Aerosol and Air Quality Research, 15: 544–555, 2015
Copyright © Taiwan Association for Aerosol Research
ISSN: 1680-8584 print / 2071-1409 online
doi: 10.4209/aaqr.2014.08.0154

Predicting Common Air Quality Index – The Case of Czech Microregions

Petr Hajek[*], **Vladimir Olej**

Institute of System Engineering and Informatics, Faculty of Economics and Administration, University of Pardubice, Studentská 84, 532 10 Pardubice, Czech Republic

ABSTRACT

This paper presents a design of models for common air quality index prediction using computational intelligence methods. In addition, the sets of input variables were optimized for each air pollutant prediction by genetic algorithms. Based on data measured by the three monitoring stations of Dukla, Rosice and Brnenska in the Czech Republic, the models were designed to predict air quality indices for each air pollutant separately and, consequently, to predict the common air quality index. Considering the root mean squared error, the results showed that the compositions of individual prediction models significantly outperform single prediction models of the common air quality index. The feature selection procedure indicates that the determinants of air quality indices were strongly locality specific. Therefore, the models can be applied to obtain more accurate one day ahead predictions of air quality indices. Here we show that the composition models achieve high prediction accuracy for maximum air quality indices (between 50.69 and 63.36%). The goal of the prediction by various methods was to compare the results of the prediction with the aim of various recommendations to micro-regional public administration management.

Keywords: Air quality index; Prediction; Fuzzy inference system; Neural network; Support vector regression.

INTRODUCTION

Air quality indices (AQIs) (Kassomenos *et al.*, 1999; Murena, 2004; Upadhyay *et al.*, 2014) were initially established in response to health issues related to the deteriorating air quality. AQIs are used to report on the state of air pollutants (APs), which are widely accepted as important determinants of adverse health effects. For example, the Aphekom project reported that reducing APs would result in significant health and monetary gains in Europe (Pascal *et al.*, 2013). Therefore, increasing attention has been paid to the prediction of individual APs recently. Recent studies have shown that predicting the APs is a complex problem and thus computational intelligence and machine learning (ML) approaches have provided promising results (Iliadis and Papaleonidas, 2009; Hajek and Olej, 2012). The problem becomes even more complex when attempting to develop models for the prediction of AQIs that combine several APs. Only a few attempts have been made to address this issue (Kyriakidis *et al.*, 2012; Kumar and Goyal, 2013). However these models have only been applied to classification tasks, although much information is lost when transforming the values of AQIs into nominal values.

The goal of this study is to design such models for the prediction of AQIs which would allow modeling complex and non-linear processes within AQIs formation issue. We show that the models are capable of: (1) learning these relations; (2) applying them later on actual data; (3) and finally processing uncertainty related to measuring both APs and meteorological variables.

There are several drawbacks of computational intelligence and ML approaches reported in the literature on APs prediction (Feng *et al.*, 2011). First, the models cannot be employed in other regions since they are developed under specific local chemical and meteorological conditions. Second, meteorological processes are usually simplified in these models. Third, they do not model the interrelationships among multiple pollutants. In our study, we compare three different stations (both urban and suburban) and perform feature selection for each locality to detect the most critical determinants of AQIs under different conditions. We further study the effect of wind direction and velocity in detail to take into account their variability and synergy effects during the day. In this study, the interrelationships among multiple pollutants are reflected in the common AQI, which combines the effects of individual pollutants. Additionally, we perform a comparison of various soft computing and ML methods (Takagi-Sugeno fuzzy inference systems (TSFISs), radial basis function neural networks (RBFNNs), multilayer

[*] Corresponding author.
Tel.: +420 466 036 074; Fax: +420 466 036 010
E-mail address: petr.hajek@upce.cz

perceptron neural networks (MLPs) and support vector regression (SVR)) over a wide range of APs, which has also not been reported in the literature so far.

PREVIOUS LITERATURE ON AIR POLLUTANT PREDICTION

The previous literature on AP prediction can be grouped into three categories (Zhang *et al.*, 2012): (1) simple empirical approaches, (2) physically-based approaches and (3) parametric or non-parametric statistical approaches. The simple empirical approaches either use the values of the present day (persistence method) to predict those of tomorrow or strictly rely on the dependence between meteorological variables and forecasted pollutants. Either way, this approach provides low accurate forecasts. Physically-based approaches, on the other hand, are more accurate mainly due to their capability of modeling temporal and spatial patterns of meteorological variables and APs (Kumar and Goyal, 2014). However, these processes are often too complex to be easily represented by physically-based models, which results in biased forecasts. Parametric or non-parametric statistical approaches such as neural networks (NNs) can outperform physically-based approaches in the accuracy of forecasts (Dutot *et al.*, 2007).

Recent research has shown that NNs predict future AP levels with significantly higher accuracy than linear regression methods (Ozbay *et al.*, 2011). The dominance of NNs over linear regression models has been demonstrated for a variety of APs in many studies. Recently, SVR models performed equally or better than NNs in AP prediction (Lin *et al.*, 2011). This is due to the capability of NNs and SVR to model the complex non-linear relationships between meteorological and air quality variables. To overcome the limitations of the individual approaches to AP prediction, deterministic and statistical methods have been previously combined (Konovalov *et al.*, 2009).

In Table 1, some of the previous literature in this field is summarized (significantly more accurate methods are marked in italics). More specifically, Jiang *et al.* (2004) developed a MLP model for the AQIs (PM_{10}, SO_2, NO_2) prediction in Shanghai, China. It was demonstrated that predicting PM_{10} is more difficult than predicting the remaining APs. This is because PM_{10} consists of multiple chemical components over a broad-sized spectrum. Agirre-Basurko *et al.* (2006) indicated improved performance for the MLP models over the multiple linear regression model in the forecasts of O_3 and NO_2 hourly levels. The same conclusion was drawn by Hrust *et al.* (2009) for four APs (NO_2, O_3, CO and PM_{10}) in the case of an urban residential area.

Recently, SVR models have been developed for the prediction of APs. For example, Osowski and Garanty (2007) demonstrate that SVR model perform significantly better than MLP models on the prediction of various APs (i.e., NO_2, SO_2, CO and dust) in Poland. This is due to the fact that the SVR model is relatively insensitive to the limited number of training data since it minimizes the structural risk unlike the MLP, which minimizes the empirical risk. Thus, the error on testing data is limited with the SVR model. Lin *et al.* (2011) showed the dominance of SVR over NNs on the forecasts of three pollutants, PM_{10}, NO_x and NO_2.

Previous literature has reported that the data on APs are non-linear, heterogeneous and uncertain. Moreover, they are often inconsistent and missing, too. To accommodate the uncertainty in the data, fuzzy logic-based models have been used to forecast APs lately. Domanska and Wojtylak (2012) proposed a model based on fuzzy numbers to predict the concentrations of PM_{10}, $PM_{2.5}$, SO_2, NO, CO and O_3 for a chosen number of hours forward. The superiority of TSFIS over NNs was demonstrated by Hajek and Olej (2012) on O_3 prediction for the city of Pardubice, Czech Republic.

Several attempts have been made to predict common AQIs. Kyriakidis *et al.* (2012) developed several common AQIs to find that MLPs performs better than decision trees and linear regression models in day-ahead predictions. Kumar and Goyal (2013) employed MLP to forecast a common AQI (consisting of three APs: NO_2, SO_2 and dust), suggesting

Table 1. Previous studies on air pollutants' prediction.

Study	Stations	Forecasted pollutants	Methods
Kukkonen *et al.* (2003)	2	NO_2, PM_{10}	*MLP*, deterministic model
Jiang *et al.* (2004)	1	PM_{10}, SO_2, NO_2	MLP
Agirre-Basurko *et al.* (2006)	4	O_3, NO_2	*MLP*, LR
Osowski and Garanty (2007)	7	NO_2, SO_2, CO, dust	SVR + wavelet decomposition
Wang *et al.* (2008)	1	NO_x, SO_2, dust	SVR
Hrust *el al.* (2009)	1	NO_2, O_3, CO, PM_{10}	MLP
Konovalov *et al.* (2009)	3	PM_{10}	LR
Moustris *et al.* (2010)	7	NO_2, CO, SO_2, O_3	MLP
Kumar and Jain (2010)	1	O_3, CO, NO, NO_2	ARIMA
Lin *et al.* (2011)	1	PM_{10}, NO_x, NO_2	GRNN, MLP, SARIMA, *SVR*
Domanska and Wojtylak (2012)	1	PM_{10}, $PM_{2.5}$, SO_2, NO, CO, O_3	fuzzy numbers
Singh *et al.* (2012)	5	NO_2, SO_2, PM_{10}	*GRNN*, LR, MLP, RBFNN
Hajek and Olej (2013)	1	O_3, NO_2, NO_x, SO_2, PM_{10}	TSFIS, *MLP*
Baawain and Al-Serihi (2014)	1	O_3, NO, NO_2, NO_x, SO_2, PM_{10}, CO, H_2S	MLP

Legend: ARIMA is autoregressive integrated moving average model, TSFIS is Takagi-Sugeno fuzzy inference system, GRNN is general regression neural network, LR is linear regression model, and SARIMA is seasonal ARIMA.

that a feature extraction procedure significantly improves accuracy.

MATERIALS AND METHODS

The data for our investigations were represented by the measurements from the three monitoring stations Dukla, Rosice and Brnenska for years 2009–2011. The data were obtained from the Czech Hydro-meteorological Institute and contain the average daily meteorological variables (such as T_{2m} – temperature 2m above terrain, V – wind velocity, θ – wind direction, H – relative air humidity and SR – solar radiation), maximum daily emission variables (NO_2, NO, NO_x, SO_2, PM_{10}, O_3, toluene (TLN) and benzene (BZN)) and other variables (working day (0,1)). The CO substance was not measured in the localities. Basic information about the studied localities is provided as follows. Dukla station (DU) is an urban residential zone, localized at 50°1′26.531 "North latitude, 15°45′48.776″East longitude, altitude = 239 m. Rosice station (RO) is a suburban residential/industrial zone, localized at 50°2′31.913″orth latitude, 15° 44′ 21.891 "East longitude, altitude = 217 m. Brnenska station (HK) is an urban residential/commercial zone, localized at 50°11′43.304″ North latitude, 15°50′46.955″East longitude, altitude = 232 m.

The input variables related to the meteorological variables for the modeling of AQIs in the DU, RO and HK stations were obtained using measurements and calculations as follows. The calculation of the standard deviation σ_θ of the θ was carried out based on the EPA's recommendation (EPA, 1999) using averages from the sequence of n angles θ_i (Yamartino, 1984; Farrugia and Micallef, 2006). The average angle θ_a was determined from the average values of (s_a,c_a), i.e. from $\sin(\theta_i)$ and $\cos(\theta_i)$, as follows

$$s_a = 1/n \sum_{i=1}^{n} \sin(\theta_i),\ c_a = 1/n \sum_{i=1}^{n} \cos(\theta_i),\ \theta_a = \mathrm{arctg}(s_a/c_a). \quad (1)$$

Average wind direction θ_a was determined as average 6 hr values of variables (s_a,c_a) $(s_{a(0\text{-}5)}, s_{a(6\text{-}11)}, s_{a(12\text{-}17)}, s_{a(18\text{-}23)}, c_{a(0\text{-}5)}, c_{a(6\text{-}11)}, c_{a(12\text{-}17)}, c_{a(18\text{-}23)})$, i.e., as 6 hr angle $\theta_a(\theta_{a(0\text{-}5)}, \theta_{a(6\text{-}11)}, \theta_{a(12\text{-}17)}, \theta_{a(18\text{-}23)})$. The corresponding standard deviations were represented by average 6 hr standard deviations $\sigma_\theta(\sigma_{\theta(0\text{-}5)}, \sigma_{\theta(6\text{-}11)}, \sigma_{\theta(12\text{-}17)}, \sigma_{\theta(18\text{-}23)})$. Then the standard deviation σ_θ can be expressed according to Yamartino (1984) and Farrugia and Micallef (2006) as

$$\sigma_\theta = \arcsin(\varepsilon)(1 + (2/\sqrt{3} - 1))\,\varepsilon^3),\ \text{where}\ \varepsilon = \sqrt{1 - (s_a^2 + c_a^2)}\,. \quad (2)$$

Similarly, wind velocity V was determined as follows

$$V_{xa} = -1/n \sum_{i=1}^{n} V_i \sin(\theta_i),\ V_{ya} = -1/n \sum_{i=1}^{n} V_i \cos(\theta_i),$$
$$\theta_v = \mathrm{arctg}(V_{xa}/V_{ya}), \quad (3)$$

where V_i is wind velocity and the +x and +y directions and velocities align with the east and north unit vectors, respectively. Again, a 6hr average was used, that is $\theta_v(\theta_{v(0\text{-}5)}, \theta_{v(6\text{-}11)}, \theta_{v(12\text{-}17)}, \theta_{v(18\text{-}23)})$. The 6 hr averages for wind direction and velocity were selected in order to gain insight into the variability and synergy effects during day. The four daily measurements also match with the corresponding transport sector trajectories (Brook *et al.*, 2002). Additionally, seasonal effect was estimated by sin(day) and cos(day), respectively. A strong seasonal variability of major APs was reported in previous studies (Cheng *et al.*, 2007; Castell-Balaguer *et al.*, 2012; Mandal *et al.*, 2014), reflecting the variability of emissions and atmospheric transport processes.

The basic descriptive statistics of the variables for individual localities DU, RO and HK are presented in Table 2. The monitoring stations are located in urban residential zone (DU), suburban residential/industrial zone (RO) and urban residential/commercial zone (HK). The characteristics of the zones largely determined the air quality. The best air quality was in the DU locality, while the RO locality faced high levels of PM_{10} and SO_2, and the HK locality in particular showed high levels of nitrogen oxides.

About 1.7% of data were missing. As shown in Junninen *et al.* (2004), an improvement in the accuracy of air quality prediction can be achieved using multiple imputations where a final estimate is composed of the outputs of several multivariate fill-in methods. Therefore, the multiple imputation scheme was used to fill in the missing values. More specifically, we employed a model-based multiple imputation (Schafer, 2010) that solves the problem of underestimation of the error variance.

The original set of input variables was optimized using a correlation-based filter (Hall, 1988), which optimizes the set of input variables so that it evaluates the worth of a subset of input variables (features) by considering the individual predictive ability of each feature along with the degree of redundancy between them. The objective function f(λ) can be expressed as

$$f(\lambda) = \frac{\lambda \times \zeta_{cr}}{\sqrt{\lambda + \lambda \times (\lambda - 1) \times \zeta_{rr}}}, \quad (4)$$

where λ is the subset of features, ξ_{cr} is the average feature to output correlation, and ξ_{rr} is the average feature to feature correlation. A genetic algorithm was used as a search method to maximize the objective function f(λ). For further modeling it was necessary to optimize the set of input variables for each AQI.

The AQIs of the Czech National Institute of Public Health were used as target variables in this study. The daily AQIs for individual APs are expressed as $AQIAP^t = AP^t/AP_{24Hmax}$, where AP^t is the maximum value of APs in day t, and AP_{24Hmax} denotes maximum allowed daily concentration of the APs. For the AQIs it holds true that the limit value is met if $AQIAP^t < 1$. A similar approach was used by van den Elshout *et al.* (2008) for the common AQI in Europe. This AQI emphasizes the role of traffic as a source of pollution, as it uses AQIs for city background and traffic situations. In addition, different time scales are used with different calculations. Specifically, hourly and daily AQIs are calculated using a grid with five classes combining

Table 2. Basic descriptive statistics (Mean ± StdDev) of monitored variables.

	DU	RO	HK
O_3 [μg/m^3]	47.233 ± 22.017	45.680 ± 19.616	45.135 ± 21.553
TLN [μg/m^3]	2.427 ± 2.279	4.617 ± 3.803	4.976 ± 4.552
BZN [μg/m^3]	1.461 ± 1.283	2.746 ± 1.714	1.928 ± 2.178
NO_x [μg/m^3]	27.898 ± 19.198	28.299 ± 21.303	49.751 ± 32.494
NO_2 [μg/m^3]	19.971 ± 9.469	18.860 ± 9.656	25.188 ± 11.169
NO [μg/m^3]	5.239 ± 7.466	6.211 ± 8.746	16.054 ± 15.025
SO_2 [μg/m^3]	6.646 ± 5.489	7.797 ± 5.475	4.809 ± 3.742
PM_{10} [μg/m^3]	29.220 ± 21.409	42.195 ± 21.885	29.458 ± 19.982
T_{2m} [K]	282.395 ± 9.568	282.496 ± 8.864	282.283 ± 8.630
SR [W/m^2]	141.697 ± 106.074	142.849 ± 94.956	123.387 ± 84.816
H [%]	73.728 ± 17.062	77.242 ± 18.231	75.084 ± 11.770
θ_a	93.785 ± 39.250	99.523 ± 42.603	86.618 ± 35.889
σ_θ	0.487 ± 0.269	0.471 ± 0.267	0.496 ± 0.248
V	0.899 ± 0.482	2.199 ± 1.420	1.411 ± 1.55
θ_v	–0.213 ± 0.482	0.065 ± 0.406	0.253 ± 0.421

linear regression (O_3) and piecewise linear regression (NO_2 and PM_{10}). In contrast, yearly AQIs are calculated relatively to European limit values, with AQI < 1 (for all APs) denoting satisfaction of the EU annual norms (for details see van den Elshout *et al.*, 2014). Note that this approach is different from traditional AQIs used in US EPA or China, which are purely based on piecewise linear function using index breakpoints. The traditional AQIs use only the maximum value of AQIAPs to calculate the common AQI. Thus, they do not account for the additive impact of different APs. Specifically, the effect of exposure to each AP is independent, and therefore the total daily health impact is expected to be the sum of the values associated with exposure to each AP (Cairncross *et al.*, 2007).

Based on the above considerations, the common AQI (further only AQI^t) is expressed as follows

$$\mathrm{AQI_{MAX}}^t = \max_i(\mathrm{AQIAP}_i^t)\,, \qquad (5)$$

where AQIAP$_i^t$ is the value of AQI^t for the i-th AP, i = 1,2, …, n, and n denotes the number of APs.

The maximum limits AP_{max} of the examined APs were developed by the Czech Hydro-meteorological Institute as follows: SO_2: 125 μg/m^3; NO_2: 200 μg/m^3; PM_{10}: 50 μg/m^3; O_3: 120 μg/m^3. Where the daily limits were not available, the value of AP_{max} for the nearest time interval was used. This evaluation takes the possible influence of human health into account. We note that alternative exposure indices can be constructed in order to consider specific health effects (Bhaskar and Mehta, 2010).

Boxplots in Fig. 1 depict the values of both the individual AQIAPs (SO_2, NO_2, PM_{10}, O_3) and the AQIs in the three localities. The AQIAPs of the RO station were generally higher when compared with those of the other two stations. The DU and HK stations, on the other hand, showed high AQIAPs for NO_2. The AQIs of the RO station were significantly higher than in the other two stations. The distributions of the AQIs in the DU and HK stations were similar.

The AQIs were further transformed into AQI^T from the interval < 0,6) using linear discontinuous functions. Based on the value of the AQI, the state of air pollution can be classified into six classes as shown in Table 3.

Fig. 2 provides more detailed information on the distribution of the AQI classes in the stations. Satisfactory and slightly polluted air was observed in the monitored period for the $\mathrm{AQI_{MAX}}$.

Fig. 3 shows that PM_{10} and O_3 play a determining role in the $\mathrm{AQI_{MAX}}$, while NO_2 and SO_2 together account for less than 1% importance. The impacts were calculated as the percentages of occurrences of APs in the $\mathrm{AQI_{MAX}}$.

To make one day ahead AQI predictions, we used various soft computing and ML methods, each with advantages and limitations. Although NNs can handle nonlinear chemical system at a locality, they have been criticized for poor generalization performance (Zhang *et al.*, 2012). SVR, on the other hand, provides efficient generalization but it may perform poorly on noisy data. Finally, although fuzzy logic systems are capable of processing inherent uncertainties in both the data and human knowledge, they may suffer from computational complexity.

The experiments with the presented methods were carried out in Windows 7, Weka 3.6.8 (correlation-based filter, RBFNNs, MLPs and SVRs), and Matlab 7.1 Fuzzy Logic Toolbox (TSFISs). For NNs and SVR, the structures of the models were optimized using genetic algorithms and grid search algorithm, respectively. This was carried out in DTREG 8.9.

DESIGN OF MODELS FOR PREDICTION

After the optimization of input variables using the correlation based filter, Eq. (4), the individual prediction models were designed (the time horizon was 1 day) (Table 4). The composition $\mathrm{AQI_{MAX}}^{t+1}$ was calculated as a maximum of the APs predictions according to Eq. (5). Table 4 reports that: (1) the variability in wind direction (σ_θ) was an important determinant of all APs but seemingly only for urban poly–storeyed built–up areas (DU and HK); this is

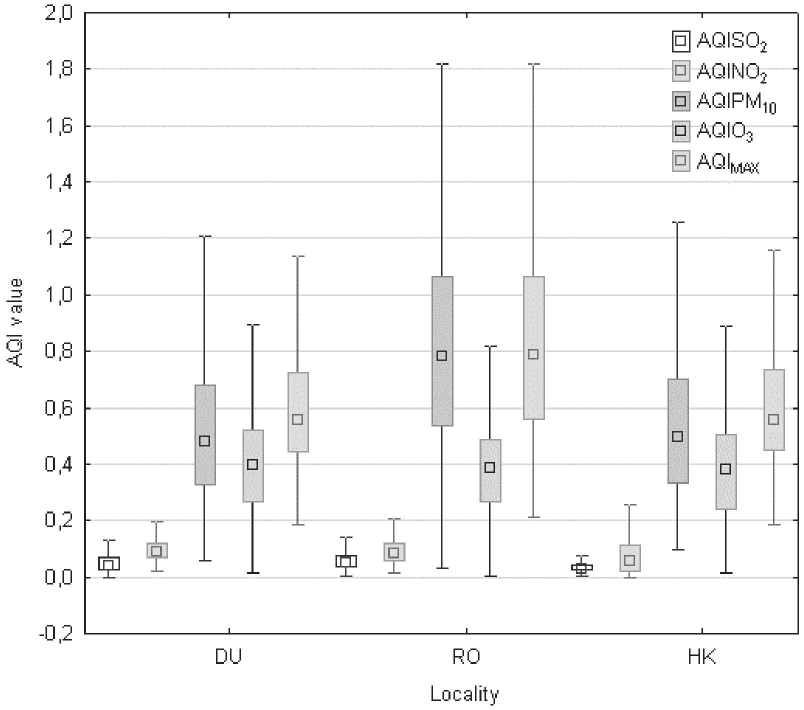

Fig. 1. Boxplots of AQIAPs and AQI.

Table 3. AQI classes of the Czech National Institute of Public Health.

AQI^T	AQI class	Class description
⟨0,1)	1	Clean air, very healthy environment.
⟨1,2)	2	Satisfactory air, healthy environment.
⟨2,3)	3	Slightly polluted air, acceptable environment.
⟨3,4)	4	Polluted air, environment dangerous for sensitive population.
⟨4,5)	5	High polluted air, environment dangerous for the whole population.
⟨5,6)	6	Very high polluted air, harmful environment.

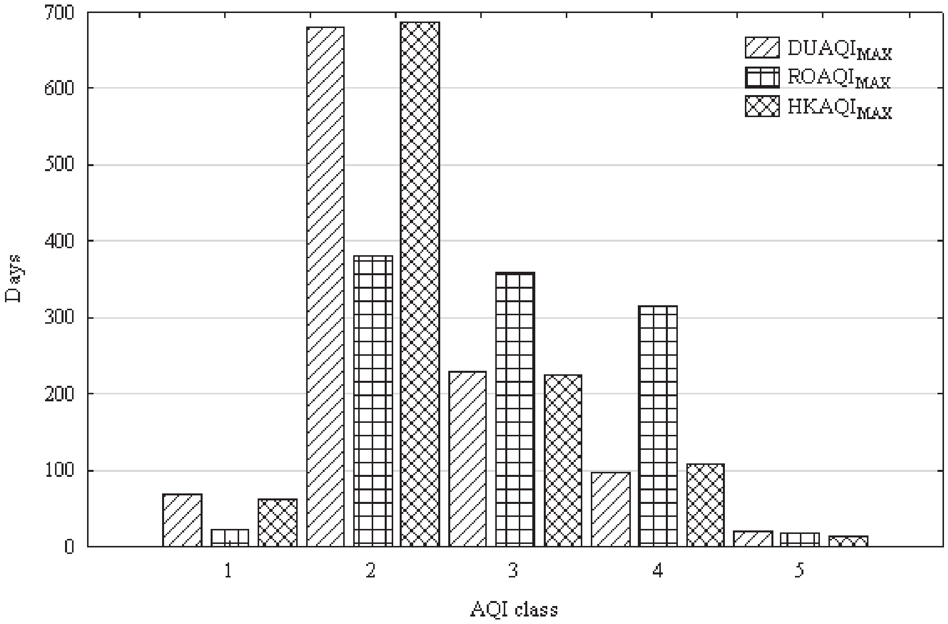

Fig. 2. Histogram of AQI_{MAX} classes.

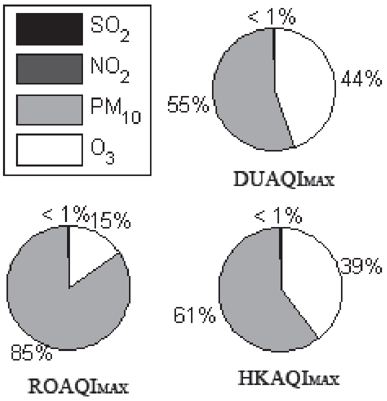

Fig. 3. Percentage of AP impact on AQIs.

because the variability in wind direction was strongly correlated with both wind velocity and wind direction in suburban RO locality and thus considered redundant by the correlation-based filter; (2) wind velocity was a relevant determinant for all APs, irrespective of locality type; (3) a high importance of BZN in the suburban area (RO) shows relation to high traffic volumes (according to the National Traffic Census in 2010, the traffic volume ranged between 15,000 and 25,000 Vehicles/24 hr. in RO, while it was between 10,000 and 15,000 in DU and HK localities); and (4) seasonal effects mattered in almost all prediction models. Generally, the most complex models (with a maximum of determinants) were generated for the industrial, suburban locality of RO located at the edge of Pardubice city (in the direction of the city of Hradec Kralove). This corroborates the high variability of ROAQIs observed in Fig. 1. ROAQIs seem to be dependent on both the traffic volumes and AP sources placed between the two cities.

For further comparison, it was also possible to design additional models for AQI^{t+1} predictions whose sets of input variables were optimized in the same manner as for the models $M_1^1, M_2^1, \ldots, M_4^3$. In further texts, these directly optimized predictions AQI^{t+1} are denoted as optimized AQI^{t+1}. After the feature selection process, the optimized DUAQI^{t+1} (ROAQI^{t+1}, HKAQI^{t+1}) can be defined as follows:

$$\mathrm{DUAQI_{MAX}}^{t+1} = \mathrm{f}(\mathrm{workingday}(0,1)^t, \mathrm{TLN}^t, \mathrm{NO_2}^t, \mathrm{T_{2m}}^t, c_{\mathrm{a}(6\text{-}11)}{}^t, c_{\mathrm{a}(12\text{-}17)}{}^t, s_{\mathrm{a}(6\text{-}11)}{}^t, \mathrm{sinday}^t, \mathrm{PM_{10}}^t, \theta_{\mathrm{a}(0\text{-}5)}{}^t, \sigma_{\theta(0\text{-}5)}{}^t) \tag{6}$$

$$\mathrm{ROAQI_{MAX}}^{t+1} = \mathrm{f}(\mathrm{NO_2}^t, \mathrm{NO_2}^t, \mathrm{SO_2}^t, \mathrm{T_{2m}}^t, c_{\mathrm{a}(12\text{-}17)}{}^t, c_{\mathrm{a}(18\text{-}23)}{}^t, s_{\mathrm{a}(12\text{-}17)}{}^t, s_{\mathrm{a}(18\text{-}23)}{}^t, \mathrm{cosday}^t, \mathrm{PM_{10}}^t, \theta_{\mathrm{a}(12\text{-}17)}{}^t) \tag{7}$$

$$\mathrm{HKAQI_{MAX}}^{t+1} = \mathrm{f}(\mathrm{SO_2}^t, \mathrm{NO}^t, \mathrm{NO_2}^t, \mathrm{T_{2m}}^t, c_{\mathrm{a}(12\text{-}17)}{}^t, c_{\mathrm{a}(18\text{-}23)}{}^t, \theta_{v(18\text{-}23)}{}^t, \mathrm{cosday}^t, \mathrm{PM_{10}}^t, \theta_{\mathrm{a}(0\text{-}5)}{}^t) \tag{8}$$

RESULTS

As presented above, the data for the AQI prediction change over time and are of nonlinear character. They are also heterogeneous, inconsistent, missing and uncertain. This character of data indicates that the models using NNs and methods with uncertainty might perform well in the case of AQI prediction. The TSFISs (Takagi and Sugeno, 1985), RBFNNs (Park and Sandberg, 1991), MLPs (Lippman, 1987)

Table 4. Prediction models.

DU locality	Marking
$\mathrm{DUAQIO_3}^{t+1} = \mathrm{f}(\mathrm{workingday}(0,1)^t, \mathrm{O_3}^t, c_{\mathrm{a}(12\text{-}17)}{}^t, s_{\mathrm{a}(6\text{-}11)}{}^t, \theta_{v(0\text{-}5)}{}^t, \mathrm{sinday}^t, \mathrm{cosday}^t, \theta_{\mathrm{a}(12\text{-}17)}{}^t, \sigma_{\theta(0\text{-}5)}{}^t, \sigma_{\theta(6\text{-}11)}{}^t, \sigma_{\theta(18\text{-}23)}{}^t)$	M_1^1
$\mathrm{DUAQINO_2}^{t+1} = \mathrm{f}(\mathrm{TLN}^t, \mathrm{BZN}^t, \mathrm{NO_2}^t, \mathrm{T_{2m}}^t, c_{\mathrm{a}(0\text{-}5)}{}^t, s_{\mathrm{a}(6\text{-}11)}{}^t, s_{\mathrm{a}(12\text{-}17)}{}^t, \mathrm{cosday}^t, \theta_{\mathrm{a}(0\text{-}5)}{}^t)$	M_2^1
$\mathrm{DUAQISO_2}^{t+1} = \mathrm{f}(\mathrm{workingday}(0,1)^t, \mathrm{BZN}^t, \mathrm{SO_2}^t, s_{\mathrm{a}(18\text{-}23)}{}^t, \mathrm{sinday}^t, \theta_{\mathrm{a}(0\text{-}5)}{}^t, \sigma_{\theta(0\text{-}5)}{}^t)$	M_3^1
$\mathrm{DUAQIPM_{10}}^{t+1} = \mathrm{f}(\mathrm{NO_2}^t, c_{\mathrm{a}(6\text{-}11)}{}^t, c_{\mathrm{a}(12\text{-}17)}{}^t, c_{\mathrm{a}(18\text{-}23)}{}^t, s_{\mathrm{a}(6\text{-}11)}{}^t, \theta_{v(18\text{-}23)}{}^t, \mathrm{cosday}^t, \mathrm{PM_{10}}^t, \theta_{\mathrm{a}(0\text{-}5)}{}^t, \sigma_{\theta(0\text{-}5)}{}^t)$	M_4^1
$\mathrm{DUAQI}^{t+1} = \mathrm{composition}(\mathrm{DUAQIO_3}^{t+1}, \mathrm{DUAQINO_2}^{t+1}, \mathrm{DUAQISO_2}^{t+1}, \mathrm{DUAQIPM_{10}}^{t+1})$	
RO locality	
$\mathrm{ROAQIO_3}^{t+1} = \mathrm{f}(\mathrm{TLN}^t, \mathrm{NO_x}^t, \mathrm{O_3}^t, \mathrm{NO}^t, \mathrm{T_{2m}}^t, \mathrm{SR}^t, s_{\mathrm{a}(18\text{-}23)}{}^t, \theta_{v(12\text{-}17)}{}^t, \mathrm{sinday}^t, \theta_{\mathrm{a}(6\text{-}11)}{}^t, \theta_{\mathrm{a}(18\text{-}23)}{}^t)$	M_1^2
$\mathrm{ROAQINO_2}^{t+1} = \mathrm{f}(\mathrm{workingday}(0,1)^t, \mathrm{NO_2}^t, \mathrm{T_{2m}}^t, c_{\mathrm{a}(6\text{-}11)}{}^t, s_{\mathrm{a}(0\text{-}5)}{}^t, s_{\mathrm{a}(18\text{-}23)}{}^t, \theta_{v(12\text{-}17)}{}^t, \mathrm{sinday}^t, \theta_{\mathrm{a}(6\text{-}11)}{}^t, \theta_{\mathrm{a}(12\text{-}17)}{}^t)$	M_2^2
$\mathrm{ROAQISO_2}^{t+1} = \mathrm{f}(\mathrm{workingday}(0,1)^t, \mathrm{TLN}^t, \mathrm{BZN}^t, \mathrm{SO_2}^t, \mathrm{T_{2m}}^t, c_{\mathrm{a}(6\text{-}11)}{}^t, c_{\mathrm{a}(12\text{-}17)}{}^t, s_{\mathrm{a}(0\text{-}5)}{}^t, s_{\mathrm{a}(18\text{-}23)}{}^t, \mathrm{sinday}^t, \theta_{\mathrm{a}(6\text{-}11)}{}^t)$	M_3^2
$\mathrm{ROAQIPM_{10}}^{t+1} = \mathrm{f}(\mathrm{workingday}(0,1)^t, \mathrm{BZN}^t, \mathrm{NO_2}^t, \mathrm{SO_2}^t, \mathrm{T_{2m}}^t, c_{\mathrm{a}(12\text{-}17)}{}^t, c_{\mathrm{a}(18\text{-}23)})^t, s_{\mathrm{a}(18\text{-}23)}{}^t, \theta_{v(18\text{-}23)}{}^t, \mathrm{PM_{10}}^t, \theta_{\mathrm{a}(6\text{-}11)}{}^t, \theta_{\mathrm{a}(12\text{-}17)}{}^t)$	M_4^2
$\mathrm{ROAQI}^{t+1} = \mathrm{composition}(\mathrm{ROAQIO_3}^{t+1}, \mathrm{ROAQINO_2}^{t+1}, \mathrm{ROAQISO_2}^{t+1}, \mathrm{ROAQIPM_{10}}^{t+1})$	
HK locality	
$\mathrm{HKAQIO_3}^{t+1} = \mathrm{f}(\mathrm{O_3}^t, \mathrm{T_{2m}}^t, \mathrm{H}^t, s_{\mathrm{a}(12\text{-}17)}{}^t, \theta_{v(6\text{-}11)}{}^{\mathrm{t}}, \mathrm{sinday}^t, \mathrm{cosday}^t, \theta_{\mathrm{a}(6\text{-}11)}{}^t, \theta_{\mathrm{a}(18\text{-}23)}{}^{\mathrm{t}}, \sigma_{\theta(0\text{-}5)}{}^t, \sigma_{\theta(6\text{-}11)}{}^t)$	M_1^3
$\mathrm{HKAQINO_2}^{t+1} = \mathrm{f}(\mathrm{SO_2}^t, \mathrm{NO}^t, \mathrm{NO_x}^t, \mathrm{O_3}^t, \mathrm{T_{2m}}^t, c_{\mathrm{a}(12\text{-}17)}{}^t, \theta_{v(18\text{-}23)}{}^t, \mathrm{sinday}^t, \mathrm{cosday}^t, \theta_{\mathrm{a}(0\text{-}5)}{}^t, \theta_{\mathrm{a}(18\text{-}23)}{}^t)$	M_2^3
$\mathrm{HKAQISO_2}^{t+1} = \mathrm{f}(\mathrm{workingday}(0,1)^t, \mathrm{SO_2}^t, \mathrm{T_{2m}}^t, c_{\mathrm{a}(6\text{-}11)}{}^t, c_{\mathrm{a}(18\text{-}23)})^t, s_{\mathrm{a}(0\text{-}5)}{}^{\mathrm{t}}, s_{\mathrm{a}(6\text{-}11)}{}^t, s_{\mathrm{a}(12\text{-}17)}{}^t, s_{\mathrm{a}(18\text{-}23)}{}^t, \mathrm{cosday}^t, \mathrm{PM_{10}}^t, \theta_{\mathrm{a}(6\text{-}11)}{}^t, \theta_{\mathrm{a}(12\text{-}17)}{}^t, \sigma_{\theta(0\text{-}5)}{}^t, \sigma_{\theta(18\text{-}23)}{}^t)$	M_3^3
$\mathrm{HKAQIPM_{10}}^{t+1} = \mathrm{f}(\mathrm{workingday}(0,1)^t, \mathrm{SO_2}^t, \mathrm{NO_2}^t, \mathrm{T_{2m}}^t, c_{\mathrm{a}(18\text{-}23)}{}^t, s_{\mathrm{a}(12\text{-}17)}{}^t, \theta_{v(0\text{-}5)}{}^t, \theta_{v(12\text{-}17)}{}^t, \mathrm{sinday}^t, \mathrm{PM_{10}}^t, \theta_{\mathrm{a}(0\text{-}5)}{}^t, \sigma_{\theta(0\text{-}5)}{}^t, \sigma_{\theta(18\text{-}23)}{}^t)$	M_4^3
$\mathrm{HKAQI}^{t+1} = \mathrm{composition}(\mathrm{HKAQIO_3}^{t+1}, \mathrm{HKAQINO_2}^{t+1}, \mathrm{HKAQISO_2}^{t+1}, \mathrm{HKAQIPM_{10}}^{t+1})$	

and SVR (Vapnik, 1995) were employed to predict the target AQI^{t+1}. The most usual procedure in air quality prediction was followed to determine the relation between training and testing set (O_{train}:O_{test}). Therefore, the set of training data O_{train} (used in the process of learning) contained the data from the years 2009 and 2010 (730 data points) and the set of testing data O_{test} covered the year 2011 (365 data points). To avoid overlearning, the 10-fold cross-validation was used to find the optimum values of the methods' parameters.

TSFISs used in this study are first-order fuzzy rule-based systems with fuzzy sets in the antecedents and linear regression functions in the consequents of the if-then rules. The identification of the TSFISs was carried out using subtractive clustering algorithm (Chiu, 1994). The number of clusters was equal to the number of if-then rules, and it was determined by the radius of influence of a cluster. In order to avoid over-fitting, we tested different values of the radius $r = \{0.1,0.2, \ldots ,0.9\}$. The numbers of resulting if-then rules ranged from 8 to 12. The parameters of the linear regression functions were optimized using the ANFIS method, namely with the hybrid algorithm combining back-propagation with least square method (Jang, 1993).

RBFNN refers to any kind of feed-forward NNs that uses RBF (usually Gaussian) as an activation function. The structures and parameters of the RBFNNs were found using a genetic algorithm for the following values: (1) the maximum number of neurons in the hidden layer was set to 100; and (2) the radius of RBF ranged from 0.01 to 400. The initial centres for the Gaussian RBFs were found using the *k*-means algorithm.

MLP used in this study is a feed-forward NN with one hidden layer consisting of neurons with sigmoid activation functions. MLPs were trained using the back-propagation algorithm, where the structures and parameters were found using grid search for the following values: (1) the number of neurons in the hidden layer ranged from 1 to 20 with the step set to 1 and maximum steps without change set to 4; (2) the number of cycles $n_c = \{50, 100, \ldots, 2000\}$; (3) and learning rate $\eta = \{0.001, 0.01, 0.05, 0.1, 0.2, 0.3\}$.

SVR employed here uses ε-insensitive loss function in the regularized risk functional that ensures maximum generalization performance. The SVR was trained by the sequential minimal optimization (SMO_{reg}) (Shevade *et al.*, 2000) using second degree polynomial kernel with complexity parameter $C = \{1, 2, 4, \ldots, 256\}$. Again, grid search was used to find the optimum complexity parameter. The first and third degree kernel and RBF kernel function (with radius ranging from 0.01 to 20) were also tested without improvement.

Fig. 4 provides the $RMSE_{test}$ of predictions for partial models, composition models and for the single optimized models. It is obvious that the lowest value of the $RMSE_{test}$ for the AQI_{MAX} was achieved by the SVR trained on the composition model. The highest error for the composition model, on the other hand, was obtained using the RBFNN. The $RMSE_{test}$ on the directly optimized model was also the lowest one for the SVR. This suggests that in the process of learning and testing, the SVR is able to cope with the lower amount of data when compared with the methods with uncertainty and NNs, respectively. Surprisingly, the MLP performed best on the ozone AQI prediction for the DU locality while the TSFIS performed best for remaining localities. The results for the HK locality are slightly different from those for the DU and RO localities. First, the SVR prediction models outperformed other models in all cases, except for $HKAQIO_3^{t+1}$. Second, the error for NO_2 was, for all methods, almost two times as high as in the case of DU and RO, respectively. For the RO locality, the high error on PM_{10} seems to generate a high $RMSE_{test}$ on the AQIs. This

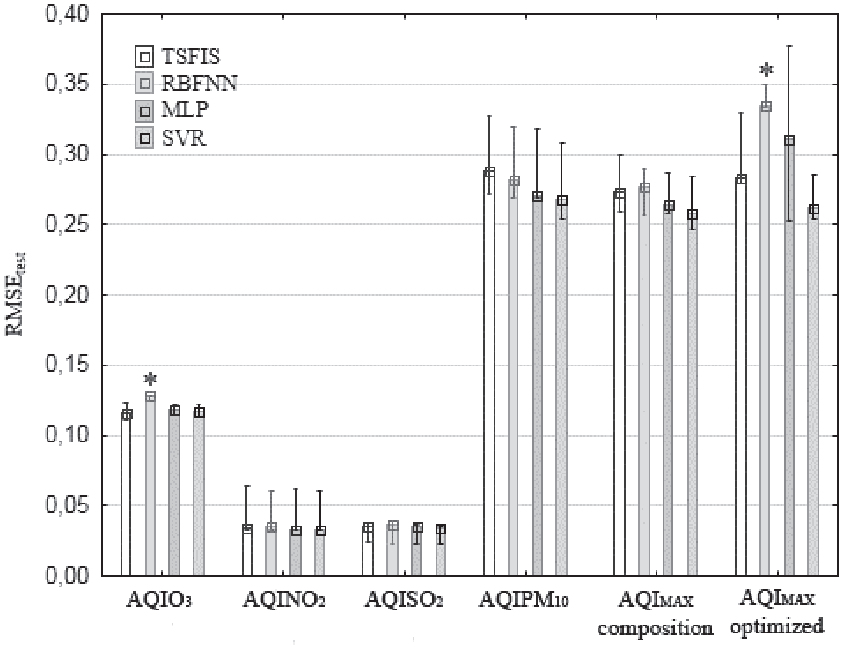

Fig. 4. $RMSE_{test}$ for AQIs (Legend: * denotes significantly lower $RMSE_{test}$ at $P < 0.10$ (paired *t*-test)).

occurred despite the low errors on ROAQISO$_2^{t+1}$, ROAQIO$_3^{t+1}$ and ROAQINO$_2^{t+1}$. To sum up, these results suggest that the prediction of O_3 and NO_2 pollutants is less complex than the remaining APs because the SVR models performed worse. However, SVR models performed best for both composition and optimized AQIs.

To compare the errors, we performed paired *t*-tests, both for individual APs and for common AQIs. While TSFISs, MLP and SVR provided significantly lower error for O_3 (at $P < 0.05$), all methods performed statistically similar for NO_2, SO_2 and PM_{10}. For the composition models of AQIs, all methods performed statistically similar, while RBFNN models performed significantly worse for the optimized models of AQIs. More importantly, as shown in Fig. 5, the best SVR models successfully dealt with the extreme values of AQIs.

Previous results have shown that the prediction models were able to handle the lowest (highest) levels of AQIs. However, most of the extreme values were exaggerated in Fig. 5. Therefore, we tested if the models correctly classify the patterns into AQI classes. We calculated the classification accuracy (percentage of correctly classified patterns) on testing data to assess the classification performance of the prediction models. As shown in Table 5, in the testing data O_{test} only the patterns from the AQI classes 1–5 were present for AQI$_{MAX}$. Again, we compared the performance of the composition and optimized models. The classification performance was measured in terms of classification accuracy, type I error (false positive rate) and type II error (false negative rate).

The total classification accuracy ranged between 50.69% (RO) and 63.36% (HK) (Table 5). The composition ANFIS model performed best. The accuracies were lower for the extreme AQI classes. However, the best model was able to classify the class 4 with a high classification accuracy of 55.10% (composition ANFIS for the HK locality). In fact, the capability of predicting the AQI level, which is dangerous to the population, is one of the most desired features of an AQI prediction model.

As can be seen from Fig. 2, the vast majority of DUAQI, ROAQI and HKAQI measurements can be classified into four AQI classes. Further, the analysis of the classification results in Table 5 shows that the incorrectly classified measurements were mostly classified into a neighboring class, i.e., from 1 → 2, from (2 → 1 or 2 → 3), from (3 → 2 or 3 → 4), Only in rare cases was the classification incorrect in terms of two classes away.

CONCLUSION

We designed two models for the prediction of AQI^{t+1}. The first (composition) model was based on (1) partial AQIs for the localities DU, RO and HK; and (2) on the consequent composition (maximum). To achieve low prediction errors it was necessary to design more complex hierarchical models combining predictions of individual

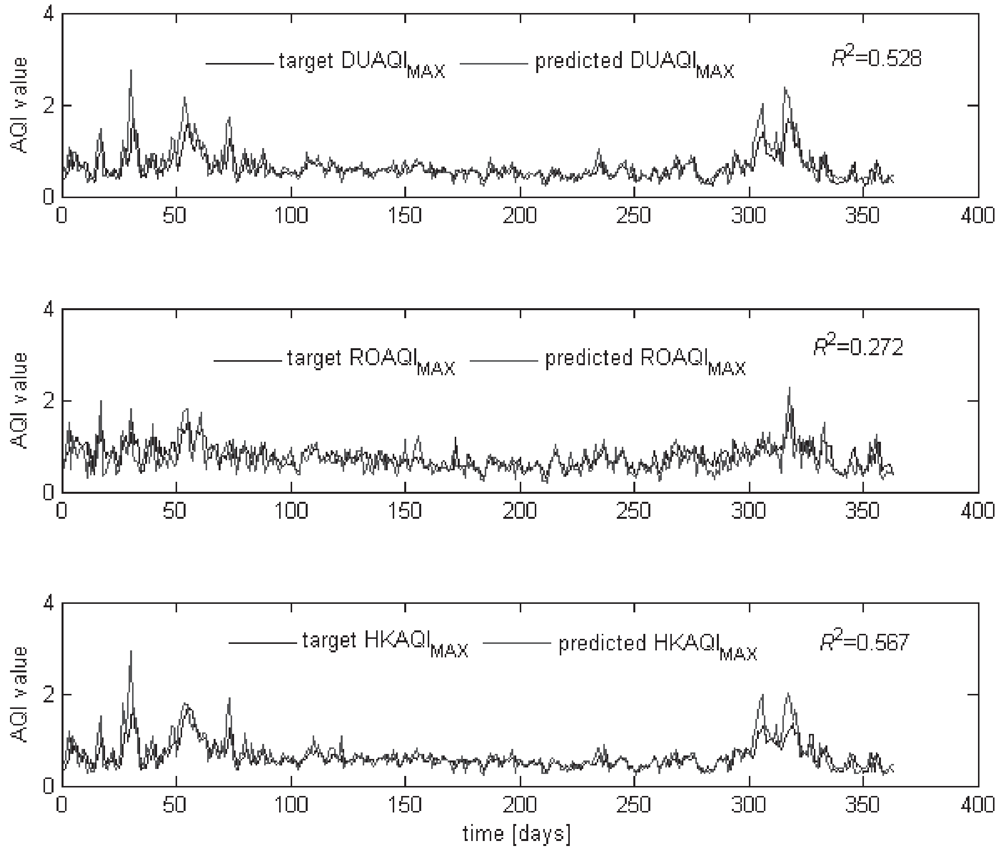

Fig. 5. Prediction of AQI$_{MAX}$ by SVR composition model.

Table 5. Classification accuracy [%] on testing data for AQI_{MAX}.

	DU locality							
AQI class	composition				optimized			
	ANFIS	RBFNN	MLP	SVR	ANFIS	RBFNN	MLP	SVR
1	8.00	12.00	12.00	**28.00**	12.00	0.00	0.00	0.00
2	82.86	80.48	**83.33**	79.05	73.81	66.51	78.95	81.82
3	37.04	35.80	29.63	37.04	46.91	**69.14**	49.38	38.27
4	46.34	41.46	39.02	**48.78**	36.59	0.00	46.34	41.46
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Total	**61.98**	60.06	60.06	61.43	58.13	53.72	61.71	60.33
Total (1class away)	97.52	97.52	97.52	97.52	96.69	94.77	97.52	97.80
Type I error	21.49	24.24	24.52	24.52	19.83	19.83	**17.08**	21.49
Type II error	16.53	15.70	15.43	14.05	**13.77**	**13.77**	21.21	18.18
	RO locality							
AQI class	composition				optimized			
	ANFIS	RBFNN	MLP	SVR	ANFIS	RBFNN	MLP	SVR
1	6.67	0.00	0.00	**13.33**	6.67	0.00	0.00	0.00
2	43.10	43.10	45.40	**56.90**	30.46	0.00	0.00	49.71
3	54.78	**63.48**	61.74	56.52	47.83	59.48	57.76	56.90
4	43.10	36.21	37.93	31.03	46.55	60.34	**62.07**	29.31
5	100.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
Total	45.45	46.56	47.66	**50.69**	37.47	28.65	28.37	46.56
Total (1class away)	90.36	92.84	93.11	92.56	90.36	83.75	84.30	**94.21**
Type I error	14.60	16.25	15.70	19.56	17.08	6.61	**6.34**	20.66
Type II error	39.97	37.19	36.64	**29.75**	45.45	64.74	65.29	32.78
	HK locality							
AQI class	composition				optimized			
	ANFIS	RBFNN	MLP	SVR	ANFIS	RBFNN	MLP	SVR
1	20.00	3.33	16.67	23.33	13.33	0.00	**53.33**	6.67
2	83.33	85.71	**86.19**	80.48	71.43	71.90	77.62	84.76
3	30.56	37.50	27.78	31.94	**47.22**	45.83	13.89	41.67
4	**55.10**	42.86	40.82	48.98	46.94	0.00	24.49	38.78
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Total	**63.36**	63.09	62.26	61.43	58.13	50.69	55.37	63.09
Total (1class away)	96.14	96.69	96.97	97.25	95.04	94.21	95.59	**97.80**
Type I error	21.76	20.94	22.87	23.42	**18.18**	24.79	37.47	19.83
Type II error	14.88	15.98	14.88	15.15	23.69	24.52	**7.16**	17.08
Avg (methods)*	*56.93*	*56.57*	*56.66*	*57.85*	*51.24*	44.35	48.48	*56.66*
Avg rank	2.33	3.83	3.83	3.00	6.33	7.67	5.67	3.33
Avg (comp. vs. opt.)*	*57.00*				50.19			

* significantly higher accuracy at $P < 0.10$ (Friedman test) is marked in italics.

APAQIs. When compared with previous literature, this is the first attempt to employ hierarchical regression models to predict AQIs. The models are more complex since the sets of input variables were optimized separately for each subsystem. Following this procedure, TSFISs, RBFNNs, MLPs and SVRs performed better than the single (optimized) prediction models, where the set of input variables was optimized only once from the original set of all input variables. The results show that the predicted composition $DUAQI^{t+1}$, $ROAQI^{t+1}$ and $HKAQI^{t+1}$ well approximate the target $AQIs^{t+1}$.

The procedure of input variables' optimization using the correlation-based filter and genetic algorithm has provided important determinants of individual APAQIs. The determinants were strongly locality specific. Thus, these subsets of input variables add substantially to our understanding of APAQIs formation. In addition, we performed the optimization procedure for different stations to detect the most critical determinants of APAQIs under different conditions. The study has also gone some way towards enhancing our understanding of the role of wind direction and velocity in APAQIs prediction. Both their variability and synergy effect during day were studied for the three localities. The results of this study indicate that the variability (standard deviation) of wind direction was critical for both urban residential and suburban residential/industrial localities although this effect may be suppressed in feature selection owing to the strong correlations with wind velocity and wind direction, respectively. Furthermore, wind velocity was a relevant determinant for all APs, irrespective of locality type. It is, however, important to note that urban stations may encounter irregular wind flows because of more complex construction settings when compared with rural stations. More specifically, the terrain roughness and atmospheric mixing following a daily cycle driven by solar heating have been found to be the main determinants affecting wind flows (Katinas *et al.*, 2013). Therefore, small differences in local settings (topography, building geometry and dimensions, streets, traffic, trees, etc.) can largely affect air flows (Georgakis and Santamouris, 2006).

The high importance of volatile organic compounds (BZN) in the prediction models suggests that traffic volume represents a critical determinant in the prediction of APAQIs in the three localities. This is in line with the results of the National Traffic Census in 2010, indicating that the traffic volume ranged between 10,000 and 25,000 (Vehicles/24 hr.) in the localities.

In the common AQI we reflected the effects of multiple APs. Finally, we performed a comparison of several soft computing and machine learning methods over the AQIs. To achieve low prediction errors it was necessary to design more complex hierarchical models combining predictions of individual APs. As expected, we have shown that the soft computing methods can be effectively used to accurately process the given non-linear, inconsistent, missing and uncertain data. This result corroborates the findings of a great deal of the previous work in this field as the dependencies of urban APs concentrations on both the relevant APs and meteorological variables are taken as strongly non-linear. SVR performed particularly well due to its remarkable generalization performance. However, the remaining methods showed promising results in several cases, for example TSFIS for O_3 prediction, which seems to be related to a greater uncertainty in O_3 data. Another promising result has been shown in the case of the classification into the AQI classes, where the incorrectly classified instances were placed into neighboring classes in the worst cases. In this case, SVR and ANFIS models performed best.

Finally, a number of important limitations of this study need to be considered. The main limitation of the chosen statistical approach is that the application of the models is completely local due to specific chemical and meteorological conditions and thus the models cannot be applied in other stations, even in the same city. The limitation of the composition approach is that it takes into account the errors for individual APs but not the final error between the target and predicted value of AQI^{t+1}. Therefore, future research should be done using multiple output SVRs. Further, we replaced the missing measurements with the multiple imputation approach. In the case of a high proportion of missing data, this procedure may result in biased results. Therefore, a future study employing ε-insensitive SVR with semi-supervised learning would be interesting because this method can be effectively used due to its possibility to make use of unlabelled data (with missing output values). Finally, the classification performance was strongly affected by the balance of the class sizes. Therefore, in cases of imbalanced data we suggest using classifiers optimized to address this issue.

ACKNOWLEDGMENTS

We gratefully acknowledge the help provided by constructive comments of the anonymous referees. This work was supported by the scientific research project of the Technology Agency of the Czech Republic under Grant No: TD-010130 and we thank the Czech Hydro-meteorological Institute for providing the data.

APPENDIX – LIST OF ABBREVIATIONS

ANFIS	adaptive neuro-fuzzy inference system
AP	air pollutant
AP_{24Hmax}	maximum allowed daily concentration of AP
AQI	air quality index
AQIAP	the daily AQI for AP
AQI_{MAX}	maximum AQI
DU	Dukla station
DUAQI	AQI in DU station
EPA	Environmental Protection Agency
HK	Brnenska station
HKAQI	AQI in HK station
ML	machine learning
MLP	multilayer perceptron
NN	neural network
RBF	radial basis function
RMSE	root mean squared error
RO	Rosice station
ROAQI	AQI in RO station
SMO	sequential minimal optimization
SVR	support vector regression
TSFIS	Takagi-Sugeno fuzzy inference system

REFERENCES

Agirre-Basurko, E., Ibarra-Berastegi, G. and Madariaga, I. (2006). Regression and Multilayer Perceptron-based Models to Forecast Hourly O_3 and NO_2 Levels in the Bilbao Area. *Environ. Modell. Softw.* 21: 430–446.

Baawain, M.S. and Al-Serihi, A.S. (2014). Systematic Approach for the Prediction of Ground-Level Air Pollution (around an Industrial Port) Using an Artificial Neural Network. *Aerosol Air Qual. Res.* 14: 124–134.

Bhaskar, B.V. and Mehta, V.M. (2010). Atmospheric Particulate Pollutants and Their Relationship with Meteorology in Ahmedabad. *Aerosol Air Qual. Res.* 10: 301–315.

Brook, J.R., Lillyman, C.D., Shepherd, M.F. and Mamedov, A. (2002). Regional Transport and Urban Contributions to Fine Particle Concentrations in Southeastern Canada. *J. Air Waste Manage. Assoc.* 52: 855–866.

Cairncross, E.K., John, J. and Zunckel, M. (2007). A Novel Air Pollution Index based on the Relative Risk of Daily Mortality Associated with Short-Term Exposure to Common Air Pollutants. *Atmos. Environ.* 41: 8442–8454.

Castell-Balaguer, N., Téllez, L. and Mantilla, E. (2012). Daily, Seasonal and Monthly Variations in Ozone Levels Recorded at the Turia River Basin in Valencia (Eastern Spain). *Environ. Sci. Pollut. Res. Int.* 19: 3461–3480.

Cheng, C.S., Campbell, M., Li, Q., Li, G., Auld, H., Day, N., Pengelly, D., Gingrich, S. and Yap, D. (2007). A Synoptic Climatological Approach to Assess Climatic Impact on Air Quality in South-Central Canada. Part I: Historical Analysis. *Water Air Soil Pollut.* 182: 131–148.

Chiu, S. (1994). Fuzzy Model Identification Based on Cluster Estimation. *J. Intell. Fuzzy Syst.* 2: 267–278.

Domańska, D. and Wojtylak, M. (2012). Application of Fuzzy Time Series Models for Forecasting Pollution Concentrations. *Expert Syst. Appl.* 39: 7673–7679.

Dutot, A.L., Rynkiewicz, J., Steiner, F.E. and Rude, J. (2007). A 24-h Forecast of Ozone Peaks and exceedance Levels Using Neural Classifiers and Weather Predictions. *Environ. Modell. Softw.* 22: 1261–1269.

EPA (1999). Guideline for Reporting of Daily Air Quality – Air Quality Index, US Environmental Protection Agency.

Farrugia, P.S. and Micallef, A. (2006). Comparative Analysis of Estimators for Wind Direction Standard Deviation. *Meteorol. Appl.* 13: 29–41.

Feng, Y., Zhang, W., Sun, D. and Zhang, L. (2011). Ozone Concentration Forecast Method Based on Genetic Algorithm Optimized Back Propagation Neural Networks and Support Vector Machine Data Classification. *Atmos. Environ.* 45: 1979–1985.

Georgakis, C. and Santamouris, M. (2006). Experimental Investigation of Air Flow and Temperature Distribution in Deep Urban Canyons for Natural Ventilation Purposes. *Energ. Buildings* 38: 367–376.

Hajek, P. and Olej V. (2012). Ozone Prediction on the Basis of Neural Networks, Support Vector Regression and Methods with Uncertainty. *Ecol. Inf.* 12: 31–42.

Hajek, P. and Olej V. (2013). Prediction of Air Quality Indices by Neural Networks and Fuzzy Inference Systems – The Case of Pardubice Microregion. In *14th Engineering Applications of Neural Networks*, Iliadis, L., Papadopoulos, H. and Jayne, Ch. (Eds.), p. 302–312.

Hall, M.A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*, University of Waikato, Hamilton.

Hrust, L., Klaić, Z.B., Križan, J., Antonić, O. and Hercog, P. (2009). Neural Network Forecasting of Air Pollutants Hourly Concentrations Using Optimised Temporal Averages of Meteorological Variables and Pollutant Concentrations. *Atmos. Environ.* 43: 5588–5596.

Iliadis, L.S. and Papaleonidas, A. (2009). Intelligent Agents Networks Employing Hybrid Reasoning: Application in Air Quality Monitoring and Improvement. In *Communications in Computer and Information Science*, Vol. 43, Palmer-Brown, D., Draganova, Ch., Pimenidis, E. and Mouratidis, H. (Eds.), Springer, Berlin, p. 1–16.

Jang, J.S. (1993). ANFIS: Adaptive-Network-based Fuzzy Inference System. *IIEEE Trans. Syst. Man Cybern.* 23: 665–685.

Jiang, D., Zhang, Y., Hu, X., Zeng, Y., Tan, J. and Shao, D. (2004). Progress in Developing an ANN Model for Air Pollution Index Forecast. *Atmos. Environ.* 38: 7055–7064.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. and Kolehmainen, M. (2004). Methods for Imputation of Missing Values in Air Quality Data Sets. *Atmos. Environ.* 38: 2895–2907.

Kassomenos, P., Skouloudis, A.N., Lykoudis, S. and Flocas, H.A. (1999). Air Quality Indicators for Uniform Indexing of Atmospheric Pollution over Large Metropolitan Areas. *Atmos. Environ.* 33: 1861–1879.

Katinas, V., Sankauskas, D., Perednis, E. and Vaitiekūnas, P. (2013). Investigation of the Wind Characteristics and Prospects of Wind Power Use in Lithuania. *J. Environ. Eng. Landsc. Manage.* 21: 209–215.

Konovalov, I.B., Beekmann, M., Meleux, F., Dutot, A. and Foret, G. (2009). Combining Deterministic and Statistical Approaches for PM_{10} Forecasting in Europe. *Atmos. Environ.* 43: 6425–6434.

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R. and Cawley, G. (2003). Extensive Evaluation of Neural Network Models for the Prediction of NO_2 and PM_{10} Concentrations, Compared with a Deterministic Modelling System and Measurements in Central Helsinki. *Atmos. Environ.* 37: 4539–4550.

Kumar, A. and Goyal, P. (2013). Forecasting of Air Quality Index in Delhi Using Neural Network based on Principal Component Analysis. *Pure Appl. Geophys* 170: 711–722.

Kumar, A. and Goyal, P. (2014). Air Quality Prediction of PM_{10} through an Analytical Dispersion Model for Delhi. *Aerosol Air Qual. Res.* 14: 1487–1499.

Kumar, U. and Jain, V.K. (2010). ARIMA Forecasting of Ambient Air Pollutants (O_3, NO, NO_2 and CO). *Stoch. Environ. Res. Risk Assess.* 24: 751–760.

Kyriakidis, I., Karatzas, K., Papadourakis, G., Ware, A. and Kukkonen, J. (2012). Investigation and Forecasting of the Common Air Quality Index in Thessaloniki, Greece. In *IFIP Advances in Information and Communication Technology*, Vol. 382, Iliadis, L., Maglogiannis, I., Papadopoulos, H., Karatzas, K. and Sioutas, S. (Eds.), Springer, Berlin, p. 390–400.

Lin, K.P., Pai, P.F. and Yang, S.L. (2011). Forecasting Concentrations of Air Pollutants by Logarithm Support Vector Regression with Immune Algorithms. *Appl. Math. Comput.* 217: 5318–5327.

Lippman, R.P. (1997). An Introduction to Computing with Neural Nets. *IEEE ASSP Mag.* 4: 4–22.

Mandal, P., Saud, T., Sarkar, R., Mandal, A., Sharma, S.K., Mandal, T.K. and Bassin, J.K. (2014). High Seasonal Variation of Atmospheric C and Particle Concentrations in Delhi, India. *Environ. Chem. Lett.* 12: 225–230.

Moustris, K.P., Ziomas, I.C. and Paliatsos, A.G. (2010). 3-Day-ahead Forecasting of Regional Pollution Index for the Pollutants NO_2, CO, SO_2, and O_3 Using Artificial Neural Networks in Athens, Greece. *Water Air Soil Pollut.* 209: 29–43.

Murena, F. (2004). Measuring Air Quality over Large Urban Areas: Development and Application of an Air Pollution Index at the Urban Area of Naples. *Atmos. Environ.* 38: 6195–6202.

Osowski, S. and Garanty, K. (2007). Forecasting of the Daily Meteorological Pollution using Wavelets and Support Vector Machine. *Eng. Appl. Artif. Intell.* 20: 745–755.

Ozbay, B., Keskin, G.A., Dogruparmak, S.C. and Ayberk,

S. (2011). Predicting Tropospheric Ozone Concentrations in Different Temporal Scales by using Multilayer Perceptron Models. *Ecol. Inf.* 6: 242–247.

Park, J. and Sandberg, I.W. (1991). Universal Approximation Using Radial-Basis-Function Networks. *Neural Comput.* 3: 246–257.

Pascal, M., Corso, M., Chanel, O., Dexlercq, C., Badaloni, C., Cesaroni, G., Henschel, S., Meister, K., Haluza, D., Martin-Olmedo, P. and Medina, S. (2013). Assessing the Public Health Impacts of Urban Air Pollution in 25 European Cities: Results of the Aphekom Project. *Sci. Total Environ.* 449: 390–400.

Schafer, J.L. (2010). *Analysis of Incomplete Multivariate Data*, CRC Press, London.

Shevade, S.K., Keerthi, S.S., Bhattacharyya, C. and Murthy, K.R.K. (2000). Improvements to the SMO Algorithm for SVM Regression. *IEEE Trans. Neural Networks* 11: 1188–1193.

Singh, K.P., Gupta, S., Kumar, A. and Shukla, S.P. (2012). Linear and Nonlinear Modeling Approaches for Urban Air Quality Prediction. *Sci. Total Environ.* 426: 244–255.

Takagi, T. and Sugeno, M. (1985). Fuzzy Identification of Systems and Its Applications to Modeling and Control. *IIEEE Trans. Syst. Man Cybern.* 15: 116–132.

Upadhyay, A., Kanchan, Goyal, P., Yerramilli, A. and Gorai, A.K. (2014). Development of a Fuzzy Pattern Recognition Model for Air Quality Assessment of Howrah City. *Aerosol Air Qual. Res.* 14: 1639–1652, doi: 10.4209/aaqr.2013.04.0118.

van den Elshout, S., Léger, K. and Nussio, F. (2008). Comparing Urban Air Quality in Europe in Real Time: A Review of Existing Air Quality Indices and the Proposal of a Common Alternative. *Environ. Int.* 34: 720–726.

van den Elshout, S., Léger, K. and Heich, H. (2014). CAQI Common Air Quality Index - Update with $PM_{2.5}$ and Sensitivity Analysis. *Sci. Total Environ.* 488: 461–468.

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*, Springer, New York.

Wang, W., Men, Ch. and Lu, W. (2008). Online Prediction Model Based on Support Vector Machine. *Neurocomp.* 71: 550–558.

Yamartino, R.J. (1984). A Comparison of Several Single Pass Estimators of the Standard Deviation of Wind Direction. *J. Climate Appl. Meteor.* 23: 1362–1366.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C. and Baklanov, A. (2012). Real-Time Air Quality Forecasting, Part I: History, Techniques, and Current Status. *Atmos. Environ.* 60: 632–655.

Received for review, August 1, 2014
Revised, November 18, 2014
Accepted, January 2, 2015