Atmospheric Pollution Research xxx (xxxx) xxx–xxx

HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

Atmospheric Pollution Research

journal homepage: www.elsevier.com/locate/apr

Statistical models for multi-step-ahead forecasting of fine particulate matter in urban areas

Ida Kalate Ahani, Majid Salari*, Alireza Shadman

Department of Industrial Engineering, Ferdowsi University of Mashhad, P.O. Box 91779-48951, Mashhad, Iran

ARTICLE INFO

Keywords:
Multi-step-ahead
$PM_{2.5}$ forecasting
Fine particulate matter
Statistical models
ARIMAX
ANN
Long-term forecasting

ABSTRACT

In recent years, the atmospheric pollution in most metropolitan cities has become a crisis and the necessity of air quality forecasting has increased. Among different air pollutants, $PM_{2.5}$ is considered as the major air pollutant in urbanized regions, especially because of serious harmful health effects on human being. So, there is an urgent need to develop air quality forecast programs capable of providing accurate predictions over a long future horizon. Predicting $PM_{2.5}$ concentrations for several steps ahead of time is of great interest, especially in decision-making related to control policies and emergency measures such as traffic limitations, school closures, or temporarily shutting down major polluting industrial units. In this paper, commonly used multi-step ahead prediction strategies, including Recursive (Rec), Direct (Dir), Direct-Recursive (DirRec), Multi-Input Multi-Output (MIMO) and Direct-MIMO (DIRMO) along with Autoregressive integrated moving average with exogenous variables (ARIMAX) and Multi-Layer Perceptron (MLP) modelling techniques are examined. Also, the independent variables are considered as time series variables and are forecasted using ARIMA/MLP model in order to be used for prediction of the dependent variables in multi-steps ahead of time. The experimental study is performed using $PM_{2.5}$ data in Mashhad, Iran. Daily $PM_{2.5}$ forecasts for this city is provided for the next 10 days. Four different feature selection methods are also implemented and compared. The results indicate that recursive strategy with LASSO feature selection in ARIMAX model overcomes in most of time steps.

1. Introduction

In this section, the research background, literature review and the aims and innovations of this paper are described in detail.

1.1. Background

In the recent era, air pollution issues have attracted worldwide attention, because it has a great impact on human, other living organisms and environment. Major air pollutants include SO_2, NO_2, NO, O_3, CO, PM_{10} and $PM_{2.5}$. Among them, $PM_{2.5}$, atmospheric particulate matter with a diameter less than 2.5 μm, is considered as the major air pollutant in urbanized regions, and has the greatest effect on health. Exposure to this pollutant is associated with cardiovascular diseases (Brook et al., 2010; Lippmann, 2014), respiratory diseases (Guaita et al., 2011; Tecer et al., 2008), cerebrovascular diseases (Santibañez et al., 2013), low birth weight in infants (Coker et al., 2016), diabetes (Pearson et al., 2010), etc. The significant increase in air pollution during recent years, especially in most metropolitan cities has become a crisis and endangers human health. In order to respond to this problem, there is an urgent need to develop air quality forecast programs. Development of an accurate and robust air quality forecasting system with the ability of providing long-term predictions, can improve public awareness about severe pollution episodes. Therefore, susceptible individuals with cardiovascular or respiratory disease, children, pregnant women and older adults should take precautions to avoid exposure to air pollution and reduce outdoor activities (Lv et al., 2016; Zhang et al., 2012). When air quality alert is issued, ordinary people should also avoid outdoor exercise and physical activities. Furthermore, control policies such as temporarily shutting down major polluting industrial units or traffic limitations can be taken into account (Zhang et al., 2012), especially in populous areas where a large number of people are affected.

1.2. Literature review

An extensive literature review reveals that there has been a significant growth in the number of published articles dealing with $PM_{2.5}$

Peer review under responsibility of Turkish National Committee for Air Pollution Research and Control.
* Corresponding author.
E-mail address: msalari@um.ac.ir (M. Salari).

https://doi.org/10.1016/j.apr.2018.11.006
Received 28 August 2018; Received in revised form 13 November 2018; Accepted 13 November 2018

forecasting (Jiang et al., 2017; Niska et al., 2005; Voukantsis et al., 2011; Wang et al., 2018), which shows understanding the need for prediction systems to reduce the harmful effects on human health and other living organisms. In this section, literature on $PM_{2.5}$ forecasting and also multistep ahead prediction in different fields are reviewed.

1.2.1. Review on $PM_{2.5}$ prediction

Different methods have been used for $PM_{2.5}$ forecasting in urban areas. Neural Networks, regression, time series analysis and fuzzy approaches are the most commonly used techniques to predict $PM_{2.5}$ levels. In 2000, Pérez et al. proposed MLP to predict hourly concentrations of $PM_{2.5}$ for the next hours in Santiago de Chile (Pérez et al., 2000). Ordieres et al. compared three different neural network methods, MLP, radial basis function, and square MLP to predict daily averages of $PM_{2.5}$ concentrations along the US-Mexico border (Ordieres et al., 2005). In another study for the city of Santiago (Perez and Salini, 2008), MLP, a linear algorithm and a hybrid clustering algorithm have been employed to forecast maximum of the 24 h moving average of $PM_{2.5}$ for the next day. The results obtained by hybrid clustering algorithm were found to be more accurate for detecting high concentration situations. Recent studies in air pollution forecasting have focused on applying hybrid models to improve the accuracy of predictions. Prakash et al. proposed a wavelet based Elman recurrent neural network (ERNN) to predict hourly concentrations of five different pollutants ($PM_{2.5}$, CO, NO_2, NO, O_3, SO_2) one hour in advance, in Delhi, India (Prakash et al., 2011). A hybrid model based on back-propagation neural network optimized by cuckoo search and ensemble empirical mode decomposition was addressed by Qin et al. to forecast $PM_{2.5}$ and PM_{10} concentrations in four major cities of China, Beijing, Shanghai, Guangzhou and Lanzhou (Qin et al., 2014). Ni et al. proposed correlation analysis model of $PM_{2.5}$ to meteorological data, pollutant concentration data and social media data by applying multivariate statistical analysis model and back propagation neural network. Also, ARIMA model was developed to forecast hourly and daily concentrations of $PM_{2.5}$ in Beijing, China (Ni et al., 2017). Wang et al. put forward a novel hybrid-Garch model based on ARIMA and support vector machine (SVM) for predicting hourly $PM_{2.5}$ concentrations in Shenzhen, China (Wang et al., 2017b). In (Zhai and Chen, 2018) an ensemble model based on the stacking of least absolute shrinkage and selection operator (LASSO), Adaboost, XGBoost, MLP optimized by the genetic algorithm and support vector regression (SVR) was proposed to forecast daily average $PM_{2.5}$ concentrations in Beijing, China. In this study special feature extraction procedures, stability feature selection and tree-based feature selection methods were used to choose important variables. There are some other methods in the literature which have been employed in different studies in order to forecast $PM_{2.5}$ concentrations. Markov models (Dong et al., 2009; Sun et al., 2013; Xu and Wang, 2016), Bayesian models (Corani and Scanagatta, 2016; Yu et al., 2016), decision trees and random forests (Kamińska, 2018; Kleine Deters et al., 2017; Sekar et al., 2015; Yu et al., 2016; Zhao and Hasan, 2013) are some of the examples.

Previous studies show that the issue of dealing with long term $PM_{2.5}$ forecasting has not been addressed adequately in the literature. In fact, most of the studies focus on one-step ahead forecasting rather than multi-step ahead forecasting. Multi-step-ahead forecasts of air pollution in urban areas are of greater importance than one-step-ahead ones. Actually, evaluation of the behavior of $PM_{2.5}$ in an urban area for several days ahead, can be used widely by decision makers and people to take some precautionary measures for reducing the harmful effects of $PM_{2.5}$ in advance. This paper addresses multi-step ahead forecasting of fine particulate matter in urban areas. Therefore, in the following, studies on multi-step ahead prediction are reviewed briefly.

1.2.2. Review on multi-step ahead prediction

When stabilisingstablishing a model for multi-step ahead prediction, besides choosing a suitable forecasting method, we need to select a multi-step forecasting strategy. Different strategies have been proposed in the literature for dealing with multi-step ahead prediction. The oldest forecasting strategy is Rec strategy which is often referred to in standard time series text books (Box and Jenkins, 1976). In Rec strategy which is also called iterated or multi-stage strategy, a single model is trained for one step ahead forecasting. After the learning process, H next values are returned by using the trained model H times and the predicted values as inputs of the model to forecast the subsequent points (Taieb et al., 2012). The second strategy, which is called Dir strategy, first proposed by Cox in 1961 (Cox, 1961). In this strategy, in contrast to previous one, H different models are developed for each prediction horizon. A combination of the two mentioned strategies was suggested by Sorjamaa and Lendasse in 2006, which is called DirRec strategy (Sorjamaa and Lendasse, 2006). In this strategy, H forecasting models are estimated for each prediction time step the same as Dir strategy. Also, the forecasts of the previous steps are added to the set of inputs for every horizon like Rec strategy. The three mentioned strategies are classified as single output models. For the first time in 2004 Kline and Zhang proposed a multi-output strategy which they called joint strategy (Kline and Zhang, 2004). Later in 2008 Bontempi introduced it as MIMO strategy (Bontempi, 2008). In this strategy one multi-output model is employed and the forecasts are returned in one step by using this model (Taieb et al., 2012). The DIRMO strategy also called MISMO strategy (Taieb et al., 2009) takes a middle approach of the Dir and MIMO strategies. In this strategy the H step-ahead forecasting task is decomposed into n multiple output forecasting tasks ($n = \frac{H}{s}$), each with an output of size of s ($s\varepsilon\{1.....H\}$). The mathematical details of implementing multi-step ahead strategies will be discussed in Section 2.7.

Different studies have compared between the performance of multi-step ahead approaches. In 2003, Dir and Rec strategies are used for multi-step ahead forecasting of monthly data on various US economic time series. According to the obtained results, the Dir method may or may not improve forecast accuracy upon the Rec method, depending on optimal order selection criteria, forecast periods, forecast horizons and time series to be forecasted (Kang, 2003). In 2009 multi-step ahead forecasting of six different time series (Box and Jenkins, 1976) with artificial neural network (ANN) is considered using Rec and Dir strategies. The results point out the superiority of Dir method against Rec method in this study (Hamzaçebi et al., 2009). Ben Taieb et al. compared between single-output and multi-output strategies for multi-step ahead forecasting (Taieb et al., 2010). The results of the approaches over the series of the NN3 competition (Crone, 2009a) show that multi-output strategies outperform single-output ones. In another study by Bontempi and Ben Taieb, a comparison between Rec, Dir and MIMO strategies has been made which was in favour of MIMO (Bontempi and Taieb, 2011). In (Taieb et al., 2012) existing strategies for multi-step ahead forecasting has been reviewed and compared using a large experimental benchmark (namely the 111 series from the NN5 forecasting competition) (Crone, 2009b), and showed that multi-output strategies are invariably better than single-output approaches.

Multi-step ahead prediction has been used in different applications. In (Niu et al., 2010) a hybrid model based on self-organizing map (SOM) and SVM optimized by particle swarm optimization algorithm is proposed for predicting electricity price. Dir strategy is considered for presenting the forecasts for the next 24 h. Xiong et al. suggested a novel hybrid method based on feedforward neural network for multi-step ahead forecasting of crude oil prices. Rec, Dir and MIMO strategies were examined and compared in this study (Xiong et al., 2013). An et al. employed a MIMO feedforward neural network with empirical mode decomposition for the electricity demand forecasting in New South Wales. In this study the data were half hourly and 1 to 48 steps ahead prediction was considered (An et al., 2013). In another study, Rec, DirRec and DIRMO strategies were adopted to perform multi-step ahead prediction of global solar radiation over the horizon. A hybrid

model combining SOM with optimally pruned extreme learning machine (SOM-OPELM) was proposed as the prediction technique (Wu and Wang, 2016).

As stated earlier, there are a few studies on multi-step ahead prediction of $PM_{2.5}$ in urban areas and to the best of our knowledge none of them examined the performance of different multi-step strategies for $PM_{2.5}$ forecasting. Pérez et al. proposed MLP to predict $PM_{2.5}$ concentrations for 24 h ahead in Santiago de Chile. In this study for every $t\varepsilon\{1,...,24\}$ a different model was developed (Pérez et al., 2000). Feng et al. introduced air mass trajectory analysis and wavelet transformation for improvement of artificial neural network accuracy in $PM_{2.5}$ forecasting. They implemented Rec strategy to present the predictions for two days in advance (Feng et al., 2015). In another study by Pérez and Gramsch, an MLP model was suggested to forecast $PM_{2.5}$ concentrations for the next 21 h, especially high concentrations episodes during winter in city of Santiago. Multiple linear regression (MLR) method was also utilized in order to make a comparison with the results of MLP model. A different network was trained for each time delay for multi-step prediction in this study (Perez and Gramsch, 2016). Biancofiore et al. utilized MLR, ERNN and neural network without the recurrent architecture to forecast daily $PM_{2.5}$ and PM_{10} concentrations for the next three days in the seaside town of Pescara in central Italy. Different meteorological parameters and pollutants at time t was considered as inputs and $PM_{2.5}$ concentrations at time $t + \Delta t$ as output (Biancofiore et al., 2017). In (Ong et al., 2016) a novel pre-training method using auto-encoder is introduced to enhance the performance of deep recurrent neural network for $PM_{2.5}$ prediction. Twelve hours ahead forecasting is considered which is employed by MIMO strategy.

It is worth mentioning that none of the studies in $PM_{2.5}$ forecasting literature addressed the issue of multi-step ahead prediction and the comparison of different strategies in this field. Therefore, this paper tries to fill this gap by comparatively examining the performance of frequently used strategies of multi-step forecasting along with ARIMAX and MLP modelling techniques.

1.3. Aims and innovations

The aims and contributions of this research are listed as below:

- As a major contribution, in this paper long-term prediction of $PM_{2.5}$ is addressed, which can be widely used in decision-making related to control policies and emergency measures such as traffic limitations, school closures, or temporarily shutting down major polluting industrial units at the right time.
- In this study, the performance of five frequently used strategies for multi-step ahead forecasts, namely Rec, Dir, DirRec, MIMO and DIRMO, are examined and compared utilizing ARIMAX and MLP as two prediction tools.
- In the validation phase, the exogenous variables are considered as time series variables and forecasted using ARIMA and MLP models. The forecasted values are used for prediction of the dependent variable in multi-steps ahead of time instead of real values. Therefore, the prediction error related to exogenous variables also affects the final forecasts of $PM_{2.5}$. So, the validation of different prediction methods will be more realistic. Also, in order to examine the effect of using forecasted data instead of real values, two scenarios are defined. In the first scenario real values of exogenous variables and in the second one the prediction values are used. Finally, the results of strategies are compared in these two scenarios.
- For further improvement of the prediction accuracy, different feature selection methods including stepwise regression, LASSO, minimax concave penalty (MCP), and smoothly clipped absolute deviation (SCAD) methods are employed, and the results of multi-step strategies are evaluated using different set of input variables.
- All of the multi-step ahead forecasting strategies together with different prediction tools are applied using data from Mashhad, one of the most polluted cities in Iran.

The rest of the paper is as follows. In Section 2, the related materials and methods are presented. The study region and data sources are described in this section. Also, related approaches used in this study, including different feature selection methods, ARIMAX and MLP modelling techniques, and multi-step ahead forecasting strategies are introduced. In Section 3, the experimental results and discussion are given. Finally, the conclusions are presented in Section 4.

2. Material and methods

In this section, description of study area and the data used are first given. Then, preprocessing steps taken before training the models, are reported. Related approaches used in this study are discussed briefly, including different feature selection methods, ARIMAX and MLP modelling techniques, and multi-step ahead forecasting strategies.

A common notation is used where f and $\hat{f}$ denote the model and estimated model between past and future observations respectively, Y_t and $\hat{Y}_t$ represent the real value and the predicted value of time series at time t respectively, X_t^i denotes ith exogenous variable at time t, d refers to the number of past values used to predict future values, e denotes modelling error, disturbances or noise, h is the prediction time step and H is the maximum time steps to be forecasted.

2.1. Study region

Mashhad is the capital of Razavi Khorasan Province, located in the northeast of Iran. This city is the second most populous city of the country with population of 3,012,090 in urban areas according to 2016 census. Mashhad is an important pilgrimage site in Iran and annually hosts millions of tourists. Mashhad has become one of the most polluted cities of Iran. Considerable population together with millions of tourists travelling to this city, have caused significant increase in transportation, traffic and building different factories which all result in the increase of fossil oil consumption and significant air pollution in this city. In Fig. 1 the geographic location of the study area is shown.

2.2. Data description

The air quality data including NO_2, SO_2, CO, O_3, PM_{10} and $PM_{2.5}$ concentrations were obtained from Environment Pollution Monitoring Centre of Mashhad. The pollutants data are recorded hourly at eleven monitoring stations all over the city. The daily average pollutants of the city are subsequently calculated from hourly readings in different stations. The data cover four years from 21st March 2014 to 20th March

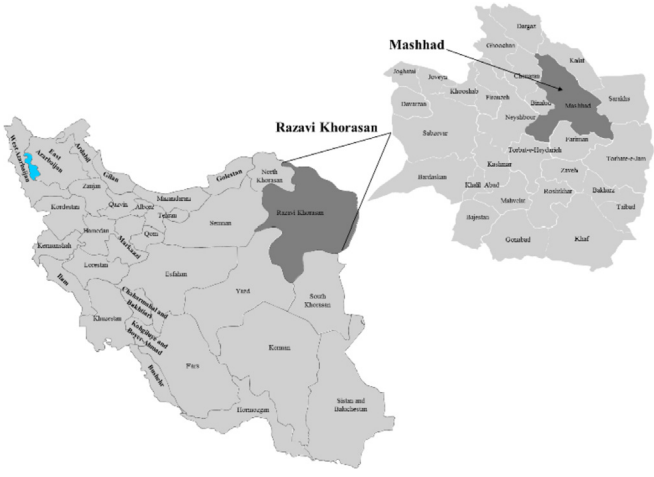

Fig. 1. The geographic location of the study area.

Table 1
Summary statistics of air quality and meteorological parameters available.

Variable	Unit	Max	Min	Avg	Sd	Skewness	Kurtosis	Missed Data
Min temperature	°C	29.20	−13.00	9.77	8.60	−0.12	−1.10	0
Max temperature	°C	41.60	−3.20	23.26	10.57	−0.28	−1.09	0
Min relative humidity	%	100.00	3.00	27.13	21.03	1.38	1.42	0
Max relative humidity	%	100.00	18.00	66.51	25.65	−0.06	−1.46	0
Precipitation	mm	35.90	0.00	0.66	2.74	6.76	59.24	0
Sunshine duration	h	13.60	0.00	8.29	3.90	−0.73	−0.42	0
Evaporation	mm	19.60	0.00	5.20	5.01	0.39	−1.22	0
Max wind speed	m/s	25.00	2.00	6.97	2.71	1.19	3.61	0
Wind direction	°C	360.00	10.00	152.26	97.94	0.83	−0.62	0
Average wind speed	m/s	10.00	0.25	3.24	1.12	0.57	1.49	0
Wet bulb temperature	°C	19.46	−8.45	9.07	5.60	−0.44	−0.76	0
Dry bulb temperature	°C	34.56	−6.66	16.39	9.75	−0.13	−1.24	0
Dew point temperature	°C	15.06	−19.68	1.09	5.27	−0.11	0.13	0
Pressure at QFE station	hpa	920.00	888.13	903.37	4.64	0.42	−0.05	0
Pressure at QFF station	hpa	1041.34	991.91	1014.80	8.37	0.35	−0.56	0
NO_2	ppb	85.18	8.25	32.11	12.88	0.94	0.83	166
SO_2	ppb	48.45	3.34	11.84	4.86	1.81	8.10	53
O_3	ppb	56.06	0.03	18.75	6.56	0.38	0.81	55
CO	ppm	13.94	0.83	1.70	0.55	8.82	176.27	0
$PM_{2.5}$	$\mu g/m^3$	106.03	4.91	30.08	12.79	1.43	3.75	0

2018. Meteorological parameters during the same period of time were also obtained from Razavi Khorasan Meteorological Organization. Meteorological data are recorded every 3-h at synoptic station of Mashhad city and the daily data are calculated from 3-h readings in this station. Descriptive statistics of meteorological and air quality data are shown in Table 1. As can be seen, NO_2, SO_2 and O_3 have 166, 53, and 55 missed data, respectively, which will be addressed in the next subsection. The data from 21st March 2014 to 20th March 2017 are used to fit the models; the remainder of the data from 21st March 2017 to 20th March 2018, is used to test the models.

Besides the parameters summarized in Table 1, temporal variables including day of the week, holiday, and season of the year are considered as additional input variables for the developed models. In Fig. 2, time series of $PM_{2.5}$ over four years is depicted.

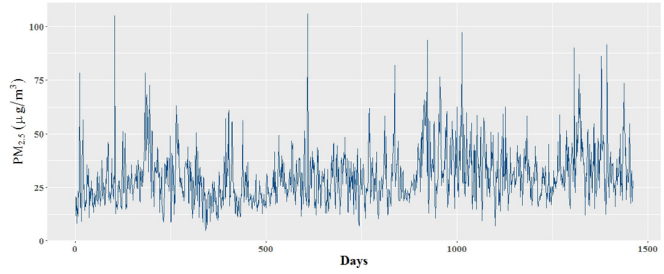

Fig. 2. Time series of $PM_{2.5}$ over four years in Mashhad.

2.3. Handling missed values

In this section, dealing with missing data is addressed. As stated in Section 2.2, measurements of some air quality data are missing. The percentage of missed data for NO_2, SO_2, and O_3 pollutants is 11.36%, 3.63%, and 3.76% respectively. In Fig. 3, the pattern of missing data is shown. According to this figure, in almost 80% of data, none of the samples have missing values. Missing values of these air pollutants, especially NO_2, extend across several days which is common in air quality monitoring data due to malfunctions that affect equipment. In (Junger and de Leon, 2015) a missing value imputation method is introduced which is based on expectation-maximization algorithm and considers the correlations among covariates and the temporal components of the time series. Missing data imputation for NO_2, SO_2, and O_3 was carried out by means of this procedure which is available in the mtsdi library of *R* software.

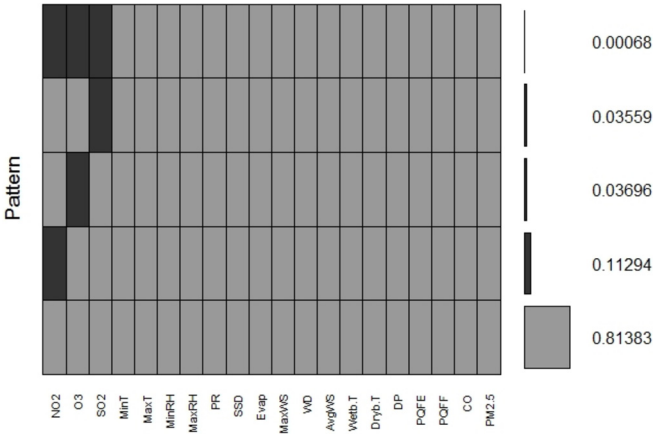

Fig. 3. Pattern of missing values.

2.4. Input selection techniques

Feature selection is an important step in development of prediction models. The aim of feature selection is to remove irrelevant variables thereby building a simpler and more comprehensive models, improving model performance and computational difficulty (Li et al., 2017). In this paper, four feature selection techniques including stepwise regression by BIC criterion, LASSO, SCAD and MCP are implemented.

Stepwise regression is a standard procedure for variable selection which was developed to economize on computational efforts as compared with the various all possible regression procedure (Kutner, 2005). This procedure starts with a model having no predictors. Then, at each step the variable that gives the largest improvement to the fit (based on a specified criterion) is added to the model. However, after adding each predictor, the possibility of removing any variables which no longer provide an improvement, is considered (James et al., 2013). Here, BIC is used as a criterion for adding or removing the variables.

LASSO is another feature selection method which was first introduced in 1996 by Robert Tibshirani (1996). In this method the mean squared residual error is minimized with an upper bound on the sum of the absolute values of coefficients. In order to solve the problem, LASSO applies a regularization process and penalizes the coefficients of

variables. Therefore, this method forces some of the coefficients of variables to zero and is categorized as a shrinkage method. While selecting features for model development, the variables with zero coefficients are dropped. SCAD is another shrinkage method with a non-concave penalty function proposed by Fan and Li (2001). A good penalty function should result in an estimator with unbiasedness, continuity and sparsity properties, which are satisfied by SCAD penalty function. MCP, as another feature selection method, has a penalty function with similar properties as SCAD (Zhang, 2010).

2.5. ARIMAX modelling technique

Air pollution concentrations, commonly measured at equally spaced time intervals, is a kind of time series data and time series forecasting methods can be used for predicting air quality levels. These methods have been used in different studies to predict air pollutant concentrations (Díaz-Robles et al., 2008; Kumar and Jain, 2010). Among different time series modelling tools applied in prediction of $PM_{2.5}$ concentrations, ARIMA model is the most frequently used one (Chen et al., 2017; Liu and Li, 2015; Wang et al., 2017a; Zhou et al., 2014). Autoregressive moving average process of orders p and q abbreviated to ARMA(p, q), is formulated as in equation (1), in which φ_i, μ and θ_i are the parameters to be estimated:

$$Y_t = \mu + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + ... + \varphi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - ... - \theta_q e_{t-q} \quad (1)$$

If the series is non-stationary, by taking the dth difference of the variables, one can make it stationary and use the stationary time series in the above formula. In this case, the model is called ARIMA with orders of p, d and q. In ARIMA modelling technique, information from the past observations of a series is included. But, sometimes it is useful to include other relevant information to improve prediction accuracy. ARIMA models can be extended by combining with regression models to give regression with ARIMA errors (ARIMAX) (Hyndman and Athanasopoulos, 2014). A regression model with ARMA(p, q) errors can be formulated as in equations (2) and (3):

$$Y_t = \beta_1 X_t^1 + \beta_1 X_t^2 + ... + \beta_1 X_t^m + e_t \quad (2)$$

$$e_t = \varphi_1 e_{t-1} + \varphi_2 e_{t-2} + ... + \varphi_p e_{t-p} + z_t - \theta_1 z_{t-1} - ... - \theta_q z_{t-q} \quad (3)$$

In this paper, a regression model with 29 predictors is estimated by means of OLS method. These predictors include 10 temporal variables (6 dummy variables for day of the week, 3 dummy variables for season of the year, 1 dummy variable for holidays), 15 meteorological variables for one previous day (minimum and maximum temperature, minimum and maximum relative humidity, precipitation, sunshine duration, evaporation, maximum and average wind speed, wind direction, wet bulb temperature, dry bulb temperature, dew point, pressure at QFE and QFF stations), and 4 air quality variables for one previous day (NO2, SO2, O3 and CO). The regression model is shown in equation (4):

$$Y_{t+1} = \mu + \sum_{i=1}^{n_0} \beta_i X_{t+1}^i + \sum_{i=n_0+1}^{n_0+n_1} \beta_i X_t^i + e_{t+1} \quad (4)$$

Note that temporal variables are considered with lag 0, while air quality and meteorological parameters are considered with lag 1. In the above formula, n_0 and n_1 indicate the number of variables with lag 0 and lag 1, respectively.

In order to formulate ARIMA model for error terms, the Box-Jenkins methodology is used (Box and Jenkins, 1976). Autocorrelation and partial autocorrelation functions for the error series are depicted in Fig. 4. A single large spike in PACF plot and a decaying pattern in ACF plot are the signs of ARMA(1,0) process (Enders, 2014). In order to check the stationarity of process and determining the order of parameter d, ADF test is employed. The result of ADF test in R software shows that p-value is equal to 0.01, and the null hypothesis is rejected.

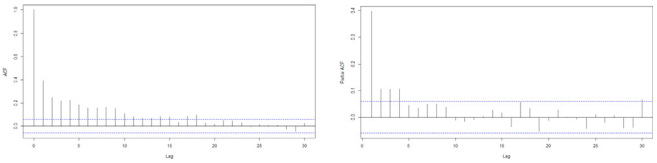

Fig. 4. ACF and PACF plots for error series of regression model.

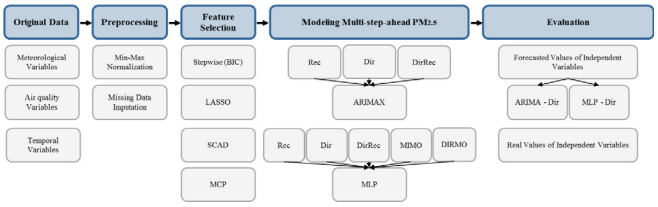

Fig. 5. Summary of the paper.

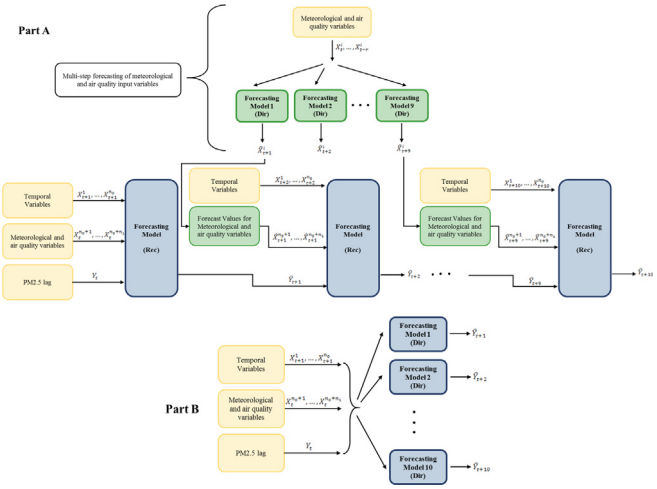

Fig. 6. Architectures of multi-step forecasting approaches for Rec (Part A), and Dir (Part B) strategies in this study.

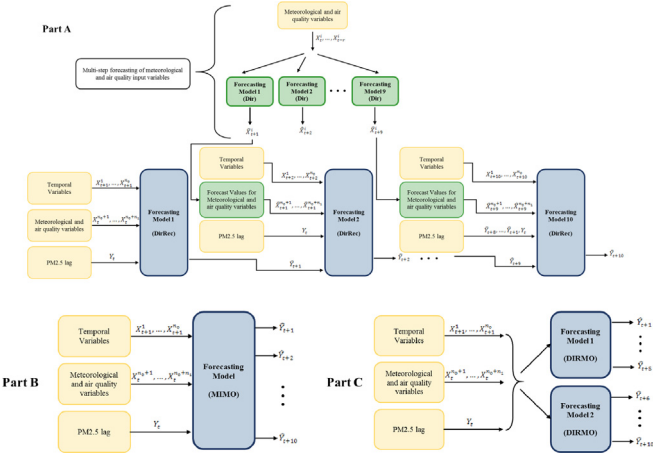

Fig. 7. Architectures of multi-step forecasting approaches for DirRec (Part A), MIMO (Part B) and DIRMO (Part C) in this study.

So, the final model will be ARIMAX(1,0,0).

2.6. MLP modelling technique

ANNs are capable of simulating complex problems with nonlinear relationships between input and output data and found to be a useful

Table 2
Comparison of multi-step strategies for MLP.

H	Metric	Rec (S1)	Rec (S2)	DirRec (S1)	DirRec (S2)	Dir	MIMO	DIRMO
1	RMSE	11.570	11.570	11.570	11.570	11.570	11.299	11.712
	MAE	8.218	8.218	8.218	8.218	8.218	8.422	8.104
	MAPE	27.378	27.378	27.378	27.378	27.378	29.170	24.562
2	RMSE	12.700	12.372	12.051	12.495	13.035	11.789	12.514
	MAE	9.495	9.294	9.044	9.190	9.610	8.843	8.791
	MAPE	32.370	31.001	30.559	31.105	33.244	30.729	27.391
3	RMSE	13.304	13.161	13.179	13.428	12.821	12.607	12.291
	MAE	9.910	9.735	10.028	10.046	9.520	9.138	8.612
	MAPE	33.958	32.851	33.939	34.276	31.604	31.329	26.854
4	RMSE	13.947	13.203	13.983	14.189	13.507	12.312	12.724
	MAE	10.167	9.728	10.449	10.569	9.638	8.913	9.041
	MAPE	34.952	32.713	36.042	35.460	31.667	30.431	27.753
5	RMSE	14.175	13.452	13.761	15.004	14.599	12.227	13.833
	MAE	10.222	9.841	10.556	11.188	10.465	8.869	9.626
	MAPE	35.233	32.571	36.038	37.613	34.205	29.577	29.084
6	RMSE	14.163	13.798	13.162	16.704	13.894	12.766	13.475
	MAE	10.267	10.275	9.640	12.408	9.963	9.408	9.680
	MAPE	35.894	35.140	32.554	43.808	33.805	30.948	31.864
7	RMSE	13.971	15.031	13.262	17.404	13.956	12.503	12.561
	MAE	10.229	11.339	10.170	12.668	10.213	9.167	9.075
	MAPE	35.809	39.273	35.145	45.563	32.746	30.469	29.578
8	RMSE	13.874	15.346	13.863	18.607	15.569	12.631	12.678
	MAE	10.216	11.552	10.443	13.183	11.404	9.281	9.423
	MAPE	35.712	39.895	36.874	47.491	35.793	30.424	30.417
9	RMSE	13.459	15.488	14.919	24.504	13.393	12.788	13.192
	MAE	10.119	11.707	11.233	16.799	9.830	9.468	9.857
	MAPE	35.039	39.876	39.892	62.918	31.419	30.439	31.316
10	RMSE	13.428	15.206	13.215	23.789	13.534	12.651	13.317
	MAE	10.108	11.466	10.540	15.788	10.260	9.245	10.011
	MAPE	34.932	39.020	37.074	58.879	33.440	29.480	31.737
Avg. RMSE		13.459	13.863	13.297	16.769	13.588	12.357	12.830
Avg. MAE		9.895	10.315	10.032	12.006	9.912	9.075	9.222
Avg. MAPE		34.128	34.972	34.550	42.449	32.530	30.300	29.056

Table 3
Comparison of Multi-step strategies for ARIMAX.

h	Metric	Rec (S1)	Rec (S2)	DirRec (S1)	DirRec (S2)	Dir
1	RMSE	10.463	10.463	10.463	10.463	10.463
	MAE	7.198	7.198	7.198	7.198	7.198
	MAPE	23.142	23.142	23.142	23.142	23.142
2	RMSE	11.359	12.277	11.329	12.253	11.675
	MAE	8.030	8.787	8.006	8.784	8.455
	MAPE	25.976	28.731	25.833	28.706	27.914
3	RMSE	11.531	13.183	11.425	13.162	12.165
	MAE	8.040	9.630	7.948	9.639	8.784
	MAPE	25.941	31.660	25.587	31.613	29.295
4	RMSE	11.683	14.763	11.785	15.439	12.273
	MAE	8.152	11.073	8.223	11.553	8.766
	MAPE	25.970	36.644	26.033	38.149	29.148
5	RMSE	11.717	14.882	11.800	15.152	12.271
	MAE	8.182	10.915	8.164	11.010	8.907
	MAPE	26.041	35.145	25.890	35.416	30.068
6	RMSE	11.728	15.298	11.725	15.437	12.590
	MAE	8.190	11.270	8.178	11.310	9.190
	MAPE	26.077	36.484	26.156	36.799	31.113
7	RMSE	11.750	15.650	11.858	15.922	12.839
	MAE	8.210	11.551	8.252	11.603	9.414
	MAPE	26.125	37.332	26.250	37.717	31.269
8	RMSE	11.766	15.529	11.916	15.700	12.788
	MAE	8.222	11.436	8.290	11.441	9.265
	MAPE	26.159	37.017	26.327	37.120	30.202
9	RMSE	11.780	15.880	11.742	15.910	12.618
	MAE	8.235	11.752	8.182	11.699	9.002
	MAPE	26.206	38.042	26.116	38.033	28.595
10	RMSE	11.784	15.913	11.705	16.045	12.530
	MAE	8.230	11.885	8.174	11.935	8.947
	MAPE	26.183	38.607	25.963	38.848	28.524
Avg. RMSE		11.556	14.384	11.575	14.548	12.221
Avg. MAE		8.069	10.550	8.061	10.617	8.793
Avg. MAPE		25.782	34.280	25.730	34.554	28.927

tool in classification, function approximation and prediction problems. This method can be used as an effective alternative to traditional statistical forecasting methods (Gardner and Dorling, 1998) and has been employed in various areas for prediction purposes (Chitsaz et al., 2015; Kara et al., 2011; Panapakidis and Dagoumas, 2016; Yu and Xu, 2014). The basic information processing unit in an ANN is called neuron. Each neuron is consisted of a set of connections with their associated weights, which are combined linearly as $z_j = \sum_{i=1}^{n} w_{ij} u_i + b_j$, where b_j denotes the bias for the neuron j, u_i refers to ith neuron input and w_{ij} represents the weight of the link between neuron input i and neuron j. Then, activation function (φ) is used to produce the neuron output (a_j) using weighted sum of the inputs (z_j) as in equation (5):

$$a_j = \varphi(z_j) \tag{5}$$

MLP is a combination of these neurons with the power of classification and prediction. The MLP network is consisted of an input layer, hidden layer(s) and output layer. All the connections between the layers are of forward kind, which means all the links between neurons of the same level and the signal backward propagation are not allowed.

ANNs has been widely used in air pollution forecasting with different structures and MLP is the most popular structure used for $PM_{2.5}$ forecasting (de Mattos Neto et al., 2014; McKendry, 2002; Pérez et al., 2000; Voukantsis et al., 2011). In this paper, a MLP type of back-propagation neural network is chosen. The parameters including number of hidden layers, number of hidden neurons in each layer, weights and training algorithm, which defines MLP, are optimized using genetic algorithm. The number of hidden layers can be one or two, and the number of hidden neurons in each layer can be up to 20. Four training algorithms are considered including Levenberg–Marquardt (Moré, 1978), Scaled Conjugated Gradient (Møller, 1993), Resilient Back Propagation (Riedmiller and Braun, 1993) and One Step Secant Conjugate Gradient (Battiti, 1992). The initial population is generated randomly. The population size of 50 was selected after examination of

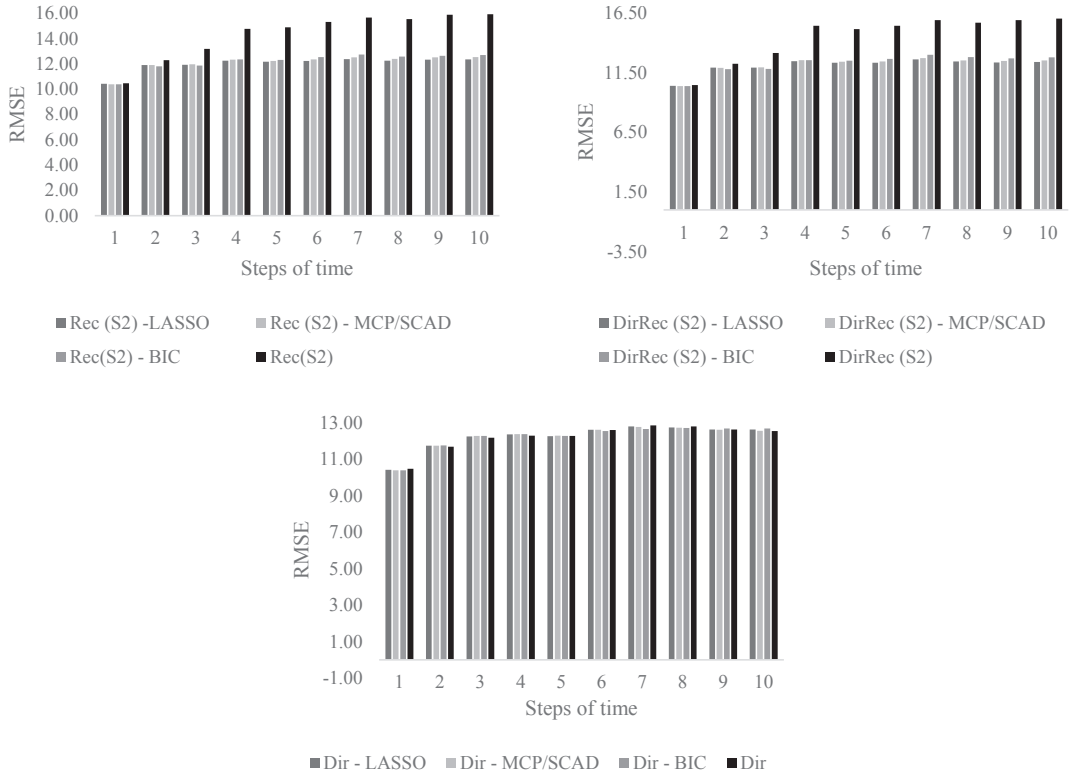

Fig. 8. Results of multi-step strategies for ARIMAX with different input selection methods.

Table 4
Comparison of multi-step ahead forecasting by ARIMAX with different input selection methods.

	Metrics	LASSO	MCP/SCAD	Stepwise BIC	No feature selection
Rec (S2)	RMSE	12.012	12.106	12.182	14.384
	MAE	8.678	8.744	8.809	10.550
	MAPE	28.713	28.900	29.090	34.280
DirRec (S2)	RMSE	12.125	12.198	12.308	14.548
	MAE	8.725	8.769	8.887	10.617
	MAPE	28.862	28.968	29.485	34.554
Dir	RMSE	12.232	12.225	12.223	12.221
	MAE	8.795	8.800	8.689	8.793
	MAPE	29.010	29.037	27.979	28.927

various sizes. Each chromosome of the population consists of the algorithm parameters, which should be tuned. One-point crossover is chosen as the operator for recombination of the solutions. In this procedure, a crossover site is selected randomly over the chromosome length, and all data beyond this point is swapped between the two parent solutions. Also, changing an arbitrary bit in a chromosome from its original state, is chosen as mutation procedure. Finally, the stopping criterion is specified using the number of generations which is set to 30.

Input variables for MLP modelling technique are the same as ARIMAX, including temporal, meteorological and air quality variables. Besides, $PM_{2.5}$ value for the previous day is considered as another input variable for MLP models.

2.7. Multi-step ahead forecasting strategies

Forecasting the next H values of a time series is called multi-step ahead (also called long-term) time series prediction. As mentioned in Section 1.2.2, five common strategies for multi-step ahead forecasting are proposed in the literature. In the following, Rec, Dir, DirRec, MIMO and DIRMO strategies are discussed in detail.

2.7.1. Rec strategy

In Rec strategy, a single model (f) is trained for one step ahead prediction as in equation (6):

$$Y_{t+1} = f(Y_t, ..., Y_{t-d+1}) + e \tag{6}$$

After the learning process, this model is iteratively used for forecasting each horizon h. The forecasts are given by equation (7):

$$\hat{Y}_{t+h} = \begin{cases} \hat{f}(Y_t, ..., Y_{t-d+1}) & if\ h = 1 \\ \hat{f}(\hat{Y}_{t+h-1}, ..., \hat{Y}_{t+1}, Y_t, ..., Y_{t-d+h}) & if\ h\varepsilon\{2, ..., d\} \\ \hat{f}(\hat{Y}_{t+h-1}, ..., \hat{Y}_{t-d+h}) & if\ h\varepsilon\{d+1, ..., H\} \end{cases} \tag{7}$$

It should be noted that only in the first step all of the inputs are from the original dataset, but for the next steps some ($h\varepsilon\{2, ..., d\}$) or all ($h\varepsilon\{d+1, ..., H\}$) of the inputs are the prediction values from previous steps. So, the prediction accuracy is degraded due to accumulation of errors, especially when forecasting horizon h exceeds the number of inputs d (Taieb et al., 2012). Considering exogenous variables in this paper, the forecasts for several steps ahead in Rec strategy will be obtained as in equation (8):

$$\hat{Y}_{t+h} = \begin{cases} \hat{f}\,(X_{t+1}^1, ..., X_{t+1}^{n_0}, X_t^{n_0+1}, ..., X_t^{n_0+n_1}, Y_t) & if\ h = 1 \\ \hat{f}\,(X_{t+h}^1, ..., X_{t+h}^{n_0}, \hat{X}_{t+h-1}^{n_0+1}, ..., \hat{X}_{t+h-1}^{n_0+n_1}, \hat{Y}_{t+h-1}) & if\ h > 1 \end{cases} \tag{8}$$

2.7.2. Dir strategy

In Dir strategy H models are trained independently for each horizon with the same input vector, as shown in equation (9):

$$Y_{t+h} = f_h(Y_t, ..., Y_{t-d+1}) + e, \quad h\varepsilon\{1, ..., H\} \tag{9}$$

After training process, the learned models are used to obtain the forecasts for each time step as shown in equation (10):

$$\hat{Y}_{t+h} = \hat{f}_h(Y_t, ..., Y_{t-d+1}), \quad h\varepsilon\{1, ..., H\} \tag{10}$$

As can be seen, in contrast to Rec strategy no prediction values are used as input of the models. Therefore, there will be no accumulation of

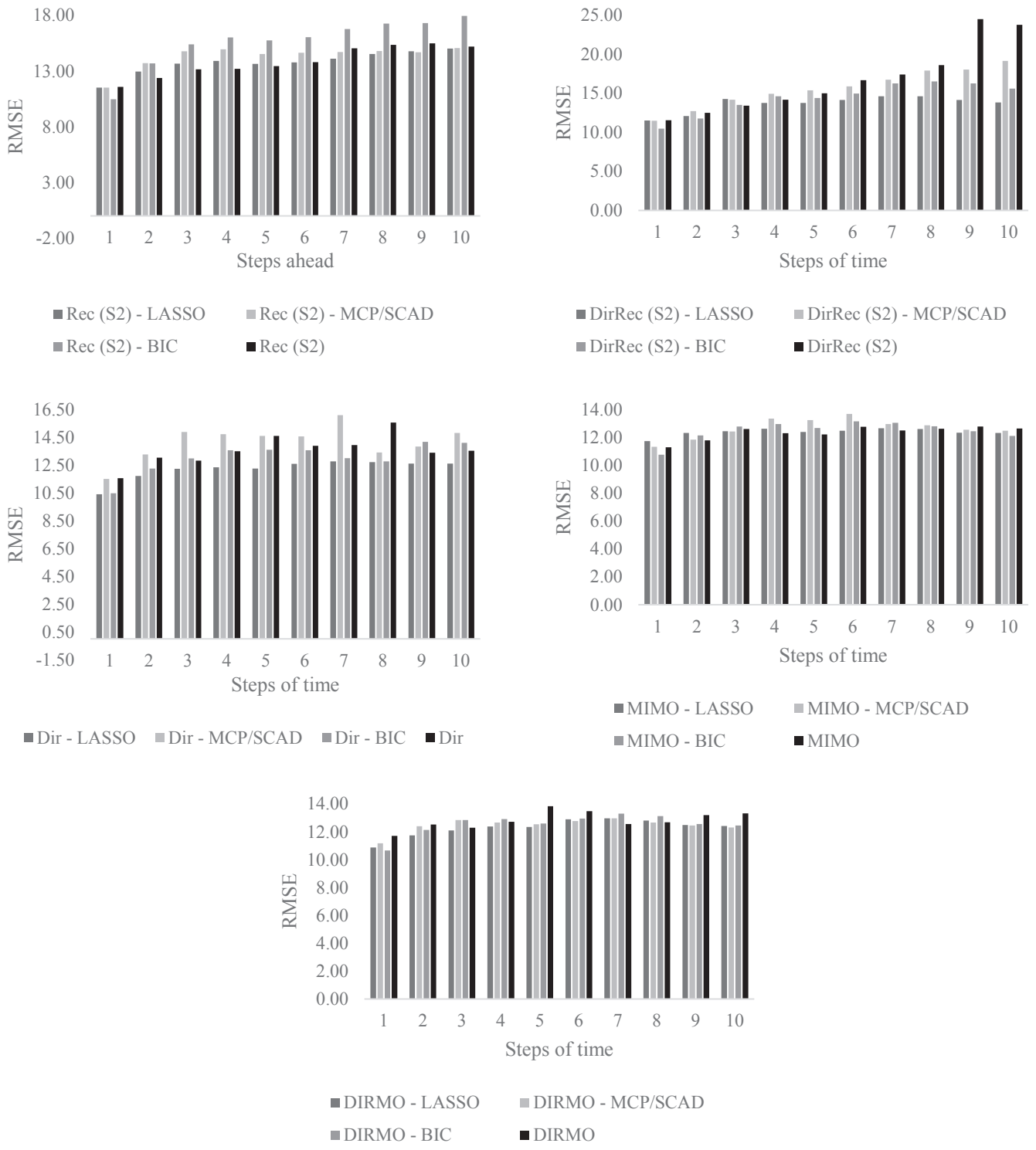

Fig. 9. Results of multi-step strategies for MLP with different input selection methods.

error in this strategy (Taieb et al., 2012). Whereas different models are learned for each horizon independently, a conditional independence of future values is assumed in this strategy, which may affect the forecast accuracy (Bontempi, 2008; Bontempi and Taieb, 2011; Kline and Zhang, 2004). In this paper, the prediction values for each horizon using exogenous variables can be obtained as equation (11):

$$\hat{Y}_{t+h} = \hat{f}_h\left(X_{t+h}^1,\ldots,X_{t+h}^{n_0}, X_t^{n_0+1},\ldots,X_t^{n_0+n_1}, Y_t\right), \quad h\varepsilon\{1,\ldots,H\} \tag{11}$$

2.7.3. DirRec strategy

Two previous strategies are combined in the DirRec strategy. One separate model is trained for each time horizon and the forecasts obtained from previous steps are used as inputs of the next steps. The DirRec strategy can be written as equation (12):

$$Y_{t+h} = f_h(Y_{t+h-1},\ldots,Y_{t-d+1}) + e, \quad h\varepsilon\{1,\ldots,H\} \tag{12}$$

After the learning process, the learned models are used to obtain the forecasts for each time step, as in equation (13):

Table 5
Comparison of multi-step strategies for MLP with different input selection methods.

	Metrics	LASSO	MCP/SCAD	Stepwise BIC	No feature selection
Rec (S2)	RMSE	13.790	14.340	15.666	13.863
	MAE	10.292	11.195	10.884	10.315
	MAPE	33.409	39.838	38.207	34.972
DirRec (S2)	RMSE	13.687	15.653	14.451	16.769
	MAE	10.194	12.000	11.026	12.006
	MAPE	35.259	42.760	38.442	42.449
Dir	RMSE	14.116	14.176	13.055	13.588
	MAE	10.218	10.492	9.370	9.912
	MAPE	33.838	34.681	30.277	32.530
MIMO	RMSE	12.397	12.682	12.493	12.357
	MAE	9.113	9.341	9.130	9.075
	MAPE	30.521	31.111	30.223	30.300
DIRMO	RMSE	12.298	12.477	12.550	12.830
	MAE	9.028	9.197	9.070	9.222
	MAPE	30.315	31.039	29.843	29.056

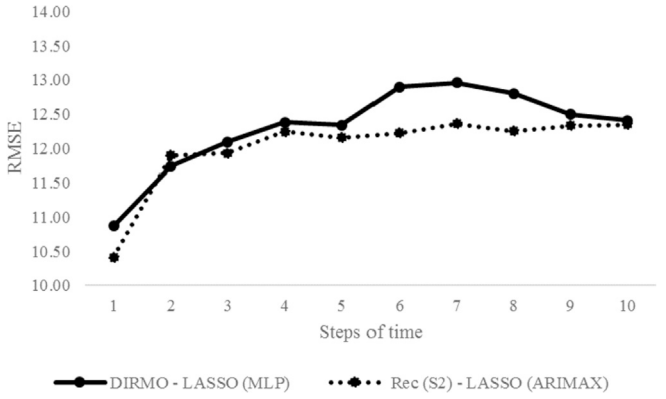

Fig. 10. Comparison of best results of multi-step strategies for MLP and ARIMAX.

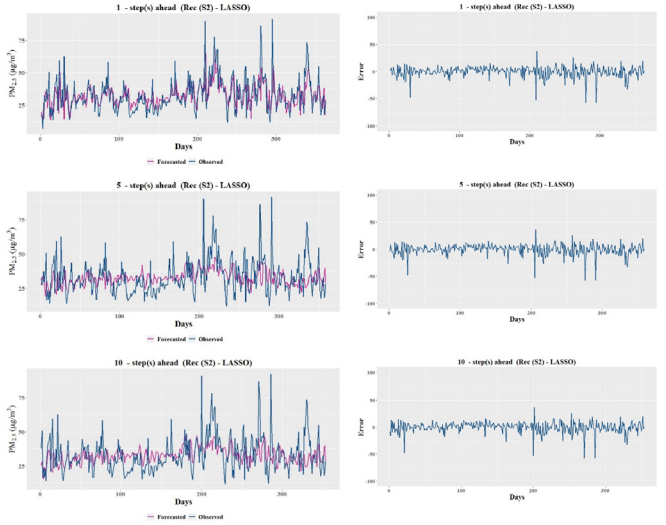

Fig. 11. Time series plots between observed and predicted $PM_{2.5}$ and error series plots for 1, 5 and 10 days ahead.

$$\hat{Y}_{t+h} = \begin{cases} \hat{f}_h(Y_t, \ldots, Y_{t-d+1}) & \text{if } h = 1 \\ \hat{f}_h(\hat{Y}_{t+h-1}, \ldots, \hat{Y}_{t+1}, Y_t, \ldots, Y_{t-d+1}) & \text{if } h \varepsilon \{2, \ldots, H\} \end{cases} \quad (13)$$

It should be noted that, in this strategy unlike the two previous ones, the number of model inputs increases. So, when the forecasting horizon exceeds the embedding size d, there will be real data as well as predicted values for model inputs, in contrast to Rec strategy in which all of the inputs will be the predicted values from previous steps. Equation (14) shows how to obtain the forecasts for different prediction horizons in DirRec strategy, considering exogenous variables in this paper:

$$\hat{Y}_{t+h} = \begin{cases} \hat{f}_1\left(X_{t+1}^1, \ldots, X_{t+1}^{n_0}, X_t^{n_0+1}, \ldots, X_t^{n_0+n_1}, Y_t\right) & \text{if } h = 1 \\ \hat{f}_h\left(X_{t+h}^1, \ldots, X_{t+h}^{n_0}, \hat{X}_{t+h-1}^{n_0+1}, \ldots, \hat{X}_{t+h-1}^{n_0+n_1}, \hat{Y}_{t+h-1}, \ldots, Y_t\right) & \text{if } h \varepsilon \{2, \ldots, H\} \end{cases} \quad (14)$$

2.7.4. MIMO strategy

Unlike Dir strategy which estimates the future values using H models, in MIMO strategy only one multi-output model is learned from the time series, as in equation (15):

$$(Y_{t+1}, \ldots, Y_{t+H}) = f(Y_t, \ldots, Y_{t-d+1}) + e \quad (15)$$

Therefore, the forecasts for all horizons are returned in one step by using equation (16) (Taieb et al., 2012):

$$(\hat{Y}_{t+1}, \ldots, \hat{Y}_{t+H}) = \hat{f}(Y_t, \ldots, Y_{t-d+1}) \quad (16)$$

In contrast to single output strategies, this strategy considers the existence of stochastic dependencies between future values, which is neglected in single-output strategies. But, in multi-output strategies are taken into account. Actually, conditional independence assumption in Dir strategy, also accumulation of error in Rec strategy are avoided in MIMO strategy. One of the drawbacks of MIMO is limiting all the horizons to be forecasted with the same model structure which results in the reduction of the flexibility of this approach (Bontempi, 2008; Bontempi and Taieb, 2011).

Considering exogenous variables, MIMO strategy will return the forecast for different time steps using equation (17):

$$(\hat{Y}_{t+1}, \ldots, \hat{Y}_{t+H}) = \hat{f}\left(X_{t+1}^1, \ldots, X_{t+1}^{n_0}, X_t^{n_0+1}, \ldots, X_t^{n_0+n_1}, Y_t\right) \quad (17)$$

2.7.5. DIRMO strategy

In DIRMO, the H step-ahead forecasting task is decomposed into n multiple output forecasting tasks ($n = \frac{H}{s}$), each with an output of size s ($s \varepsilon \{1, \ldots, H\}$). Actually, this strategy provides a trade-off between flexibility of the prediction and preserving stochastic dependencies between future values. So, n multiple output model is learned as in equation (18):

$$(Y_{t+(p-1)*s+1}, \ldots, Y_{t+p*s}) = f_p(Y_t, \ldots, Y_{t-d+1}) + e, \; p \in \{1, \ldots, n\} \quad (18)$$

Therefore, the forecasts for different horizons are returned in n steps by using the learned models as in equation (19):

$$(\hat{Y}_{t+(p-1)*s+1}, \ldots, \hat{Y}_{t+p*s}) = \hat{f}_p(Y_t, \ldots, Y_{t-d+1}), \quad p \in \{1, \ldots, n\} \quad (19)$$

Equation (20) shows how to obtain the forecasts for different prediction horizons in this paper, using DIRMO strategy and considering exogenous variables:

$$(\hat{Y}_{t+(p-1)*s+1}, \ldots, \hat{Y}_{t+p*s}) = \hat{f}_p\left(X_{t+1}^1, \ldots, X_{t+1}^{n_0}, X_t^{n_0+1}, \ldots, X_t^{n_0+n_1}, Y_t\right), \quad p \in \{1, \ldots, n\} \quad (20)$$

The overall structure of the materials and methods applied in this paper is shown in Fig. 5.

3. Results and discussion

In this section, computational results of the developed models are presented and compared with each other, conducting two experiments. In Experiment I, forecasted exogenous variables and real values of them are used in validation phase and compared with each other. In Experiment II, implementing different feature selection techniques for

improving the prediction results are examined and discussed.

The forecasting performance of the strategies is measured using three metrics, including root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). These performance measures are commonly used in the literature of time series forecasting and are defined by equations (21)–(23), respectively. In these formulas, n is the number of observations, Y_i and $\hat{Y}_i$ indicate the real and the predicted value at period i, respectively.

$$RMSE = \sqrt{\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2/n} \tag{21}$$

$$MAE = \sum_{i=1}^{n} |\hat{Y}_i - Y_i|/n \tag{22}$$

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} |\frac{\hat{Y}_i - Y_i}{Y_i}| \tag{23}$$

3.1. Experiment I: using forecasted values of input variables instead of real values

When evaluating the developed models in Rec and DirRec strategies, predicted values from previous steps need to be used. There is a similar issue relating to the meteorological and other air quality input variables. In other words, from the second step onwards ($h > 1$), we don't have access to prediction values for air quality and meteorological input variables. So, in order to have a more realistic evaluation of the multi-step forecasting methods, predicted values of input variables must be used instead of real ones. Therefore, two different scenarios are considered. In the first scenario (S1), real values for these variables are used in the test phase. In the second scenario (S2), the variables are predicted, and the predicted values are employed in the test phase. So in S2, the meteorological and air quality variables for the next time steps need to be predicted and for each of the variables we are encountering another multi-step ahead prediction problem. When implementing MLP model, the meteorological and air quality variables for the next steps are forecasted using Dir strategy and MLP as the prediction tool. The number of lags as input nodes, number of hidden layers, number of units in each hidden layer, weights, and the training algorithm are optimized for each of the input variables using genetic algorithm. Also, when using ARIMAX modelling technique, the multi-step forecasting of input variables is obtained employing Dir strategy and ARIMA as the prediction tool. Using ACF and PACF plots, the orders of ARIMA is determined for each of the variables. In Figs. 6 and 7, the structures of created models are shown. For Rec and DirRec strategies, S2 is considered in these figures.

In Tables 2 and 3, the performance of multi-step strategies for MLP and ARIMAX are reported. In the first column prediction horizon is given. From the third column to the last one the average values of RMSE, MAE and MAPE metrics for each of the strategies in each time step is shown. It should be noted that for ARIMAX only single-output strategies can be employed, but for MLP all five strategies including single-output and multi-output strategies are implemented. According to Table 2, the following observations can be deducted:

- The multi-output strategies (MIMO and DIRMO) outperform the single-output strategies, including Dir, Rec and DirRec in both scenarios. Actually, single-output approaches suffer either from accumulation of errors, or from neglecting the existence of stochastic dependencies between future values, which may result in lower accuracy compared to multi-output strategies.
- In terms of RMSE, MIMO shows a lower error in comparison with all other strategies, even better than Rec and DirRec strategies in S1, which have used real values of exogenous variables. Only in the third step of time, DIRMO has a slightly lower RMSE than MIMO. Considering MAE and MAPE metric, DIRMO shows better performance in initial steps, but in final steps the lower values belong to MIMO strategy. Although, it should be noted that for DIRMO strategy the parameter s has a great impact on the results and it's better to be optimized (Taieb et al., 2012).
- Comparing Rec and DirRec strategies in two scenarios, it is obvious that on average the results of these strategies are worse in S2 because of using predicted values instead of real ones. As can be seen, in final steps the values of RMSE, MAE and MAPE measures become much larger for Rec and DirRec in S2.
- Among the single-output strategies, on average Dir strategy has a smaller RMSE, MAE and MAPE than the Rec and DirRec strategies in S2. Unlike Rec and DirRec, in Dir strategy no predicted value is used as input variable, which may result in better performance compared to Rec and DirRec strategies in S2.
- Comparing Dir strategy with Rec and DirRec approaches in S1, it is observed that on average Dir strategy has a slightly higher MAE and RMSE, but in terms of MAPE this strategy shows lower values. It should be noted that Rec and DirRec methods in S1 use real recent values of meteorological and air quality as input variables which should result in better accuracy compared to Dir strategy, using the same input variables for all horizons of prediction. However, the obtained results show that Dir strategy can be a better choice when there is no access to real values of meteorological and other air quality data, which is the case in long-term predictions.
- DirRec strategy in S2 is the worst strategy overall in terms of RMSE, MAE and MAPE metrics, which is because of accumulation of error caused by input variable predicted values including lags of response variable (which is increased for each time step), meteorological and air quality input variables.

According to Table 3, the following observations can be deducted:

- Rec and DirRec strategies in S2 perform worse than S1, because of using predicted values of meteorological and air quality variables instead of real values. Similar to MLP models, especially in final steps the values of RMSE, MAE and MAPE metrics get much larger for Rec and DirRec in S2. Actually, in final steps the errors for predicted meteorological and air quality variables are accumulated which result in lower accuracy.
- Based on the performed comparison, superiority of Dir method against Rec and DirRec strategies in S2 is pointed out. Considering RMSE, MAE and MAPE measures in all steps of time Dir strategy achieves lower values. This is because of the fact that Rec and DirRec strategies in S2 suffer from accumulation of errors related to predicted values of exogenous variables as well as lagged values of dependent variable.
- Rec and DirRec strategies in S1 perform better than Dir method on average in terms of RMSE, MAE and MAPE. Although the judge between these strategies doesn't seem fair.
- The worst strategy overall is DirRec strategy in S2, similar to MLP results. Since, the number of inputs increases for each time step in DirRec strategy, accumulation of error for lagged values of dependent variable increases in each time step and results in lower accuracy compared to Rec strategy in S2.

3.2. Experiment II: implementing feature selection techniques

In Experiment II different multi-step strategies using feature selection methods are implemented. It should be noted that Rec and DirRec strategies using feature selection methods are reported only for S2. As stated, in S2 the forecasted values are used instead of real values for prediction of the dependent variable in multi-steps ahead of time. So, the validation of different prediction methods will be more realistic. In Fig. 8, the results of multi-step ahead strategies for ARIMAX using different feature selection methods are shown in terms of RMSE, and in

Table 4 the average results based on RMSE, MAE and MAPE metrics are presented. It should be noted that MCP and SCAD have selected the same variables, so the results of them will be the same. It is observed that:

- Input selection methods in Rec and DirRec improves prediction accuracy considering RMSE, MAE and MAPE measures.
- In Rec and DirRec strategies, different input selection techniques perform almost the same. But, as can be seen in Table 4, LASSO has a lower RMSE, MAE and MAPE for both strategies on average.
- According to Fig. 8, for both Rec and DirRec strategies, stepwise BIC has the lowest RMSE in the first steps of time (1–3), and in the final steps (4–10) LASSO performs better in comparison with other feature selection methods.
- In Dir strategy the results don't seem to be improved after employing input selection methods. Considering MAPE and MAE metrics stepwise BIC shows slightly lower values.
- The overall best method for ARIMAX, is Rec strategy with LASSO input selection method.

In Fig. 9, the performance of Dir, MIMO, DIRMO, Rec and DirRec strategies in S2, applying different feature selection methods are shown in terms of RMSE, and in Table 5 the average results are reported considering RMSE, MAE and MAPE metrics. Comparing the results, it is observed that:

- Input selection methods have a great impact on the accuracy of DirRec strategy, especially in final steps and improves the results.
- For the other strategies, implementing feature selection techniques doesn't improve the results much.
- For MIMO and DIRMO strategies, the results of different input selection methods do not differ by much. For DIRMO strategy, LASSO has a lower RMSE and MAE on average, but for MIMO the results obtained without feature selection seem better. Although, for MIMO strategy, LASSO seems to have better performance among different input selection methods.
- For Rec and DirRec strategies, in final steps of time LASSO performs better (the same as ARIMAX). But in the first steps for DirRec, stepwise BIC has the lowest RMSE. In addition, as can be seen in Table 5, LASSO obtains lower RMSE, MAE and MAPE on average for both Rec and DirRec strategies.
- Stepwise BIC has the lowest RMSE measure for Dir strategy in most steps of time, and according to Table 5 performs better on average based on RMSE, MAE and MAPE measures.
- The overall best method for MLP, is DIRMO strategy with LASSO input selection method.

According to the results obtained, the superiority of DIRMO method with LASSO in MLP models and Rec strategy with LASSO in ARIMAX models can be pointed out. Fig. 10 compares these two strategies with the best performance among MLP and ARIMAX models. It is observed that, almost in all steps of time (except the second step), Rec strategy with LASSO has a lower RMSE.

In Fig. 11, time series plots of predicted vs. observed $PM_{2.5}$ daily mean concentrations for 1, 5 and 10 days ahead predictions are depicted. Also, the error series for each of the time steps are shown. Considering Fig. 10 and error time series in Fig. 11, form the second step the accuracy remains almost stable, which is an indication of robustness in long-term predictions. In the first step a better match between predicted and observed values of $PM_{2.5}$ is observed but in 5 and 10 steps ahead, it gets worse and the predictions are not good enough. Actually, the uncertainty and estimation error in the values of model inputs, including the predictions of meteorological and air quality variables besides $PM_{2.5}$ prediction values from the previous steps, result in worse forecasting in long-term.

4. Conclusion

Multi-step-ahead $PM_{2.5}$ forecasting is an interesting topic in the field of air pollution forecasting, with applications in urgent decision-making related to control policies and emergency measures such as traffic limitations, school closures, or temporarily shutting down major polluting industrial units. In this study, the performances of multi-step-ahead forecasting of $PM_{2.5}$ concentrations is evaluated using ARIMAX and MLP prediction techniques. Daily $PM_{2.5}$ forecasts for the next 10 days in Mashhad, one of the most polluted cities in Iran, are provided by means of five leading multi-step ahead techniques including Rec, Dir, DirRec, MIMO and DIRMO strategies. Furthermore, the performance of these methods are examined using different feature selection methodologies. The results show that Rec strategy with LASSO feature selection method in ARIMAX has the best performance on average. Although, in final steps of time the predictions become poor and are not good enough due to the uncertainty and error in inputs of the model. Dealing with this issue and further improvement of the accuracy of predictions is a promising research area which can be considered by the interested researchers. In addition, this study is limited to point forecasting, although providing interval forecasts are of greater value to decision-makers. This subject could also be a promising research point.

Acknowledgements

Special thanks are extended to Environment Pollution Monitoring Centre of Mashhad and Razavi Khorasan Meteorological Organization, for providing data in this research. Also, we would like to thank both reviewers for their insightful comments on the paper, as these comments led us to an improvement of the work.

References

An, N., Zhao, W., Wang, J., Shang, D., Zhao, E., 2013. Using multi-output feedforward neural network with empirical mode decomposition based signal filtering for electricity demand forecasting. Energy 49, 279–288.
Battiti, R., 1992. One step secant conjugate gradient. Neural Comput. 4, 141–166.
Biancofiore, F., et al., 2017. Recursive Neural Network Model for Analysis and Forecast of PM10 and PM2. 5. Atmospheric Pollution Research.
Bontempi, G., 2008. Long term time series prediction with multi-input multi-output local learning. In: Proc. 2nd ESTSP, pp. 145–154.
Bontempi, G., Taieb, S.B., 2011. Conditionally dependent strategies for multiple-step-ahead prediction in local learning. Int. J. Forecast. 27, 689–699.
Box, G.E., Jenkins, G.M., 1976. Time Series Analysis: Forecasting and Control. Holden-Day.
Brook, R.D., et al., 2010. Particulate matter air pollution and cardiovascular disease. Circulation 121, 2331–2378.
Chen, Y., Li, F., Deng, Z., Chen, X., He, J., 2017. PM2. 5 forecasting with hybrid LSE model-based approach. Software Pract. Ex. 47, 379–390.
Chitsaz, H., Amjady, N., Zareipour, H., 2015. Wind power forecast using wavelet neural network trained by improved Clonal selection algorithm. Energy Convers. Manag. 89, 588–598.
Coker, E., et al., 2016. Multi-pollutant exposure profiles associated with term low birth weight in Los Angeles County. Environ. Int. 91, 1–13.
Corani, G., Scanagatta, M., 2016. Air pollution prediction via multi-label classification. Environ. Model. Software 80, 259–264.
Cox, D.R., 1961. Prediction by exponentially weighted moving averages and related methods. J. Roy. Stat. Soc. Ser. B (Methodol.) 414–422.
Crone, S., 2009a. NN3 Forecasting Competition.
Crone, S., 2009b. NN5 Forecasting Competition.
de Mattos Neto, P.S., Madeiro, F., Ferreira, T.A., Cavalcanti, G.D., 2014. Hybrid intelligent system for air quality forecasting using phase adjustment. Eng. Appl. Artif. Intell. 32, 185–191.
Díaz-Robles, L.A., et al., 2008. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: the case of Temuco, Chile. Atmos. Environ. 42, 8331–8340.
Dong, M., et al., 2009. PM 2.5 concentration prediction using hidden semi-Markov model-based times series data mining. Expert Syst. Appl. 36, 9046–9055.
Enders, W., 2014. Applied Econometric Time Series, fourth ed. Wiley.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96, 1348–1360.
Feng, X., et al., 2015. Artificial neural networks forecasting of PM 2.5 pollution using air mass trajectory based geographic model and wavelet transformation. Atmos. Environ. 107, 118–128.
Gardner, M.W., Dorling, S., 1998. Artificial neural networks (the multilayer

perceptron)—a review of applications in the atmospheric sciences. Atmos. Environ. 32, 2627–2636.
Guaita, R., Pichiule, M., Maté, T., Linares, C., Díaz, J., 2011. Short-term impact of particulate matter (PM2. 5) on respiratory mortality in Madrid. Int. J. Environ. Health Res. 21, 260–274.
Hamzaçebi, C., Akay, D., Kutay, F., 2009. Comparison of direct and iterative artificial neural network forecast approaches in multi-periodic time series forecasting. Expert Syst. Appl. 36, 3839–3844.
Hyndman, R.J., Athanasopoulos, G., 2014. Forecasting: Principles and Practice. OTexts.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer.
Jiang, P., Dong, Q., Li, P., 2017. A novel hybrid strategy for PM 2.5 concentration analysis and prediction. J. Environ. Manag. 196, 443–457.
Junger, W., de Leon, A.P., 2015. Imputation of missing data in time series for air pollutants. Atmos. Environ. 102, 96–104.
Kamińska, J.A., 2018. The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: a case study in Wrocław. J. Environ. Manag. 217, 164–174.
Kang, I.-B., 2003. Multi-period forecasting using different models for different horizons: an application to US economic time series data. Int. J. Forecast. 19, 387–400.
Kara, Y., Boyacioglu, M.A., Baykan, Ö.K., 2011. Predicting direction of stock price index movement using artificial neural networks and support vector machines: the sample of the Istanbul Stock Exchange. Expert Syst. Appl. 38, 5311–5319.
Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., Rybarczyk, Y., 2017. Modeling PM2. 5 urban pollution using machine learning and selected meteorological parameters. J. Electr. Comput. Eng. 2017.
Kline, D.M., Zhang, G., 2004. Methods for multi-step time series forecasting with neural networks. In: Neural Networks in Business Forecasting, pp. 226–250.
Kumar, U., Jain, V., 2010. ARIMA forecasting of ambient air pollutants (O3, NO, NO2 and CO). Stoch. Environ. Res. Risk Assess. 24, 751–760.
Kutner, M.H., 2005. Applied Linear Statistical Models. McGraw-Hill Irwin.
Li, J., et al., 2017. Feature selection: a data perspective. ACM Comput. Surv. (CSUR) 50, 94.
Lippmann, M., 2014. Toxicological and epidemiological studies of cardiovascular effects of ambient air fine particulate matter (PM2. 5) and its chemical components: coherence and public health implications. Crit. Rev. Toxicol. 44, 299–347.
Liu, D.-j., Li, L., 2015. Application study of comprehensive forecasting model based on entropy weighting method on trend of PM2. 5 concentration in Guangzhou, China. Int. J. Environ. Res. Publ. Health 12, 7085–7099.
Lv, B., Cobourn, W.G., Bai, Y., 2016. Development of nonlinear empirical models to forecast daily PM 2.5 and ozone levels in three large Chinese cities. Atmos. Environ. 147, 209–223.
McKendry, I.G., 2002. Evaluation of artificial neural networks for fine particulate pollution (PM10 and PM2. 5) forecasting. J. Air Waste Manag. Assoc. 52, 1096–1101.
Møller, M.F., 1993. A scaled conjugate gradient algorithm for fast supervised learning. Neural Network. 6, 525–533.
Moré, J.J., 1978. The Levenberg-marquardt Algorithm: Implementation and Theory. Numerical Analysis. Springer, pp. 105–116.
Ni, X., Huang, H., Du, W., 2017. Relevance analysis and short-term prediction of PM 2.5 concentrations in Beijing based on multi-source data. Atmos. Environ. 150, 146–161.
Niska, H., et al., 2005. Evaluation of an integrated modelling system containing a multilayer perceptron model and the numerical weather prediction model HIRLAM for the forecasting of urban airborne pollutant concentrations. Atmos. Environ. 39, 6524–6536.
Niu, D., Liu, D., Wu, D.D., 2010. A soft computing system for day-ahead electricity price forecasting. Appl. Soft Comput. 10, 868–875.
Ong, B.T., Sugiura, K., Zettsu, K., 2016. Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM2. 5. Neural Comput. Appl. 27, 1553–1566.
Ordieres, J., Vergara, E., Capuz, R., Salazar, R., 2005. Neural network prediction model for fine particulate matter (PM 2.5) on the US–Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua). Environ. Model. Software 20, 547–559.
Panapakidis, I.P., Dagoumas, A.S., 2016. Day-ahead electricity price forecasting via the application of artificial neural network based models. Appl. Energy 172, 132–151.
Pearson, J.F., Bachireddy, C., Shyamprasad, S., Goldfine, A.B., Brownstein, J.S., 2010. Association between fine particulate matter and diabetes prevalence in the US. Diabetes Care 33, 2196–2201.
Perez, P., Gramsch, E., 2016. Forecasting hourly PM2. 5 in Santiago de Chile with emphasis on night episodes. Atmos. Environ. 124, 22–27.
Perez, P., Salini, G., 2008. PM 2.5 forecasting in a large city: comparison of three methods. Atmos. Environ. 42, 8219–8224.
Pérez, P., Trier, A., Reyes, J., 2000. Prediction of PM 2.5 concentrations several hours in advance using neural networks in Santiago, Chile. Atmos. Environ. 34, 1189–1196.
Prakash, A., Kumar, U., Kumar, K., Jain, V., 2011. A wavelet-based neural network model to predict ambient air pollutants' concentration. Environ. Model. Assess. 16, 503–517.
Qin, S., Liu, F., Wang, J., Sun, B., 2014. Analysis and forecasting of the particulate matter (PM) concentration levels over four major cities of China using hybrid models. Atmos. Environ. 98, 665–675.
Riedmiller, M., Braun, H., 1993. A direct adaptive method for faster backpropagation learning: the RPROP algorithm. *Neural Net*works, 1993. In: IEEE International Conference. IEEE, pp. 586–591.
Santibañez, D.A., Ibarra, S., Matus, P., Seguel, R., 2013. A five-year study of particulate matter (PM2. 5) and cerebrovascular diseases. Environ. Pollut. 181, 1–6.
Sekar, C., Gurjar, B., Ojha, C., Goyal, M.K., 2015. Potential assessment of neural network and decision tree algorithms for forecasting ambient PM 2.5 and CO concentrations: case study. J. Hazard. Toxic Radioact. Waste 20, A5015001.
Sorjamaa, A., Lendasse, A., 2006. Time Series Prediction Using Direrec Strategy.
Sun, W., et al., 2013. Prediction of 24-hour-average PM 2.5 concentrations using a hidden Markov model with different emission distributions in Northern California. Sci. Total Environ. 443, 93–103.
Taieb, S.B., Bontempi, G., Atiya, A.F., Sorjamaa, A., 2012. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. Expert Syst. Appl. 39, 7067–7083.
Taieb, S.B., Bontempi, G., Sorjamaa, A., Lendasse, A., 2009. Long-term prediction of time series by combining direct and mimo strategies. In: Neural Networks, 2009. IJCNN 2009. International Joint Conference on, IEEE, pp. 3054–3061.
Taieb, S.B., Sorjamaa, A., Bontempi, G., 2010. Multiple-output modeling for multi-step-ahead time series forecasting. Neurocomputing 73, 1950–1957.
Tecer, L.H., Alagha, O., Karaca, F., Tuncel, G., Eldes, N., 2008. Particulate matter (PM2. 5, PM10-2.5, and PM10) and children's hospital admissions for asthma and respiratory diseases: a bidirectional case-crossover study. J. Toxicol. Environ. Health, Part A 71, 512–520.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. Ser. B (Methodol.) 267–288.
Voukantsis, D., et al., 2011. Intercomparison of air quality data using principal component analysis, and forecasting of PM 10 and PM 2.5 concentrations using artificial neural networks. In: Thessaloniki and Helsinki, Science of the Total Environment, vol. 409. pp. 1266–1276.
Wang, J., Li, H., Lu, H., 2018. Application of a novel early warning system based on fuzzy time series in urban air quality forecasting in China. Appl. Soft Comput.
Wang, J., Zhang, X., Guo, Z., Lu, H., 2017a. Developing an early-warning system for air quality prediction and assessment of cities in China. Expert Syst. Appl. 84, 102–116.
Wang, P., Zhang, H., Qin, Z., Zhang, G., 2017b. A novel hybrid-Garch model based on ARIMA and SVM for PM 2.5 concentrations forecasting. Atmos. Pollut. Res.
Wu, Y., Wang, J., 2016. A novel hybrid model based on artificial neural networks for solar radiation prediction. Renew. Energy 89, 268–284.
Xiong, T., Bao, Y., Hu, Z., 2013. Beyond one-step-ahead forecasting: evaluation of alternative multi-step-ahead forecasting models for crude oil prices. Energy Econ. 40, 405–415.
Xu, M., Wang, Y.-X., 2016. Quantifying PM 2.5 concentrations from multi-weather sensors using hidden Markov models. IEEE Sensor. J. 16, 22–23.
Yu, F., Xu, X., 2014. A short-term load forecasting model of natural gas based on optimized genetic algorithm and improved BP neural network. Appl. Energy 134, 102–113.
Yu, R., Yang, Y., Yang, L., Han, G., Move, O.A., 2016. RAQ–A random forest approach for predicting air quality in urban sensing systems. Sensors 16, 86.
Zhai, B., Chen, J., 2018. Development of a stacked ensemble model for forecasting and analyzing daily average PM 2.5 concentrations in Beijing, China. Sci. Total Environ. 635, 644–658.
Zhang, C.-H., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. 38, 894–942.
Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., Baklanov, A., 2012. Real-time air quality forecasting, part I: history, techniques, and current status. Atmos. Environ. 60, 632–655.
Zhao, Y., Hasan, Y.A., 2013. Fine particulate matter concentration level prediction by using tree-based ensemble classification algorithms. Int. J. Adv. Comput. Sci. Appl. 4.
Zhou, Q., Jiang, H., Wang, J., Zhou, J., 2014. A hybrid model for PM 2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network. Sci. Total Environ. 496, 264–274.