase.

   The first step in this phase was transformation of datasets at first into the comma-delimited files and then import into the MySQL database in order to simplify manipulation with data using the SQL scripts or external applications (i.e. SPSS Clementine in our case).

   Fog Prediction Data Preparation. After the data import we have performed the following data preprocessing steps:

   - Records with invalid dates (month or day) were removed from the dataset.
   - Duplicate records with the same time were merged.
   - Time of measurement of all records was aligned into the time intervals (1h for METAR and 3 hours for SYNOP), i.e. for example for METAR, if the time of measurement was from the interval $\langle$*HH:00*, *HH:30*$\rangle$, it was aligned to *HH:00*, otherwise to *HH+ 1:00*.
   - Records were completed into continuous time series (if some time interval was missing, we have added new record with "missing attribute" values).
   - Missing values of some attributes were replaced by the indicator value. This was the special case when the weather condition was good (indicated by attribute $CAVOK = 1$) and then attributes were not measured.

   Additionally we have added the following attributes computed from the source data:

   - target binary attribute fog/no fog (this attribute has to be extracted from the attribute which encodes special weather conditions)
   - relative humidity computed on the basis of Magnus formula [27]
   - difference between temperature and dew point (since according to the domain knowledge, fog arises when the temperature is approaching the dew point)
   - trend for some attributes.

   After we have formed the basic table with all relevant attributes, we have created tables with time window where each record for time t has N sets of historical attributes from METAR or SYNOP measured in time $t, t-1, t-2, \ldots, t-N$,

and target attribute for time $t+3$ (the goal is the short time prediction for next 3 hours). In some cases we have also added into the time window records N predicted sets of attributes computed by physical meteorological model for time $t+1, t+2, \ldots, t+M$.

Parameters M and N were estimated in the subsequent step which was the selection of attributes. First we have removed all attributes where the number of missing values was higher than 70 % and then attributes were reduced according to the Pearson correlation coefficient computed against the target attribute. The final dataset was then the input to the modelling phase.

All steps were implemented either as a command shell utility program implemented in C# or as a data streamscript in SPSS Clementine.

Low Clouds Data Preparation. In the next step we have replaced the values of Level 1, Level 2, and Level 3 attributes higher than 1 500 m with a special markup value to identify cases where clouds were detected above the height for good weather conditions. We have also assigned height 0 to cases where clouds were not detected (in original data, the values marked with /////).

After that, the attributes Detection status, Warning and Alarm info, ID and Information about ceilometer were discarded from the dataset and three tables for ceilometers data in MySQL database were integrated into one common dataset based on comparison of date and time for each record. During this aggregation, time intervals of measurements from ceilometers were aligned to the same discrete time line with the period of 15 seconds.

Finally, the processed data from ceilometers were merged with appropriate METAR records. One METAR message broadcasted every 30 minutes corresponds to 120 ceilometer records measured every 15 seconds. The final preprocessed dataset contained 1 081 columns (120 time points $\times$ 3 ceilometers $\times$ 3 attributes (Level 1, Level 2 and Level 3) for each ceilometer + 1 target attribute $CTOT$ extracted from METAR messages).

This operation resulted in new distribution of target attribute $CTOT$, see Figure 2.

4. Modeling The modeling phase deals with application of suitable data mining algorithms on the preprocessed data.

Fog Prediction Modeling. For modeling we have selected decision trees and neural networks models implemented in SPSS Clementine. All attributes of learning algorithms (attribute split criterion, pruning method for decision trees and number of neurons on hidden layer for neural networks) were systematically optimized using the 10-fold cross validation method on the training dataset. Besides, all created models also provide the real value output, which estimates the confidence of the model prediction. Using the 10-fold cross validation testing, we

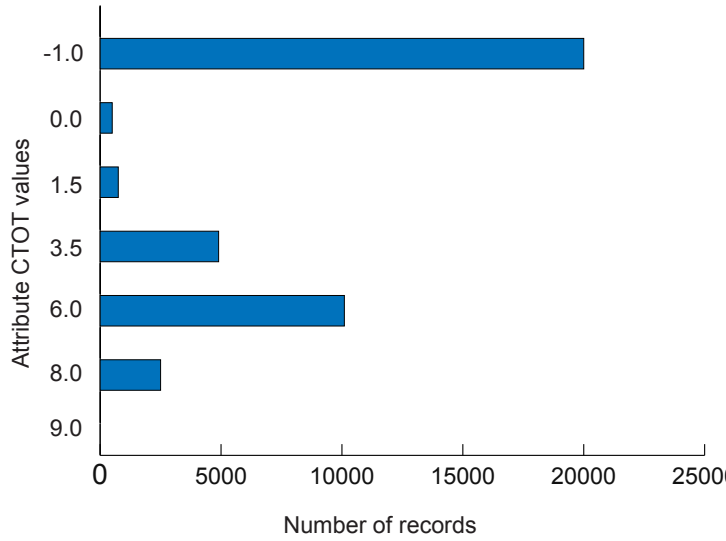


Fig. 2. The new distribution of *CTOT* attribute in merged dataset

have also optimized the threshold of the confidence estimation used for positive classification, i.e. for threshold K, if the confidence is lower than K, the record is assigned to the negative class "no fog".

Since weather conditions can be different when the fog is just forming and when it is dissipating, we have decomposed the classification problem into two sub-problems and we have created separate models for each. One model predicts start of the fog conditions after 3 hours, and the second one predicts that the fog conditions will continue after the 3 hours. Final prediction is given by combination of both models.

The main problem during fog prediction modeling was the unbalanced data, i.e. we have more than 22 000 of cases with good weather conditions but only 155 cases of fog occurrence. However, such unbalance is natural for this type of problems because we are interested in case which is special and can be rare. In order to balance data and improve efficiency of created models, we have tried two methods, which reduced the number of negative examples and improved data balancing. In the first method, we have chronologically selected only the negative cases (i.e. good weather conditions) just before the fog was formed. In the second method we have first clustered data into 155 clusters (the number of positive cases in the original dataset). Then we have formed the training dataset from positive and negative examples closest to the centroids of these clusters.

Fog Prediction Modeling Using Specialized AWS Data. Two different locations allowed performing cross-validation based on location. CHAID decision tree was trained on all data from one location and evaluated on another and vice versa.The results revealed two important discoveries: on both validation sets the score was over 50 %, AUC the over 94 % and both models were fit to training data where score was significantly higher (see Table 2).

Train Loc.	Test Loc.	Posit.	Negat.	Total	Fog Rate	Score	AUC
Moameneen	Nazwa	35	323	358	10.8 %	51.4 %	94.8 %
Moameneen	Moameneen	32	333	365	9.6 %	72.3 %	99.1 %
Nazwa	Nazwa	35	323	358	10.8 %	61.2 %	97.8 %
Nazwa	Moameneen	32	333	365	9.6 %	56.8 %	96.4 %

Table 2. Cross validation based on location (Loc.)

We decided to perform 2-fold cross validation in order to gain more confidence in predictive power of selected predictors. On validation sample the score was again around 50 % but absolute difference between training and validation was reduced to around 3 %. AUC variable was also reduced, see Table 3.

Train Loc.	Test Loc.	Posit.	Negat.	Total	Fog Rate	Score	AUC
Both 50 %	Nazwa	35	323	358	10.8 %	48.6 %	96.9 %
Both 50 %	Moameneen	32	333	365	9.6 %	51.7 %	96.3 %
Both 50 %	Both another 50 %	35	329	364	10.6 %	50.8 %	97.2 %
Both 50 %	Both 50 %	32	327	359	9.8 %	48.5 %	95.8 %

Table 3. 2-fold cross validation and validation based on location

The final model was trained on union of both locations and validated on all previously created subsets. The score is very similar to that achieved in previous experiments – slightly above 50 % and AUC above 97 % (see Table 4).

Train Loc.	Test Loc.	Posit.	Negat.	Total	Fog Rate	Score	AUC
Both 100 %	Nazwa	35	323	358	10.8 %	52.1 %	97.1 %
Both 100 %	Moameneen	32	333	365	9.6 %	56.8 %	97.6 %
Both 100 %	Both 50 %	35	329	364	10.6 %	50.8 %	97.2 %
Both 100 %	Both another 50 %	32	327	359	9.8 %	56.3 %	97.4 %
Both 100 %	Both 100 %	67	656	723	10.2 %	51.5 %	97.3 %

Table 4. Validation of final decision tree trained on union of data

In addition to overall score we include details on performance evaluation.

$Score\,(\%) = POD\,(\%) - FAR\,(\%)$ is common standard in evaluation of prediction power. Target variable was binary and binary prediction is expected on input. Our model has not produced binary prediction directly. Instead, it calculated probability of target event and we made the final binary prediction only by

setting the cut-off probability. Area Under Curve (%) was used as secondary prediction power indicator.

Decision trees estimate probabilities in leaf nodes and prediction is discrete attrib-ute. There are not more different probabilities in scored data than leaf nodes in used decision tree. We ordered model predictions by those discrete probabilities descending from highest to lowest probability and calculated cumulative positive and negative cases counts for all of them. This approach enabled score analysis over all possible probability levels where optimal cut-off score can be observed, and identification of the final set of decision rules used in positive prediction (all leaf nodes with P(Yes) $\geq$ cut-off); this improves business interpretability of the model and trust of end users.

Cutoff point may be selected based on various criteria, not only on score. Maximum score is at first 2 nodes with estimated probability higher or equal to 47.80 %. Including first four nodes into model score 41.5 % is still achieved and in addition to that 100 % (67) of positive cases are identified only on 24 % (171/723) of total observations (see Table 5).

Tree Node	Estimated Prob.	Observed Prob.	Cumul. Posit.	Cumul. Negat.	Total	Cumul. POD	Cumul. FAR	Score	Cumul. AUC
4	89.1 %	90.9 %	40	4	44	60 %	9.1 %	50.6 %	0.2 %
3	47.8 %	47.6 %	50	15	65	75 %	23.1 %	51.5 %	1.3 %
9	18.2 %	15.0 %	53	32	85	79 %	37.6 %	41.5 %	3.3 %
7	17.0 %	16.3 %	67	104	171	100 %	60.8 %	39.2 %	13.1 %
8	1.9 %	0 %	67	154	221	100 %	69.7 %	30.3 %	20.8 %
5	0.2 %	0 %	67	656	723	100 %	90.7 %	9.3 %	97.3 %

Table 5. Validation of final decision tree trained on union of data

Low Clouds Modeling. In this case we have selected decision trees as the main data mining algorithm, mainly because they are robust on data sets with missing data and unbalanced targets and because models can be simply interpreted by experts.

The proposed experiments were performed within the data mining environment called SPSS Clementine that offers several algorithms for decision trees generation, e.g. CHAID, QUEST or C5.0. The prepared dataset was divided into training and testing set with several strategies to obtain optimal conditions for modeling. In order to obtain optimal results, all parameters of the algorithms were tuned by testing several strategies to divide input dataset into training and test sets. We have tested random division (90 % for training and 10 % for testing), random division (90 % for training, 10 % for testing) with stratification and 10-fold cross validation with stratification.

The first experiments generated classification models with accuracy about 70 % which was later improved to 80 % with stratified division of training and testing examples.

Correct/Predicted value	−1.0	0.0	1.5	3.5	6.0	8.0	9.0
−1.0	**2 002**	86	114	162	197	21	1
0.0	0	**1**	0	0	0	0	0
1.5	0	0	**4**	0	0	0	0
3.5	1	0	5	**270**	9	0	0
6.0	14	9	6	56	**816**	67	0
8.0	1	0	0	0	14	**173**	0
9.0	0	0	0	0	0	0	**3**

Table 6. The first obtained results in the form of coincidence matrix

The high number of missing values in target attribute *CTOT* had strong influence on our results, because related records were eliminated from the classification model creation process. This fact decreased the amount of possible training data to 50 % and worse conditions for the algorithm to learn all necessary examples for particular categories of low cloud cover. We returned to the preprocessing phase to solve this issue based on identified dependency between *CTOT* and *CAVOK* attributes (if $CAVOK = 1$ then $CTOT = 0/8$). This operation resulted into new distribution of *CTOT* attribute with only 280 missing values (eliminated for next modeling).

This step led to an increase in accuracy of C5.0 model (about 82 %), but the problem with incorrectly classified records (mainly for class 0) still remained; see Table 7.

Correct/Predicted value	0.0	1.5	3.5	6.0	8.0	9.0
0.0	**2 001**	112	144	207	25	0
1.5	1	**7**	2	0	0	0
3.5	14	11	**261**	37	8	0
6.0	39	1	57	**748**	67	0
8.0	1	1	3	39	**174**	0
9.0	0	0	0	0	0	**3**

Table 7. The second experiment results in the form of coincidence matrix

In order to improve the final predictive accuracy we have also tested other alternatives to solve this problem. Significant improvement was obtained with approach of two-step classification. This approach consists of two main steps:

1. Binary classification with *CAVOK* as target attribute in order to separate the records describing good weather conditions ($CTOT$=0/8). These records were eliminated for the next step where only the records representing cloud cover were used for modeling. We have used algorithm C5.0 to create a model that had final accuracy around 94 %. Detailed results are provided in Table 8. It has to be noted that in our applications, effects of prediction errors of

Correct/Predicted value	0	1
0	**1 965**	175
1	42	**1 711**

Table 8. The binary classification result in the form of coincidence matrix

different types on traffic conditions are not symmetric. This means that it is less severe when the conditions are good and the classifier predicts cloud cover (175 cases) than when the classifier gives incorrect prediction of good conditions for cloudy weather (42 cases).

2. In classification of records representing detected cloud cover into four categories of $CTOT$ attribute (1.5, 3.5, 6.0, 8.0), value 9.0 was eliminated because of its low frequency in the dataset (only 3 occurrences). The generated model classified data with 86.71 % accuracy.

5. Evaluation

Fog Prediction Model Evaluation. To evaluate the model, we have used 10-fold cross validation computed 10 times. The results were then averaged for 10 repetitions of the cross validation and compared between the models. For comparison, we have also computed standard deviations. As the main efficiency measurements for fog prediction, we have evaluated the score defined as a difference between recall and number of false alarms (false positive records). Overall results are reported in Table 9, where for start of fog decision trees with clustering of negative examples for data balancing have been identified as the best model and for continuation of fog it was neural network with chronological sampling for data balancing and prediction of physical model as additional input.

	Score	Weight
Start of fog (best model)	1.12	60.3
Fog continuation (best model)	73.47	39.7
Average	29	

Table 9. Classification results for fog prediction

Accuracy of the prediction that fog will start exactly after 3 hours is quite low mainly because of higher number of false alarms. However, such a result is still plausible and overall prediction is good and higher than expected (the goal was to achieve score 20, and the average weighted score of our models is 29). We have also found that during the modeling, the standard deviations of scores were quite stable for both neural networks and decision trees, which is the indicator that our models are robust and not overfitted.

Low Clouds Model Evaluation. In this phase, we have evaluated all obtained results, used approaches and methods based on specified business goals. The whole performed data mining process represents continual and iterative activity that resulted into final classification model with 91.82 % accuracy (see Table 10). This value was computed as weighted average of both results from two-steps classification. The final result can be described as significant improvement of actual used methods for low cloud cover prediction.

Correct/Predicted value	1.5	3.5	6.0	8.0
1.5	**119**	9	2	0
3.5	5	**318**	27	1
6.0	4	83	**944**	88
8.0	0	1	15	**152**

Table 10. The final results in the form of coincidence matrix

We can see from the table above that most of the errors are for class 6.0 (the interval of cloud coverage from 5/8 to 7/8) which was incorrectly classified as full cloud coverage in 88 cases, or from 3/8 to 4/8 coverage in 83 cases. We can conclude that there are only 8 cases where the coverage prediction differs by more than 2/8 from the correct value.

Evaluation of Fog Prediction Using Specialized AWS Data. The resulting tree is quite simple but powerful in predicting, better than the previous model. The final version has 6 leafs (Figure 3). The most important predictors are current Visibility and current relative humidity. Soil temperature on level 3 has minor significance.

The revealed predictors are a subset of expected predictors. According to the authors, the two most important predictors are in accordance with expert forecasters experience. There is none of the conditions sufficient for fog formation (e.g. during many nights the visibility starts to drop, but will not reach the fog threshold of 1 000 m unless there is humidity above certain limit. The exact value of this limit (92 % in our case in Node 4 in Figure 3) is a valuable information for forecasters, as it is not the same in different locations and varies with local climatic conditions); however, this exact combination is significant and prejudices fog formation strongly.

6. Deployment The business partner of this DMM consortium plans to use the generated models within already developed and operationally running products. The fog prediction model will be incorporated into larger visibility monitoring and forecasting project IMS Model Suite [25], where physical models are already applied. It will also help as a supporting tool for manned forecasting office, where the forecaster firstly checks model outputs and then issues final forecasts. Both generated models, for fog prediction and for low clouds, extend the functionality

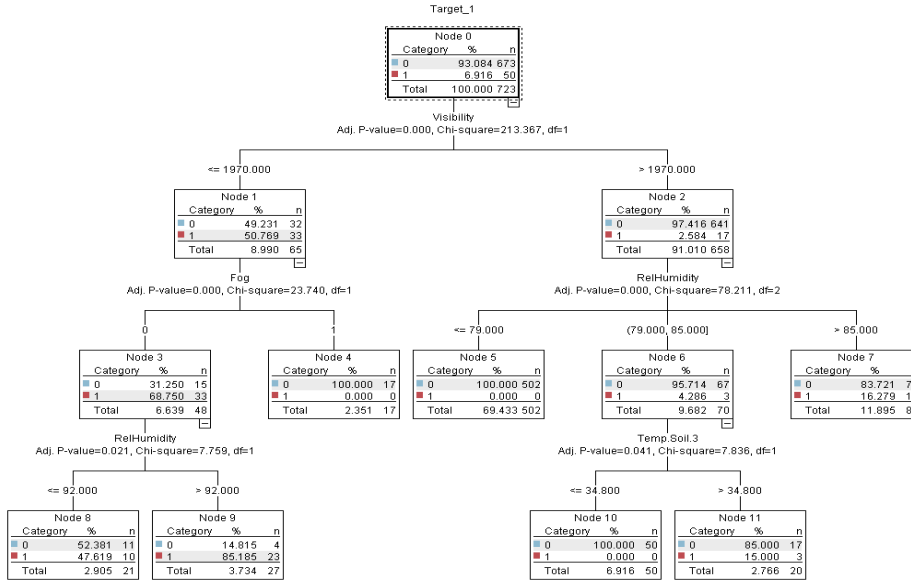


Fig. 3. Final decision tree for fog prediction

of the IMS Airport Weather Observing System [24], installed in more than 20 countries.

Before the deployment into real operation, we plan to run the system in testing operation for one year at a selected airport. The operation will be evaluated both statistically and subjectively by the airport users. We hope that this testing operation will also confirm the method as suitable for practical use.

5 CONCLUSIONS

In this paper we have described data mining approach for prediction of significant and potentially hazardous meteorological phenomena, namely fog and low clouds. According to the results, our models are comparable to the existing methods based on physical modeling and empirical rules (in case of fog) or manned observation (in case of low clouds). We have implemented whole chain of data pre-processing tasks, which extract and integrate data from various meteorological sources. Also, to some degree our models can be compared to the other data mining methods presented in the "Related work" chapter, but it has to be noted that the results are highly influenced by fog climatology in the tested area and by quality and extent of the training data. In our research, the most successful method for data mining were the decision trees. They not only yielded the highest score, but it was possible for meteorologists to compare the resulting rules with their knowledge and local experience.

The developed fog models score (POD – FAR) reached 29 % and 51 % in case of METAR dataset and specialized AWS dataset, respectively. As results form discussions with meteorological domain experts and form summarized fog models properties in [2], especially the second dataset yielded good results. The success can also be credited to goal definition. Instead of general goal in the first case, we have chosen more specific goal tailored to the end user needs in the second case. The model is already used by the business partner of the DMM consortium as a supporting tool during operative fog forecasting.

The low clouds detection model reached 91.8 % accuracy and is a good candidate for practical use.

In our future work we will integrate more data sources and balance positive examples in order to improve the quality of training data. The quantity for positive examples is crucial for the specified task. The most suitable data for mining were high frequency data with high quality stable during long term period. We also want to test our models for other airports and stations in several localities to achieve better description of fog occurrences in both similar and different climatic conditions.

Acknowledgements

The work presented in the paper was supported by the Slovak Research and Development Agency under the contract No. VMSP-P-0048-09, by the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and by Slovak Academy of Sciences under grants No. 1/1147/12 and No. 2/0054/12, and by the project KC-INTELSYS, ITMS, reg. No. 26240220072.

REFERENCES

[1] Bott, A.—Trautmann, T.: PAFOG – A New Efficient Forecast Model of Radiation Fog and Low-Level Stratiform Clouds. Atmos. Research, Vol. 64, 2002, pp. 191–203.

[2] COST 722 – Short Range Forecasting Methods of Fog, Visibility and Low Clouds. Final Report, COST Office, Brussels, Belgium, 2007.

[3] Gultepe, I.—Muller, M. D.—Boybeyi, Z.: A New Visibility Parameterization for Warm Fog Applications in Numerical Weather Prediction Models. In J. Appl. Meteor., Vol. 45, 2006, p. 1469–1480.

[4] Hluchý, L.—Habala, O.—Tran, D. V.—Ciglan, M.: Hydro-Meteorological Scenarios Using Advanced Data Mining and Integration. In The Sixth International Conference on Fuzzy Systems and Knowledge Discovery: proceedings: FSKD 2009. Editor Y. Chen, Hepu Deng, Degan Zhang, Yingyuan Xiao. Los Alamitos: IEEE Computer Society, 2009, ISBN 978-0-7695-3735-1, Vol. 7, pp. 260–264.

[5] Habala, O.—Ciglan, M.—Tran, D. V.—Hluchý, L.: Advanced Data Mining and Integration for Environmental Scenarios. In Information technologies – Applications and theory: Proc. from ITAT, Prague, ISBN 978-80-970179-1-0, 2009, pp. 98–99.

[6] Fayyad, U.—Piatetsky-Shapiro, G.—Smyth, P.: From Data Mining to Knowledge Discovery in Databases. AI Magazine, Vol. 17, 1996, No. 3, p. 37–54.

[7] Atkinson, M.—Brezany, P.—Corcho, O.—Han, L.—van Hemert, J.—Hluchý, L.—Hume, A.—Janciak, I.—Krause, A.—Snelling, D.: ADMIRE White Paper: Motivation, Strategy, Overview and Impact. Technical Report v0.9, The ADMIRE Project, January 2009.

[8] `http://www.microstep-mis.com/index.php?lang=en&site=src/references`, `http://www.microstep-mis.com/index.php?lang=en&site=src/research`.

[9] INCO – COPERNICUS GOAL – Geographic Information On-Line Analysis (GIS – Data Warehouse Integration), P. No. 977091, October 1998 – September 2001.

[10] HPSE-2001-00065 PRODGAP – The Determinants of the Productivity Gap. February – March 2004.

[11] IST-FP6-26476 SEAMLESS – Small Enterprises Accessing the Electronic Market of the Enlarged Europe by a Smart Service Infrastructure. January 2006 – June 2008.

[12] Habala, O.—Paralič, M.—Rozinajová, V.—Bartaloš, P.: Data-Aware Composition of Workflows of Web and Grid Services. In: Bubak, M., Turala, M., Wiatr, K. (eds.): Proceedings of the Cracow Grid Workshop '08. Academic Computer Centre CYFRONET AGH, Poland, 2009, ISBN 978-93-61433-00-2, pp. 120–128.

[13] Zazzaro, G.—Pisano, F. M.—Mercogliano, P.: Data Mining to Classify Fog Events by Applying Cost-Sensitive Classifier. International Conference on Complex, Intelligent and Software Intensive Systems 2010, pp. 1093–1098.

[14] Weymouth, G. T. et al.: Dealing with Uncertainty in Fog Forecasting for Major Airports in Australia. 4th Conference on Fog, Fog Collection and Dew 2007, La Serena, Chile, pp. 73–76.

[15] Ebecken, F. F.: Fog Formation Prediction in Coastal Regions Using Data Mining Techniques. International conference on environmental coastal regions No. 2, Cancun, Mexico, 1998, pp. 165–174.

[16] Acosta, G.—Tosini, M.: A Firmware Digital Neural Network for Climate Prediction Applications. Proceedings of IEEE International Symposium on Intelligent Control 2001, Mexico City, Mexico, ISBN 0-7803-6722-7, 2001.

[17] Koskela, T.—Lehtokangas, M.—Saarinen, J.—Kaski, K.: Time Series Prediction With Multilayer Perceptron, FIR and Elman Neural Networks. Proceedings of the World Congress on Neural Networks 1996, INNS Press, San Diego, USA, pp. 491–496, 1996.

[18] Fabbian, D.—de Dear, R.—Lellyett, S.: Application of Artificial Neural Network Forecasts to Predict Fog at Canberra International Airport. Weather and Forecasting, Vol. 22, 2007, No. 2, pp. 372–381.

[19] Viademonte, S.—Burstein, F.—Dahni, R.—Willians, S.: Discovering Knowledge from Meteorological Databases: A Meteorological Aviation Forecast Study. Third International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2001). Conference proceedings, LNCS 2114, pp. 61–70, Munich, Germany.

[20] Beckenkamp, F.—Pree, W.—Feldens, M. A.: Optimizations of the Combinatorial Neural Model. In Proceedings of 5th Brazilian Symposium on Neural Networks (SBRN '98), Belo Horizonte, Brazil.

[21] Rehm, F.: Prediction of Aircraft Delay at Frankfurt Airport as a Function of Weather. Presentation from German Aerospace Center, Germany, 2004.

[22] Jan, Z.—Abrar, M.—Bashir, S.—Mirza, A. M.: Seasonal to Inter-annual Climate Prediction Using Data Mining KNN Technique. Wireless Networks, Information Processing and Systems, Vol. 20, 2009, pp. 40–51.

[23] Radhika, Y.—Shashi, M.: Atmospheric Temperature Prediction Using SVM. International Journal of Computer Theory and Engineering, Vol. 1, 2009, No. 1, pp. 1793–8201.

[24] `http://www.microstep-mis.com/index.php?lang=en&site=src/products/meteorology/ims_aws/ims_aws`, 19.12.2011 [online].

[25] `http://www.microstep-mis.com/index.php?lang=en&site=src/products/meteorology/ims_model_suite`, 19.12.2011 [online].

[26] International Civil Aviation Organization: Annex 6 to the Convention on International Civil Aviation. Part I: International Commercial Air Transport – Aeroplanes. Eighth Edition, July 2001.

[27] Sobíšek, B. et al.: Meteorologický Slovník Výkladnový a Terminologický. Academia, 1st edition (1993), ISBN 80-85368-45-5.

[28] WMO No. 306 Manual on Codes, International Codes, Secretariat of the World Meteorological Organization, Geneva, Switzerland (2001).

Juraj Bartok received his Master degree in meteorology in 2000 and his Ph. D. degree in meteorological modelling in 2010, both from the Comenius University in Bratislava. He is currently head of meteorological department at MicroStep-MIS, spol. s r. o., a company specializing in the development of monitoring and information systems for aviation, meteorology and environment. His research interests currently are in the areas of numerical weather prediction, weather phenomena simulation/short-time forecasting and processing of sensor data.

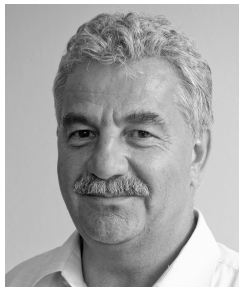
František Babič graduated at the Department of Cybernetics and Artificial Intelligence of the Faculty of Electrical Engineering and Informatics at the Technical University of Košice in 2005. In the same year, he began his Ph. D. study at the same department and successfully finished it with Ph. D. in 2009. He works also as a Researcher in the Centre for Information Technologies, common workplace of Institute of Informatics Slovak Academy of Sciences in Bratislava and Technical university of Košice. He has been participating in several international and national research projects, currently as an assistant professor at the original department. His scientific research is focusing on knowledge management, knowledge discovery, process modelling and project management.

Peter Bednár received his Master degree in 2001 and Ph. D. degree in 2010 at the Technical University of Košice. Since 2005 he is working as a researcher in the Centre of Information Technologies, common workplace of Institute of Informatics, Slovak Academy of Sciences in Bratislava, and Technical University of Košice. His scientific research is focusing on the areas of text mining, knowledge management and application of the semantic technologies in eGovernment and eBusiness.

Ján Paralič received his Master degree in 1992 and his Ph. D. degree in 1998, both from the Technical University of Košice. He is currently a professor at the Department of Cybernetics and Artificial Intelligence, Technical University of Košice and Head of the Centre for Information Technologies at the same university. He (co-)authored four books, (co-)edited 14 proceedings from various international workshops and conferences and published more than 100 scientific papers. His research interests currently are in the areas of knowledge discovery, data and text mining, semantic technologies, and knowledge management. He was editor of a special issue of this Journal devoted to knowledge technologies and related topics in 2007 (Vol. 26, No. 3). He is a member of the ACM, IEEE, Slovak Society for Computer Science and Slovak Artificial Intelligence Society.

Jozef Kováč graduated at the Department of Cybernetics and Articial Intelligence of the Faculty of Electrical Engineering and Informatics at the Technical University of Košice in 2006. He began external Ph. D. study at the same department in 2007 with expected finish in 2012. He has been working as data-mining consultant for last six years (for Adastra and IBM companies) and recently he has founded data-mining startup named 7Segments. His scientic research is focusing on applications of knowledge discovery and business analytics.

Ivana Bartoková received his Master degree in 2000 from Faculty of Mathematics, Physics and Informatics of the Comenius University in Bratislava. She is currently expert for meteorology at MicroStep-MIS, spol. s r. o. Her research interests currently are in the areas of specialized weather forecasting and climate research.



Ladislav Hluchý is the director of the Institute of Informatics of the Slovak Academy of Sciences and also the head of the Department of Parallel and Distributed Computing at the institute. He received MSc and Ph. D. degrees, both in computer science. He is R & D Project Manager, Work-package Leader in a number of 4FP, 5FP, 6FP and 7FP projects, as well as in Slovak R & D projects (VEGA, APVT, SPVV). He is a member of IEEE, ERCIM, SRCIM, and EuroMicro consortiums, the editor-in-chief of the journal Computing and Informatics. He is also (co-)author of scientific books and numerous scientific papers, contributions and invited lectures at international scientific conferences and workshops. He is also an associate professor, supervisor and consultant for Ph. D., master and bachelor studies.

Martin Gera is a lecturer at Comenius University, Faculty of Mathematics, Physics and Informatics, Bratislava (Slovakia). He received his M. Sc. and Ph. D. degrees from the same faculty, in the Department of Meteorology and Climatology. During his professional growth, he spent most of time mainly by investigating of the following scientific topics: numerical modeling of orographic waves, PBL, turbulence parameterization, downscaling of meteorological and climatological data to denser spatial grid, uncertainty analysis of produced greenhouse emissions, forecasting of hazardous phenomena, etc. He has cooperated and participated in the projects and problems connected with the ALADIN research community and in the project funded by APVV and VEGA. Within his postdoctoral visit in Royal Meteorological institute in Belgium he investigated total kinetic energy of turbulent motion. He has cooperated with Slovak and Czech Hydro-Meteorological Institute. He has also cooperated with MicroStep-MIS in modeling the hazardous phenomena. He co-authored numerous scientific papers, and is a member of Slovak Meteorological Society, Slovak National Committee for Geodesy and Geophysics and a member of the Editorial Board of Acta Meteorologica Universitatis Comenianae.