Vietnam Journal of Computer Science (2018) 5:177–184
https://doi.org/10.1007/s40595-018-0113-0

REGULAR PAPER

Revisiting urban air quality forecasting: a regression approach

Kostas Karatzas[1] · Nikos Katsifarakis[1] · Cezary Orlowski[2] · Arkadiusz Sarzyński[3]

Received: 23 August 2017 / Accepted: 11 May 2018 / Published online: 24 May 2018

Abstract
We address air quality (AQ) forecasting as a regression problem employing computational intelligence (CI) methods for the Gdańsk Metropolitan Area (GMA) in Poland and the Thessaloniki Metropolitan Area (TMA) in Greece. Linear Regression as well as Artificial Neural Network models are developed, accompanied by Random Forest models, for five locations per study area and for a dataset of limited feature dimensionality. An ensemble approach is also used for generating and testing AQ forecasting models. Results indicate good model performance with a correlation coefficient between forecasts and measurements for the daily mean PM_{10} concentration one day in advance reaching 0.765 for one of the TMA locations and 0.64 for one of the GMA locations. Overall results suggest that the specific modelling approach can support the provision of air quality forecasts on the basis of limited feature space dimensionality and by employing simple linear regression models.

Keywords Computational intelligence · Air pollution · Regression models · Ensemble

1 Introduction

In a recently published paper [1] we underlined the importance of air quality (AQ) forecasting in urban environmental management as well as in contemporary smart city development [2,3]. In the current paper we revisit and extend our initial approach, focusing on urban AQ forecasting from the regression point of view and incorporating an ensemble modelling approach. For doing so, we take into account that in the framework of smart city information systems, environmental management plays an important role [4] and air pollution abatement is one of its main targets [5]. Air Quality forecasting is among the main pillars of AQ management [6] and is materialized with the aid of appropriate AQ models. Such models are establishing a time-varying relationship between the concentration of air pollutants at a specific time and location $\boldsymbol{c}(t, \boldsymbol{x})$, and other parameters $\boldsymbol{p}(t, \boldsymbol{x})$ affecting the urban atmospheric environment. Such a relationship may be expressed with the aid of the following general function:

$$\boldsymbol{c}(t, \boldsymbol{x}) = f(\boldsymbol{p}(t, \boldsymbol{x})) \tag{1}$$

Here t represents time and $\boldsymbol{x}$ is the location vector corresponding to physical space. In this case the vector $\boldsymbol{c}(t, \boldsymbol{x})$ refers to concentration values of air pollutants like Nitrogen Dioxide (NO_2), Carbon Monoxide (CO), Ozone (O_3) and Particulate Matter (PM), while $\boldsymbol{p}(t, \boldsymbol{x})$ includes parameters like wind speed, wind direction, air temperature, solar radiation, air pollutant emissions, air pollutant concentrations, land use type, land surface height, etc. The nature of function f is dictated by the model type employed: thus, if f reconstructs the physical and chemical relationships between the parameters $\boldsymbol{p}(t, \boldsymbol{x})$ and values $\boldsymbol{c}(t, \boldsymbol{x})$, where $\boldsymbol{x}$ addresses the whole area of interest in a 3-D gridded manner, then models are said to follow an analytic-deterministic approach [7], while if f is a statistical or data-mining oriented function, then models are said to follow a data-driven approach (as reported in [8] and in references therein). In the latter case, $\boldsymbol{x}$ refers to specific areas within the studied area, which usually

✉ Kostas Karatzas
kkara@auth.gr

Nikos Katsifarakis
nikolakk@auth.gr

Cezary Orlowski
corlowski@wsb.gda.pl

Arkadiusz Sarzyński
arek3108@gmail.com

[1] Department of Mechanical Engineering, Environmental Informatics Research Group, Aristotle University, Thessaloniki, Greece

[2] Institute of Management and Finance, WSB University in Gdańsk, Gdańsk, Poland

[3] Department of Applied Business Informatics, Faculty of Management and Economics, Gdańsk University of Technology, Gdańsk, Poland

correspond to AQ measuring station locations. Thus, $\boldsymbol{x}$ is not varying with time and is excluded, leading to an equation of the form:

$$\boldsymbol{c}(t) = f(\boldsymbol{p}(t)) \qquad (2)$$

The objective of this paper is to suggest CI-based, ensemble oriented models that are able to depict as much information as possible from atmospheric quality data of low dimensionality, and to thus contribute to the scientific area of urban AQ forecasting. For this reason we employ a variety of CI methods and we suggest and test ensemble functions f in Eqs. (1) and (2). The geographic areas of interest are the Gdańsk Metropolitan Area (GMA) in Poland and the Thessaloniki Metropolitan Area (TMA) in Greece, and the parameter of interest is the daily concentration of Particulate Matter with a mean aerodynamic diameter of 10 μm (PM_{10}), approx. $1/5^{\text{th}}$ of the diameter of the human hair. The specific pollutant is able to penetrate in the bronchial part of the human lung system [9] and is one of the most important pollutants in the GMA [10] as well as in the TMA [11]. Air pollutant concentrations are addressed as numerical values. AQ forecasting follows a twofold approach:

a) Each AQ monitoring station is treated individually, i.e. AQ models are developed and tested per station location. Thus, the forecasting of the parameter of interest is performed as a regression problem.
b) Regression models are being created based on ensemble modelling principles, and are evaluated via their ability to forecast AQ levels at different locations (i.e. at each monitoring station).

The mean daily concentration level of PM_{10} one day in advance is the target of the forecasting models under development. This choice corresponds to the requirements posed by relevant legislation for citizens as well as the decision makers to be informed about the expected PM_{10} levels for the next day, not to exceed 50 $\mu g/m^3$ more than 35 days per year according to the European Regulations [9,12] and according to the World Health Organization guidelines [13]. Combustion processes, traffic and natural sources directly emit PM_{10}, while in some regions the mechanical degradation of the road surface and of winter tires also contributes to its production. PM_{10} are part of the inhalable fraction of PM and have adverse effects to human health [9].

The research question posed in the current paper moves one step ahead of our previously published results [1] and addresses (a) the ability of a low dimensionality feature space (small number of input parameters) to support effective data-driven models for PM_{10} forecasting and (b) the modelling approach to be used in terms of algorithms and their setup (single vs. ensemble oriented models). In addition, we make use of an ensemble approach based on an ANN model of simple architecture which can be applied to multiple geographic areas, thus simplifying the ensemble approach suggested by [14] and [15], while maintaining a performance comparable to the one reported by similar studies [16], and therefore providing with a novel approach to the problem at hand.

In the rest of the paper we firstly present the materials of our study (Chapter 2), followed by the computational methods (Chapter 3). Then we proceed with the presentation and the discussion of the results in Chapter 4, and we draw our conclusion in Chapter 5.

2 Materials: area of study and data made available

The areas of study as well as the AQ problem addressed have been the focus of multiple studies performed in the past.

In the case of Gdansk ANNs have been employed for AQ forecasting in [17]. The same data set has been used for PM_{10} forecasting in [18] as well as for the adaptation of an AQ forecasting model developed for Gdansk to the Thessaloniki area [19].

The air pollution of Thessaloniki has been studied and modeled with the aid of ANNs [20], with special emphasis on PM_{10} [21]. The similarity of the GMA as well as of the TMA in terms of population and existence of a sea front suggest that there might also be a similarity in the way that PM_{10} oriented air pollution can be modeled in both areas. Moreover, the need for the construction of data-driven models which use a small number of input parameters, suggested that a generalized, ensemble-based approach should be employed for the AQ modeling in both areas of interest, these being the novelty points of the research results at hand.

2.1 The two areas of interest

The city of Gdańsk is located on the Baltic coast in the southwest of the bay of Gdańsk, in the northern part of Poland. It is the capital of a tri-city metropolitan area merging with Gdynia (known for its shipyards) and Sopot (a recreational resort) and adding more than 1,000,000 residents in the GMA taking into account suburban communities also. The economy in Gdańsk is dominated by shipbuilding, petrochemicals and chemical industries, which are all concentrated quite close to the city center. The majority of air pollutant emissions originate from the industrial sector, the port activities and the city traffic [22], while the most important pollutants are PM_{10}, NO_2 and SO_2 (http://www.airqualitynow.eu).

The city of Thessaloniki faces an oval harbor bay and stands on a rising ground at the heart of a long gulf which is formed by the peninsula of Chalcidice. Various municipalities surround the city while an industrial zone is located

Table 1 The Air Quality monitoring stations used for the current study in GMA and TMA

GMA stations	AMI (Gdańsk-Śródmieście), AM2 (Gdańsk-Stogi), AM3 (Gdańsk-Nowy Port), AM4 (Gdynia-Pogórze), AM5 (Gdańsk-Szadółki)
TMA stations	Egnatia, Martiou, Lagkada, Eptapyrgiou, Malakopi

in the north-west of its outskirts. The TMA is the second largest urban agglomeration in Greece accounting for more than 1,000,000 inhabitants, with a considerable accumulation of urban traffic as well as industrial activities. The TMA is characterized by high pollution levels especially related to PM_{10} while O_3 appears to be high in suburban locations of the area and NO_2 levels are still high in dense urban areas in association with traffic [11].

2.2 The atmospheric quality data

In both the GMA and in the TMA a number of AQ monitoring stations operate (9 and 17 respectively), which routinely record concentration values of basic pollutants as well as the variation of meteorological parameters. As not all pollutants are recorded at all stations, and in order to focus on the pollutant of interest (PM_{10}), we decided to select five stations from each area of interest (included in Table 1), that were able to provide with PM_{10} concentrations as well as meteorological data, in order to come up with data sets that are identical in terms of the parameters they include. In order to deal with the non-negligible frequency of missing data, we selected data from the year 2013 which contained only daily PM_{10} concentrations as well as information for air temperature and relative humidity.

As a result and for each station, the same atmospheric parameters were used for the modelling and forecasting process: the model input or feature vector $\boldsymbol{x}$ included five parameters, namely PM_{10} concentration of the current day as well as temperature and relative humidity of the current day, complemented by the day and the month of the year. The target parameter to be forecasted $\boldsymbol{y}$ was the PM_{10} concentration of the next day. A summary of the basic statistical characteristics of the parameters involved in our study is included in Table 2.

3 Computational methods

The forecasting of the numerical value of PM_{10} concentration levels for the next day was the goal set for the development of relevant forecasting models. For this reason, we made use of the available datasets for each AQ monitoring station to develop individual (per station) AQ forecasting models.

3.1 Algorithms for single station model creation

The algorithms applied were selected based on computational experiments employing various CI methods, which were conducted with the aid of Matlab (www.mathworks.com) as well as of the WEKA computational environment [23]. On this basis, we chose the following three algorithms as the basis for AQ model development:

(i) Linear Regression (LR). Here the relationship between the forecasted parameter and the input parameters are described by an equation of the form:

$$\boldsymbol{y} = \boldsymbol{x} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

where $\boldsymbol{x}$ is the input vector, $\boldsymbol{\beta}$ is the slope vector and $\boldsymbol{\varepsilon}$ the error vector. The slope vector is commonly calculated via the least square method, thus:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{x}' \cdot \boldsymbol{x})^{-1} \cdot \boldsymbol{x}' \cdot \boldsymbol{y} \tag{4}$$

(ii) Artificial Neural Networks (ANNs). In ANNs the input vector $\boldsymbol{x}$ for each neuron k, is weighted with the aid of a weighting vector $\boldsymbol{w}_k$, and the result is summed (taking into account any bias) and then fed into a transfer function f to produce the overall output vector $\boldsymbol{y}_k$:

$$\boldsymbol{y}_k = f(\boldsymbol{w}_k^T \cdot \boldsymbol{x}) \tag{5}$$

The training of the ANN aims at reducing the error $\boldsymbol{e}_k$ between the model output $\boldsymbol{y}_k$ and the actual (real) value observed $\boldsymbol{d}_k$, which here is the PM_{10} concentration of the next day for each station.

$$\boldsymbol{e}_k = \|\boldsymbol{y}_k - \boldsymbol{d}_k\| \tag{6}$$

This error reduction is based on a number of methods all of which aim at recalculating the initial weights so that the overall network error is minimized. In the case of the gradient descent method (which is the simples of all but nevertheless representative of the way that the weights are recalculated), the relationship between the updated and the initial weighting vector for all neurons k of the ANN, is given by:

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - a(t)\boldsymbol{g}(t) \tag{7}$$

Here t and $t+1$ denote the initial and the updated weights, while the error term is described by:

$$\boldsymbol{g}(\boldsymbol{t}) = \boldsymbol{J}^{\boldsymbol{T}}(\boldsymbol{t}) \cdot \boldsymbol{e}(\boldsymbol{t}) \tag{8}$$

where $\boldsymbol{J}^T$ is the (transposed) Jacobian and $\boldsymbol{e}(t)$ is the overall error vector [1].

Table 2 Basic statistics for the AQ and meteorological parameters available for each station at GMA and TMA

Datasets	PM10 (in µg/m^3)				Temperature (in °C)				Humidity (%)			
	Min	Max	Mean	Std	Min	Max	Mean	Std	Min	Max	Mean	Std
AM1	6	92	20.55	10.48	−11.1	24.5	8.03	7.93	48	100	81.56	11.00
AM2	6	66	21.45	10.57	−11.1	24.5	7.41	7.77	48	100	82.00	10.88
AM3	3	79	16.90	10.14	−11.1	24.5	7.71	7.95	48	100	81.73	11.00
AM4	0	61	16.97	10.20	−11.1	24.5	7.82	7.92	48	100	81.65	10.97
AM5	0	55	15.01	8.35	−11.1	24.5	7.82	7.92	48	100	81.65	10.97
Egnatia	18	131	48.21	19.67	1.6	31.4	18.14	7.48	29	88	59.17	13.43
Martiou	9	113	34.44	18.69	1.3	31	18.13	7.58	33	87	60.80	13.00
Lagkada	20	244	57.04	32.62	0.8	31.6	18.01	7.76	33	89	60.56	13.39
Eptapyrgiou	8	135	28.90	18.03	−0.7	30.1	16.88	7.45	31	94	60.07	15.34
Malakopi	7	119	29.18	17.41	0.1	29.7	16.91	7.48	31	89	61.39	14.56

In this specific case a MultiLayer Perceptron Network with a feed-forward architecture and a back propagation training method was used, with an input layer consisting 5 nodes (i.e. all the input parameters per station), an output layer consisting of only one node (the PM$_{10}$ concentration of the next day) and a hidden layer with 10 nodes. The sigmoid function is employed as the transfer function while the gradient descent algorithm is used for minimizing the error function.

```
Input: Size_of_Problem, Input, Iterations,  Learning_Rate
Output: ANNetwork
ANNetwork  ← Construct_Network_Layers()
ANNetworkWeights ← Initialize_Weights(ANNetwork, Size_of_Problem)
For (j = 1 To Iterations)
   Pattern(j) ← Select_Input_Pattern(Input)
   Output(j) ← Propagate_Forward (Pattern(j), ANNetwork)
   Propagate_Error_Backwards (Pattern(j), Output(j), ANNetwork)
   Update_Weights(Pattern(j), Output(j), ANNetwork, Learning_Rate)
End
Return (ANNetwork)
```

(iii) Random Forests (RF), an ensemble method originating from the Decision Tree family of algorithms [24] that has shown high capacity to effectively model atmospheric parameters of interest [1]. The method creates N subsets of the input vector $\boldsymbol{x}$ using random selection with replacement, each subset containing 2/3 of the initial data, while the remaining data are used to estimate error and variable importance. Then for each subset, a decision tree is created with the aid of an arbitrary number of nodes, where for each node the splitting is based on a (randomly) selected subset of L attributes that optimize a target function (best split criterion). In our case $L = \text{int}[log_2(\text{Number of attributes}) + 1]$. Each of the aforementioned random trees had an unlimited number of levels and nodes. The prediction created by each tree is averaged and thus the ensemble-based overall prediction of the RF (here the PM$_{10}$ concentration of the next day) is generated. A pseudocode for this method based on http://dataaspirant.com/ is presented below:

```
Input: Features, Num_of_Nodes, Num_of_Trees, BestSplitCriterion
Output: RandomForest
For (j = 1 To Num_of_Trees)
    While(Nodes .lt.Num_of_Nodes)
        Randomly select "k" features from total "m" features.
            Where $k << m$
        Among the "k" features, calculate the node "d" using the best split point.
        Split the node into daughter nodes using the BestSplitCriterion.
    End
End
Return (RandomForest)
```

The prediction is then made on the basis of an ensemble of results based on voting for each one of the trees generated.

3.2 Ensemble models

In addition to the above approach, we investigated the possibility to develop ensemble-based models to be common for all monitoring stations. More specifically:

1. A single ensemble model was created for each one of the two areas of interest, and then applied to all individual AQ monitoring stations for the same area (local ensemble).
2. The ensemble created in the one of the geographic areas was applied to each one of AQ monitoring stations of the other geographic area (foreign ensemble).
3. Both local and foreign ensembles are combined to generate a cross ensemble model, which is then applied to each one of the AQ monitoring stations for both geographic areas of interest.

The aforementioned approach was materialized for both LR and ANN models as follows:

1. Local ensemble: In the case of LR, the parameters of the slope vector $\boldsymbol{\beta}$ of the ensemble model were calculated as weighted mean values of the parameters of each one of the individual LR models, and the local ensemble model was then applied to all stations. In the case of the ANN models, the weights of the individual models were used for the calculation of the weighted mean value of the weights of the local ensemble model. In both cases, the weighted means were calculated on the basis of the correlation coefficients of each one of the models participating in the ensemble, as resulting from their application to the monitoring station for which they were developed.
2. Foreign ensemble: the calculation was done exactly as in the case of the local ensemble, yet making use of the foreign individual model slope vectors (for LR) and weights (for ANN) instead of the local individual model characteristics.
3. Cross ensemble: the parameters of the local and the foreign ensemble models were averaged in order to calculate the parameters of the cross ensemble models.

3.3 Model validation

In order to validate the results of the PM_{10} predictions, it is important to make use of as many of the available data as possible for the training as well as for the testing phase. For this reason we followed a 10-fold cross validation procedure [25] for each one of the individual models developed: we randomly divided the initial dataset into 10 equal subsets. Then 9 out of these datasets were used for training the model, while the 10th one was used for testing, This process was repeated 10 times, each time leaving a different subset out of the training phase and using it for the test phase. The overall model results are the mean values of the statistical indices of the 10 models developed. Concerning the ensemble models, these were defined on the basis of the (pre-existing) individual models per algorithm used, and therefore no additional model validation was used.

Model results were evaluated based on the following statistical indices:

(a) Pearson's correlation coefficient $\boldsymbol{r}$ that describes the degree of linear relationship between forecasted and real PM_{10} concentration values.
(b) Mean Absolute Error (MAE), which is a measure of the mean absolute distance between forecasted and real values.
(c) Root Mean Squared Error (RMSE), which is the square of the Mean Square Error and expresses the standard devi-

Table 3 Correlation coefficient ***(r)***, Mean absolute error (MAE) and Root mean square error (RMSE) for three models per monitoring station concerning the forecast of the mean daily PM10 concentration one day in advance

Datasets	Random forest			ANN (Multilayer perceptron)			Linear Regression (Multivariate)		
	r	MAE	RMSE	*r*	MAE	RMSE	*r*	MAE	RMSE
AM1	0.530	6.441	8.947	0.226	8.244	12.168	0.545	6.380	8.767
AM2	0.456	7.322	9.843	0.361	8.202	10.696	0.479	7.206	9.599
AM3	0.401	6.105	7.785	0.233	7.528	10.572	0.406	6.758	9.306
AM4	0.601	6.007	8.235	0.427	7.379	9.787	0.641	5.581	7.821
AM5	0.592	4.853	6.821	0.301	5.373	7.400	0.607	4.754	6.690
Egnatia	0.664	10.381	14.710	0.506	12.610	16.671	0.693	9.935	14.118
Martiou	0.731	8.791	12.715	0.563	11.771	15.788	0.732	8.851	12.666
Lagkada	0.713	15.157	22.989	0.571	17.996	25.590	0.728	15.050	22.391
Eptapyrgiou	0.742	7.497	12.000	0.587	11.633	16.229	0.720	8.057	12.390
Malakopi	0.723	7.871	12.014	0.617	9.829	14.753	0.742	7.800	11.639

ation of the differences between forecasted and actual values.

4 Results and discussion

Based on the model calculations performed as described in Chapter 3, the Pearson's correlation coefficient ***r*** accompanied by the Mean Absolute Error and the Root Mean Squared Error were calculated for the three models developed and for each one of the ten AQ monitoring stations for which data were available (Table 3).

Results suggest that the algorithm leading to the best (highest) correlation coefficient between forecasted and monitored values is LR, with an ***r*** ranging from 0.406 for station AM3 up to 0.641 for station AM4 for the GMA. Concerning the TMA, LR is again the best algorithm in terms of the highest correlation coefficient achieved, with an ***r*** value ranging from 0.72 for Eptapyrgiou station up to 0.742 for the Malakopi station. The RF algorithm can be ranked as 2nd, achieving correlation coefficients very close to the ones received with the aid of LR (and surpassing it for the Eptapyrgiou station), while in some cases leading to the best (lower) MAE (like in the AM3, Martiou and Eptapyrgiou stations) and to the best (lower) RMSE (like in the AM3 and in the Eptapyrgiou stations). LR is a simple algorithm of linear logic generally considered weak in depicting nonlinear phenomena like the ones involved in AQ problems, and usually performing more poorly when compared with algorithms like ANNs or RF [1]. The success of the specific algorithm in our case has to do with the limited number of atmospheric quality parameters being available in all studied areas and stations (low number of features), thus leading to the (possible) exclusion of nonlinear dependencies from the available dataset, and dictating persistence as the main mechanism affecting the forecast of PM_{10} levels one day in advance [26].

In the case of the ensemble approach used (local, foreign and cross ensembles), the results of the two algorithms employed (LR and ANN) are presented in Table 4. The optimum ensemble approach is selected on the basis of the highest correlation coefficient achieved and taking in parallel with the lowest possible error metric values (MAE and RMSE). On this basis the local ensemble achieves the best results, followed by the cross ensemble and leaving the foreign ensemble last. The result may be attributed to the ability of the local ensemble to better represent the dependencies between the modelled parameter (mean daily PM_{10} concentration for the next day) and the parameters of the feature space (input parameters). In terms of algorithms employed, LR is always better in comparison to ANNs. Concerning the areas of study ***r***, values range from 0.505 (station AM2) up to 0.64 (station AM4) for the GMA, while ***r*** values range from 0.710 (station Egnatia) up to 0.765 (station Malakopi) for the TMA. The value range of the correlation coefficient achieved for the TMA corresponds to a value range of the coefficient of determination (which is actually the correlation coefficient squared) between 0.504 (for Egnatia) and 0.585 (for Malakopi), which are better in comparison to the values achieved for the TMA but for two different stations, as reported by [27] and [28].

By comparing ensembles with the local models, it is evident that in the case of LR-based models, the local ensemble provides with a better performance in comparison to the local models for all GMA stations with the exception of AM4, while in the case of the TMA local ensembles outperform local models for three out of five stations (Lagkada, Eptapyrgiou and Malakopi). In the case of ANN modes, both the local ensemble and the local models perform almost equally in terms of correlation coefficient values achieved.

Table 4 Results of the local, foreign and cross ensemble models for the ANN and LR algorithms in all stations of the GMA and TMA

Datasets	ANN-local ensemble			ANN-foreign ensemble			ANN-cross ensemble			LR-local ensemble			LR-foreign ensemble			LR-cross ensemble		
	r	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE
AM1	0.242	7.935	11.453	0.192	9.101	13.637	0.220	8.432	12.684	0.561	6.018	8.852	0.525	6.512	9.165	0.558	6.161	8.962
AM2	0.360	8.210	10.801	0.332	8.793	11.249	0.342	8.402	10.835	0.505	6.935	9.719	0.361	8.935	11.292	0.442	7.422	10.219
AM3	0.232	7.612	10.713	0.202	8.234	11.183	0.230	7.712	10.812	0.453	5.922	9.752	0.337	7.815	10.621	0.405	6.760	9.412
AM4	0.433	7.254	9.610	0.400	8.013	10.615	0.416	7.531	10.258	0.640	5.601	8.016	0.584	6.215	9.121	0.624	5.686	8.174
AM5	0.308	5.310	7.284	0.282	5.619	7.735	0.294	5.593	7.634	0.629	4.611	6.731	0.441	6.125	8.017	0.545	4.951	7.192
Egnatia	0.511	12.518	16.619	0.422	14.195	18.490	0.482	13.151	16.215	0.710	9.911	14.183	0.681	10.165	14.317	0.705	9.932	14.209
Martiou	0.563	11.767	15.731	0.442	14.015	18.183	0.542	12.835	16.015	0.741	8.841	12.673	0.712	9.252	12.851	0.737	8.853	12.672
Lagkada	0.572	18.104	26.107	0.451	22.526	30.015	0.532	20.015	28.124	0.749	15.001	22.843	0.727	15.053	22.481	0.743	15.009	22.912
Eptapyrgiou	0.588	11.652	16.310	0.452	13.124	19.258	0.551	12.106	17.514	0.742	7.951	12.414	0.721	8.053	12.415	0.743	7.939	12.420
Malakopi	0.616	9.848	14.981	0.442	13.253	18.315	0.561	11.257	16.938	0.765	7.716	11.863	0.727	8.183	12.285	0.757	7.731	11.915

5 Conclusions

In this paper, we address the problem of air quality forecasting for two different geographical areas of interest, the GMA and the TMA, by employing a regression approach, making use of a limited dimension feature space, and targeting at the forecast of the mean daily PM_{10} concentration of the next day. We initially develop location specific models by employing ANNs, LR and RF, and achieving correlation coefficients between 0.406 and 0.641 for the GMA stations, and between 0.693 and 0.742 for the TMA stations. The best performance was provided by the LR models, followed by the RF and the ANN models. In addition, we developed and tested three types of ensemble models per area, namely the local, the foreign and the cross ensemble models. Their application proved the local ensemble models to be the superior for both ANNs and LR algorithms. These results indicate that even when the feature space is of limited dimensionality, the best individual model outperforms the common model for all the monitoring stations, making use of the ensemble principle, and employing the recalculation of weights in a simple LR model. This suggests that city authorities may develop effective AQ models by targeting their investment in AQ monitoring to the parameters of interest, a vast feature space not being necessary for the success of the modelling approach.

In terms of geographic area of interest, models for the GMA present with a lower overall performance in comparison to TMA models, regardless of the algorithm employed. Taking into account that in both areas the same features were made available and used for the development of the relevant models, this result indicates the importance of additional feature space parameters (reflecting atmospheric mechanisms) in order to further improve modelling performance. When coming to the choice of algorithms for the development of AQ models, the superiority of LR-based models in our study supports the finding that in the case of feature spaces of low dimension, the basic mechanisms which influence the quality of the atmospheric environment are persistence and linear dependencies. This result is of use for those wishing apply AQ models in the frame of an urban environmental management system, having a low-dimension feature space available for model deployment.

References

1. Karatzas, K., Katsifarakis, N., Orlowski, C. Sarzyński A.: Urban air quality forecasting: a regression and a classification approach. In: In Nguyen N.T. et al. (eds.): Intelligent information and database systems, 9th Asian Conference on Intelligent Information and Database Systems, Part II, Lecture Notes in Artificial Intelligence vol. 10192, pp. 1–10 (2017). https://doi.org/10.1007/978-3-319-54430-4_52
2. Riffat, S., Powell, R., Aydin, D.: Future cities and environmental sustainability. Future Cities Environ. **2**, 1 (2016). https://doi.org/10.1186/s40984-016-0014-2
3. Webel, S.: Forecasting Software that's a Breath of Fresh Air. Pictures of the Future Siemens Magazine, (2016) http://www.siemens.com/innovation/en/home/pictures-of-the-future/infrastructure-and-finance/smart-cities-air-pollution-forecasting-models.html. Accessed 18 Aug 2017
4. Dawe, S. Paradice, D.: A systems approach to smart city infrastructure: a small city perspective. In: Proceedings of the Thirty Seventh International Conference on Information Systems, Dublin, http://iot-smartcities.lero.ie/wp-content/uploads/2016/12/A-Systems-Approach-to-Smart-City-Infrastructure-A-Small-City-Perspective.pdf. Accessed 18 Aug 2017
5. Marinov, M.B., Topalov, I., Gieva, E., Nikolov, G.: Air quality monitoring in urban environments. In: 39th International Spring Seminar on Electronics Technology (ISSE), Pilsen, pp. 443–448. (2016). https://doi.org/10.1109/ISSE.2016.7563237
6. Bukoski, B., Taylor, E.M.: Air quality forecasting. Air quality management 129–138 (2014)
7. Kukkonen, J., Olsson, T., Schultz, D.M., Baklanov, A., Klein, T., Miranda, A.I., Monteiro, A., Hirtl, M., Tarvainen, V., Boy, M., Peuch, V.-H., Poupkou, A., Kioutsioukis, I., Finardi, S., Sofiev, M., Sokhi, R., Lehtinen, K.E.J., Karatzas, K., San José, R., Astitha, M., Kallos, G., Schaap, M., Reimer, E., Jakobs, H., Eben, K.: A review of operational, regional-scale, chemical weather forecasting models in Europe. Atmos. Chem. Phys. **12**, 1–87 (2012)
8. Karatzas, K., Kaltsatos, S.: Air pollution modelling with the aid of computational intelligence methods in Thessaloniki, Greece. Simul. Modelling Pract. Theory **15**(10), 1310–1319 (2007)
9. EEA, 2016: Air quality in Europe—2016 report, European Environment Agency, https://doi.org/10.2800/80982. https://www.eea.europa.eu//publications/air-quality-in-europe-2016. Accessed 18 Aug 2017
10. Juda-Rezler, K., Trapp, W., Reizer, M.: Modelling the impact of climate changes on particulate matter levels over Poland. In: Steyn, D.G., Rao, S.T. (eds.) Air pollution modeling and its application XX, pp. 499–450 (2010)
11. Moussiopoulos, N., Vlachokostas, C., Tsilingiridis, G., Douros, I., Hourdakis, E., Naneris, C., Sidiropoulos, C.: Air quality status in Greater Thessaloniki Area and the emission reductions needed for attaining the EU air quality legislation. Sci. Total Environ. **407**(4), 1268–1285 (2009)
12. Andrews, A.: The clean air handbook, a practical guideline to EU air quality law, https://www.clientearth.org/reports/20140515-clientearth-air-pollution-clean-air-handbook.pdf. Accessed 18 Aug 2017
13. WHO: Air Quality Guidelines, global update 2005, ISBN 92 890 2192 6 via http://www.euro.who.int. Accessed 18 Aug. 2017
14. Siwek, K., Osowski, S.: Improving the accuracy of prediction of PM10 pollution by the wavelet transformation and an ensemble of neural predictors. Eng. Appl. Artif. Intel. **25**(6), 1246–1258 (2012)
15. Zhou, Q., Jiang, H., Wang, J., Zhou, J.: A hybrid model for PM2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network. Sci. Total Environ. **496**, 264–274 (2014)
16. Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., Di Tommaso, S., Colangeli, C., Rosatelli, G., Carlo, P.: Recursive neural network model for analysis and forecast of PM10 and PM2.5. atmospheric. Pollut. Res. **8**(4), 652–659 (2017)
17. Khokhlov, V.N., Glushkov, A.V., Loboda, N.S., Bunyakova, Y.Y.: Short-range forecast of atmospheric pollutants using non-linear prediction method. Atmos. Environ. **42**(31), 7284–7292 (2008)
18. Orłowski, C., Sarzyński, A.: A model for forecasting pm10 levels with the use of artificial neural networks. In: Information Systems Architecture and Technology—the use of IT Technologies to Support Organizational Management in Risky Environment, Wrocław (2014)
19. Orłowski, C., Sarzyński, A., Karatzas, K., Katsifarakis, N., Nazarko J.: Adaptation of an ANN-based air quality forecasting model to a new application area. In: Król D., Nguyen N., Shirai K. (eds) Advanced Topics in Intelligent Information and Database Systems 479-488 (2017)
20. Karatzas, K., Kaltsatos, S.: Air pollution modelling with the aid of computational intelligence methods in Thessaloniki, Greece. Simul. Model. Pract. Theory **15**(10), 1310–1319 (2007)
21. Voukantsis, D., Karatzas, K., Kukkonen, J., Räsänen, T., Karppinen, A., Kolehmainen, M.: Intercomparison of air quality data using principal component analysis, and forecasting of PM10 and PM2.5 concentrations using artificial neural networks, in Thessaloniki and Helsinki. Sci. Total Environ. **409**, 1266–1276 (2011)
22. Szczepaniak, K., Astel, A., Bode, P., Sârbu, C., Biziuk, M., Raińska, E., Gos, K.: Assessment of atmospheric inorganic pollution in the urban region of Gdańsk. J. Radioanal. Nuclear Chem. **270**(1), 35–42 (2006)
23. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: an update. SIGKDD Explorations **11**(1), 10–18 (2009)
24. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
25. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. Proc. Fourteenth Int. Joint Conf. Artif. Intel. **2**(12), 1137–1143 (1995)
26. EPA: Guidelines for developing an air quality (ozone and PM2.5) forecasting program, U.S. Environmental Protection Agency report EPA-456/R-03-002, https://www3.epa.gov/airnow/aq_forecasting_guidance-1016.pdf. Accessed 18 Aug 2017
27. Voukantsis, D., Niska, H., Karatzas, K., Riga, M., Damialis, A., Vokou, D.: Forecasting daily pollen concentrations using data-driven modeling methods in Thessaloniki, Greece. Atmos. Environ. **44**(39), 5101–5111 (2010)
28. Tzima, F., Mitkas, P., Voukantsis, D., Karatzas, K.: Sparse episode identification in environmental datasets: the case of air quality assessment. Expert Syst. with Appl. **38**(5), 5019–5027 (2011)