Contents lists available at ScienceDirect

Environmental Pollution

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

Spatial prediction of PM_{10} concentration using machine learning algorithms in Ankara, Turkey[☆]

Aslı Bozdağ [a], Yeşim Dokuz [b], Öznur Begüm Gökçek [c, *]

[a] Faculty of Engineering, Department of Geomatics Engineering, Nigde Omer Halisdemir University, 51240, Nigde, Turkey
[b] Faculty of Engineering, Department of Computer Engineering, Nigde Omer Halisdemir University, 51240, Nigde, Turkey
[c] Faculty of Engineering, Department of Environmental Engineering, Niğde Ömer Halisdemir University, 51240, Nigde, Turkey

ARTICLE INFO

Article history:
Received 18 February 2020
Received in revised form
14 April 2020
Accepted 18 April 2020
Available online 21 April 2020

Keywords:
Air pollution parameter
Machine learning
Spatial distribution
Artificial intelligence
Predictive modeling

ABSTRACT

With the increase in population and industrialization, air pollution has become one of the global problems nowadays. Therefore, air pollutant parameters should be measured at regular intervals, and the necessary measures should be taken by evaluating the results of measurements. In order to prevent air pollution, pollutant parameters must be evaluated within the framework of a model. Recently, in order to obtain objective and more sensitive results with regard to air pollution nowadays, studies, which use machine learning algorithms in artificial intelligence technologies, have been carried out. In this study, PM_{10} concentrations, which are obtained from 7 stations in Ankara province in Turkey, were trained with machine learning algorithms (LASSO, SVR, RF, kNN, xGBoost, ANN). The PM_{10} concentrations of the years 2009−2017 of 6 stations in Ankara were given as input, and the PM_{10} concentrations of the seventh station for the year 2018 were predicted. The model development stage was repeated for each station, and the performance and error rates of the algorithms were determined by comparing the results produced by the algorithms with the actual results. The best results were provided with ANN ($R^2 = 0.58$, RMSE = 20.8, MAE = 14.4). The spatial distribution of the estimated concentration results was provided through Geographic Information System (GIS), and spatial strategies for improving air pollution over land use were established.

1. Introduction

Nowadays, most people live in cities with high population for many reasons, such as work and education. As the population increases in cities, many environmental problems emerge with the impact of the industry and traffic. Air pollution is the main of these problems. Air pollution occurs when the amount and concentration of foreign substances in the air, which adversely affect the life of living creatures, reach the levels higher than they should be. The most important air pollutant parameters are Particulate matter (PM), Sulfur dioxide (SO_2), Carbon monoxide (CO), Carbon dioxide (CO_2), Ozone (O_3), Nitrogen oxides (NO_x), and Hydrocarbons (HC) and these parameters are collected with ambient information systems (Hazlewood and Coyle, 2009). The Particulate matters, which are among the air pollutant parameters in powder form, are called PM_{10} and $PM_{2.5,}$ according to their aerodynamic diameters. The source of these pollutants can usually be anthropogenic activities such as factories, power plants, incineration plants, construction, as well as natural sources such as fire and dust transport. Pollutant sources causing the formation of particulate matter are domestic heating, industrial activities, and traffic(Tosun, 2017). Short-term exposure to particulate matters causes lung diseases, and long-term exposure to low concentrations results in cancer and infant mortality. According to the Air Pollution Control Regulation, the concentration of PM_{10} should not exceed 50 μg/m^3 more than 35 times in a year, and according to the World Health Organization (WHO), the 24-h average should not exceed 50 μg/m^3. In order to prevent air pollution caused by PM_{10} concentration, the air quality parameters constituting the pollution should be measured at certain times from the points representing the region, and necessary measures should be taken according to the results of these measurements (Ayturan, 2019). The results expected from these measurements should be kept under a certain level with new technologies, and healthy predictions should be made for the future. Nowadays, in order to predict air quality, besides analytical

[☆] This paper has been recommended for acceptance by Pavlos Kassomenos
[*] Corresponding author.
E-mail address: begumgokcek@ohu.edu.tr (Ö.B. Gökçek).

https://doi.org/10.1016/j.envpol.2020.114635

methods, approaches concerning machine learning algorithms, which are artificial intelligence technologies, are gradually increasing. With these approaches, studies, which yield objective and more sensitive results with regard to air pollution, can be conducted.

Machine learning is one of the methods of computer engineering domain that model a given problem according to the data obtained related to the problem. In the literature, for the prediction of air quality parameters, studies have been conducted by using various machine learning algorithms. Examples of these algorithms include LASSO Regression (Chu et al., 2019; Son et al., 2018; Xu et al., 2020), Support Vector Machines (SVM) (Fan et al., 2019; Murillo-escobar et al., 2019; Saxena and Shekhawat, 2017; Zhu et al., 2018, 2017), Random Forest (Kaminska, 2018; Rubal, 2018; Sun et al., 2016; Wang et al., 2019), and k-nearest neighbor (kNN) (Fan et al., 2018; Wen et al., 2019) algorithms.

In the studies, in order to interpret the prediction values better, it is essential to conduct regional analyses, to examine the spatial distribution, and to form spatial strategies for air quality. The visualization of the obtained data through mapping and performing analyses on it require a system. Geographic Information Systems are widely used to provide data for modeling studies or visualization of the results obtained from modeling studies (Ataol, 2010; Ertürk et al., 2004). In the literature, it is observed that spatial and temporal changes are revealed by benefiting from maps created with GIS in prediction and modeling studies for air quality (Lim et al., 2019; Reid et al., 2015; Requia et al., 2019).

Unlike the studies in the literature, in this study, the estimated PM_{10} concentrations for the year 2018 were obtained by using machine learning algorithms, and the spatial distribution of these values was examined together with the land use characteristics (population density, income change, natural gas use, transportation, and industrial density) of the region. In the study, PM_{10} concentrations for the years 2009–2017 obtained from 7 stations in Ankara, the capital of Turkey, were used. Using these values, performance metrics of the regional prediction of PM_{10} concentrations for the year 2018 were revealed with machine learning algorithms (LASSO, SVR, RF, kNN, xGBoost, ANN). For this purpose, PM_{10} concentrations for the years 2009–2017 of 6 stations in Ankara were firstly given as input by means of the used algorithms, and the PM_{10} concentrations of the seventh station for the year 2018 were predicted. The model development stage was repeated for seven stations, and performances and error rates of the algorithms were determined by comparing the results produced by the algorithms with the actual results. Consequently, spatial strategies were developed according to the spatial distribution of the obtained estimated values over land use for the improvement of air pollution.

This study contributes to the literature by performing regional analysis with machine learning algorithms to determine the areas of land use causing air pollution in high-density cities and to produce sustainable spatial strategies.

1.1. *Literature summary*

There are some factors that affect the success of methods and algorithms used in air quality parameter prediction. These can be listed as the amount of data used and the time when it is obtained, the land use of the investigated region, and the affecting meteorological variables. In the literature, some studies use algorithms for predicting PM_{10} concentration change with high accuracy by using different factors (Choubin et al., 2020; Sharma et al., 2019; Stafoggia et al., 2019; Taşpınar, 2015). A summary of the conducted studies is presented in Table 1.

On the other hand, several studies are performed that use deep learning techniques for time series forecasting ((Nilanjan et al., 2018)(Qiu et al., 2017), including air quality prediction (Athira et al., 2018; Kök et al., 2017; Li et al., 2016).

2. Materials and methods

2.1. *Study area and data collection*

The province of Ankara, which is selected as the application area, is located in the Central Anatolia Region of Turkey. Although there are climate differences across the province due to its large

Table 1
Summary of studies in the literature.

Reference	Used Parameters	Machine Learning Algorithms	Used constraint (data type, duration, data volume, examined region, meteorological variables)
Debry and Mallet (2014)	ozone, nitrogen dioxide and PM_{10}	discounted ridge regression (DRR)	Hourly, daily
Tamas et al. (2016)	O_3, NO_2, PM_{10}	ANN, clustering	*Duration* (24 h)
(García Nieto et al., 2018)	PM_{10}	(Vector autoregressive moving-average (VARMA), autoregressive integrated moving average (ARIMA), multilayer perceptron (MLP) neural networks and support vector machines (SVMs))	*Examined region* (Metropolitan region) *Duration* (Monthly)
Ren et al. (2018)	PM_{10}	Random Forest, Gradient boosting	
Suleiman et al. (2019)	PM_{10} and $PM_{2.5}$	ANN, BRT, and SVM	Meteorological variables (wind velocities, wind direction, solar radiation, relative humidity, and ambient temperature) and the data type (Traffic volume, sound level, and speeds)
Irmak and Aydilek (2019)	(SO_2), (NO_2), (O_3), (CO), and (PM_{10})	Random forest, decision tree, support vector, k-nearest neighbor, linear, artificial neural network, stacking, AdaBoost, gradient boosting and bagging regression	Duration: the hourly mean values of SO_2 and NO_2 Duration: the mean values of CO and O_3 gases for the last 8 h Duration: the mean values of PM_{10} dust value for the last 24 h
Cujia et al. (2019)	PM_{10}	SARIMA model (mathematical model for Seasonal Auto-Regressive Integrated Moving Average)	24-h mean PM_{10} concentration
Stafoggia et al. (2019)	$PM_{2.5}$ and PM_{10}	Random Forest	Monthly $PM_{2.5}$ and PM_{10} concentration
Choubin et al. (2020)	PM_{10}	Random forest, Bagged Classification and Regression Trees (Bagged CART), and Mixture Discriminant Analysis (MDA)	Annual PM_{10} concentration
This Study	PM_{10}	LASSO, SVR, RF, kNN, xGBoost, ANN	24-h PM_{10} concentration

area, generally, terrestrial climate prevails (2038 Ankara Environmental Plan). The population of the province in 2018, of which urbanization rate is gradually increasing due to its being the capital city, is 5,503,985 (*Turkish Statistical Institute, 2019*). In the central districts of the province, there are three organized industrial zones besides small industrial estates. Identifying the environmental problems brought about by the gradually increasing urbanization and industrialization in the province is quite essential for establishing livable and sustainable spatial targets. Within the scope of the study, in order to determine the level of air pollution, which is one of the environmental problems in the city, central districts, and station points, where the urban population is high, were identified. Station points, where data for air pollution prediction are provided, are concentrated in some central districts of the city. Therefore, the application area was restricted to these central districts. The map of the application area is given in Fig. 1.

In order to measure air pollution correctly, the Turkey Ministry of Environment and Urbanization formed the National Air Quality Monitoring Network throughout Turkey. Air pollutant parameters that are monitored from the established air pollution measurement stations can be measured fully automatically. The measurement data collected at the measurement stations are transferred to the Ministry's Environmental Reference Laboratory Data Operation Center over this network (VPN) via GSM Modems and monitored and simultaneously broadcast. In this study, the PM_{10} concentration data of 7 station points in Ankara province for the years 2009–2018 were taken.

2.2. Machine learning algorithms

In this study, LASSO, Support Vector Machines, Random Forest, k-Nearest Neighbor, eXtreme Gradient Boosting algorithms, and Artificial Neural Networks were used. In this section these algorithms are introduced.

2.2.1. LASSO regression algorithm

The LASSO (Least Absolute Shrinkage and Selection Operator) regression algorithm is a statistical regression algorithm proposed to reduce the model complexity of linear regression and to prevent data-dependent over-fitting of the model (Robert, 2019). The LASSO regression algorithm provides the regression model's production of better results by increasing and decreasing the importance of input parameters. In this way, it both over-fitting of the model and makes the parameter selection in itself. The LASSO regression algorithm incorporates a ratio of the absolute value of its coefficients into the optimization process by using the L1 regularization approach. In this way, the effect of the parameters on the result is regulated.

In this study, the alpha parameter, which determines the effect of the L1 regularization approach for the LASSO regression algorithm, was selected to be 0.1, and the coefficients were determined to be positive.

2.2.2. Support vector machine algorithm

The SVM algorithm is a discriminatory classification algorithm that attempts to make classification by producing a line, plane, or hyperplane that separates points at two or more dimensions from each other (Drucker et al., 1997; Vapnik, 1995). The SVM algorithm attempts to determine the distinction between classes as well as possible by trying to find the most appropriate line that will maximize the distance between the points of different classes. The SVM algorithm uses three parameters called the function type, C, and gamma, to make a classification. The function type can be selected as linear, non-linear, polynomial, or radial-based function according to the characteristic of input data. C and gamma parameters are the parameters used to prevent excessive or poor learning of the SVM algorithm.

In this study, the radial-based function was used for the SVM algorithm, and C and gamma parameters were selected to be 100 and 0.01, respectively.

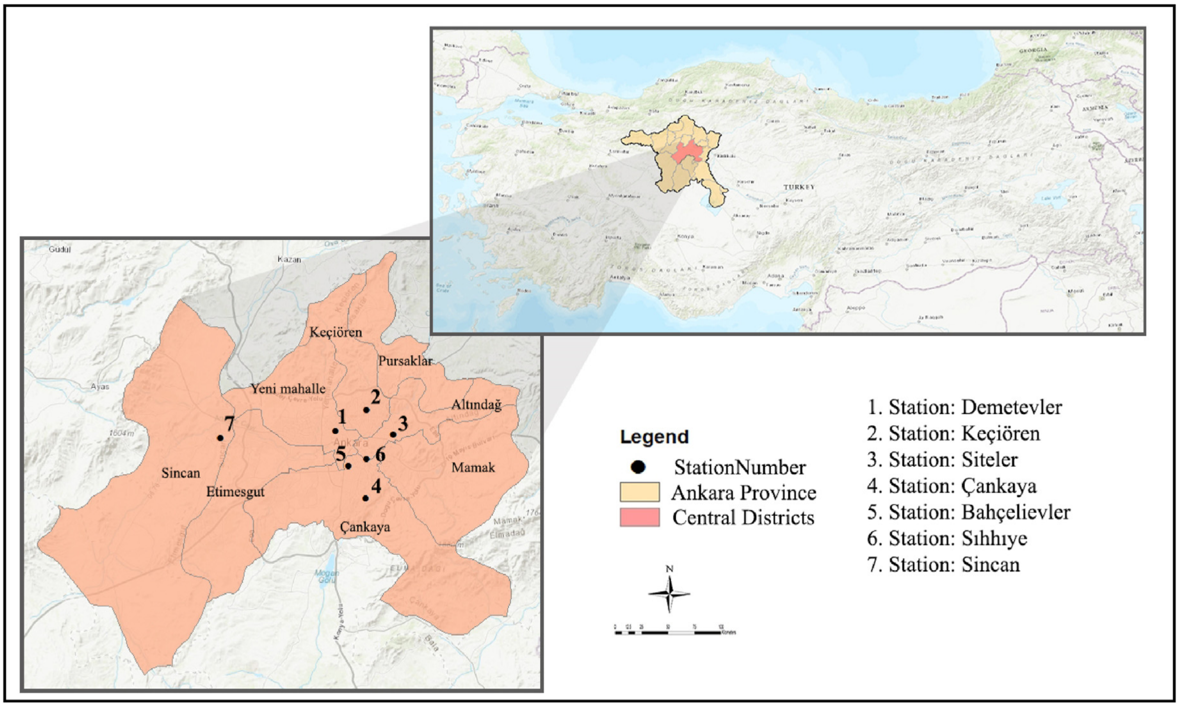

Fig. 1. Application overview.

2.2.3. Random forest algorithm

The Random Forest algorithm is a community-based algorithm that performs classification or regression tasks based on decision trees (Breiman, 2001). Decision Trees are the basis of the Random Forest algorithm. The random forest algorithm selects random parameters from the set of input parameters and creates a large number of decision trees with these parameters. Using the results of these decision trees, the classification output of the algorithm is calculated. In this way, the random forest algorithm produces successful results because parameters independent or irrelevant from each other are evaluated with different decision trees.

In this study, 200 trees were used for the Random Forest algorithm.

2.2.4. K-nearest neighbor algorithm

The k-Nearest Neighbor algorithm is an algorithm that is widely used in the literature and performs classification and regression tasks based on lazy learning (Yao and Ruzzo, 2006). The kNN algorithm determines the k class centers determined during the training phase into account and performs the classification process according to the distance of the test values to these class centers. Different distance metrics such as the Euclidian, Minkowski, and Manhattan distances are used as a criterion of proximity to class centers. The kNN algorithm begins the algorithm by randomly defining k class centers and classifies the training data according to their proximity to these class centers. Afterward, it iteratively shifts the class centers to the middle of the training data and performs reclassification. When satisfactory performance is achieved, the kNN algorithm produces the classification model.

In this study, the number of neighbors for the kNN algorithm was determined to be 20.

2.2.5. eXtreme gradient boosting (XGBoost) algorithm

The XGBoost algorithm is a machine learning algorithm, which is based on the decision tree approach and which uses a gradient boosting system (Chen and Guestrin, 2016). The XGBoost algorithm performs classification and regression tasks with the community-based weak learning approach. The XGBoost algorithm produces more successful results than other incremental algorithms since it creates tree structures with a parallel approach by considering the software and hardware configurations and makes predictions according to these structures.

In this study, the number of trees for the XGBoost algorithm was determined to be 500, the maximum tree depth 3, and the learning rate 0.1.

2.2.6. Artificial neural networks

Artificial neural networks are computational systems developed which are inspired from neurons in the human brain and from the connections established by these neurons (Jain and Mao, 1996). Artificial neural networks learn from the examples shown to them and also incorporate new examples into the learning system. In artificial neural networks, there is one input layer, one or more hidden layers, and one output layer. In the input layer, application input parameters are taken, and an output is produced by making predominantly calculations up to the output layer. According to the difference between the produced output and the actual output, the error is spread back in the network, and this process continues until the desired level of learning of the network is reached. After it is determined that the network has learned sufficiently, prediction outputs can be produced by giving the test inputs to the network.

In this study, three hidden layers, each with three neurons, were used as the artificial neural network model, the number of iterations was 500, and the Limited Memory Broyden−-Fletcher−Goldfarb−Shanno (LBFGS) algorithm was selected for weighted optimization.

2.3. Model training and testing

PM$_{10}$ values collected from seven different regions of Ankara were used in this study. A prediction model based on machine learning algorithms was tried to be established between the change in PM$_{10}$ values in 6 regions and the change in PM$_{10}$ values in the seventh region. The PM$_{10}$ values of the six regions in Ankara between the years 2009−2017 were given as input to these algorithms, and the PM$_{10}$ values between the years 2009−2017 in the seventh region were modeled. Afterward, the PM$_{10}$ values of 6 regions for the year 2018 were given as input, and it was expected to produce output according to the models developed by the algorithms. By comparing the results produced by the algorithms with the actual results, the performances and error rates of the algorithms were determined. The model development phase was repeated for each region, and the performances of the models were analyzed. The algorithmic approach used in this study was presented in Algorithm 1.

Algorithm 1. Regional PM$_{10}$ value prediction algorithm

```
Input
PM$_{10}$Train: PM$_{10}$ values between the years 2009−2017 in 7 regions
PM$_{10}$Test: PM$_{10}$ values in 7 regions for the year 2018 algorithm: the machine
   learning algorithm to be used
Output
PM$_{10}$Pred: PM$_{10}$ prediction values of the year 2018 for the selected region
Algorithm [target_train, reference_train] = target-reference-sec(PM$_{10}$Train)
   [target_test, reference_test] = target-reference-sec(PM$_{10}$Test)
   model = algorithm-train(algorithm, reference_train, target_train)
target_pred = model.prediction(reference_test)
metrics = performance-calculate(target_pred, target_test)
return target_pred
```

As shown in Algorithm 1, the training and test sections of the PM$_{10}$ values in 7 regions and the machine learning algorithm to be used are given as input. As the output, the PM$_{10}$ value in the selected region is predicted, and the performance of the machine learning algorithm is analyzed. In the first and second steps of the algorithm, the target region and the reference regions to be used for

Table 2
PM$_{10}$ concentration prediction test results with machine learning algorithms.

		LASSO	SVR	RF	kNN	xGBoost	ANN
1	**R^2**	0.32	0.34	**0.43**	**0.43**	0.35	0.39
	RMSE	25.6	25.2	**23.5**	23.5	25.0	24.3
	MAE	17.3	16.8	15.8	**15.6**	16.8	16.5
2	**R^2**	0.5	0.56	**0.61**	0.57	0.52	0.46
	RMSE	24.6	22.5	**22.1**	23.4	23.4	24.8
	MAE	15.4	14.6	**14.4**	15.5	16.0	15.7
3	**R^2**	−0.13	−0.06	−0.09	**−0.07**	−0.23	−0.5
	RMSE	30.4	**29.3**	29.8	29.5	31.7	34.9
	MAE	20.8	**20.7**	21.2	20.7	21.8	22.1
4	**R^2**	0.56	0.46	0.49	0.4	0.37	**0.57**
	RMSE	27.1	30.0	29.3	31.7	32.5	**26.8**
	MAE	18.0	18.6	18.0	19.6	18.6	**17.9**
5	**R^2**	0.47	0.42	0.49	0.39	0.43	**0.58**
	RMSE	23.2	24.3	22.8	25.0	24.3	**20.8**
	MAE	15.3	15.4	14.9	16.0	15.5	**14.4**
6	**R^2**	0.17	0.13	0.01	0.01	−0.01	**0.18**
	RMSE	32.7	33.6	35.8	35.8	36.2	**32.6**
	MAE	24.7	**24.3**	27.1	26.7	26.9	24.4
7	**R^2**	**0.21**	0.15	0.17	0.14	0.17	**0.21**
	RMSE	67.4	69.9	69.1	70.3	69.2	**67.4**
	MAE	**21.8**	22.9	23.2	24.4	23.4	**21.8**

the prediction of the target region, from the PM_{10}Train and PM_{10}Test sequences, are turned into separate sequences. In the third step of the algorithm, a model is developed using the machine learning algorithm, reference_train, and target_train sequences. In the fourth step of the algorithm, the model developed is given the reference_test sequence to produce a prediction. In the fifth step of the algorithm, the error rate between the target_pred and the actual target_test sequences produced by the modesl and their performances are calculated. Finally, in the sixth step of the algorithm, the estimated values produced by the model developed by

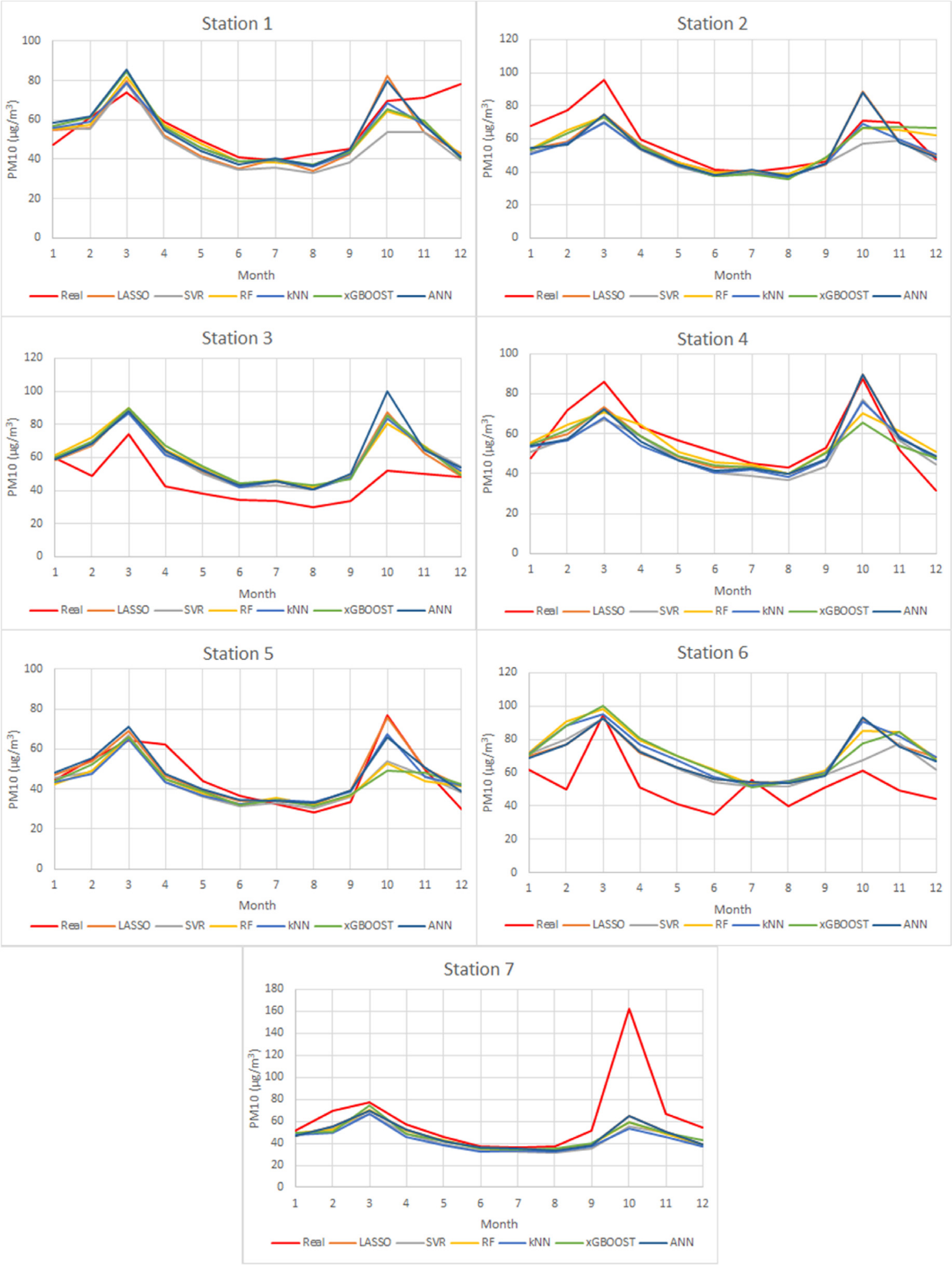

Fig. 2. The monthly comparison of the results of the stations obtained by the machine learning algorithms.

the machine learning algorithm are returned as the result of the algorithm.

2.4. Performance metrics

Three metrics were used to analyze the performance of the machine learning algorithms used in this study. These metrics are R^2, RMSE, and MAE. R^2 is a performance metric that measures how successful the model established with training data on test data is. RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) are error metrics that measure the difference between the actual test values and the test values found by the algorithms. The formulae of R^2, RMSE, and MAE were given in Equations (1)–(3), respectively.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i| \tag{3}$$

3. Results and discussion

3.1. Machine learning results

The results of the prediction algorithm of the PM_{10} concentration obtained by using the machine learning algorithms presented under the methods section are given in Table 2. In the table, for the results of each region, prediction models were established by using PM_{10} concentration values of other regions, and the performance metrics of the prediction of the PM_{10} concentration in the related region for the year 2018 were shown. When the results are examined, the best result is observed to be produced by ANN ($R^2 = 0.58$, RMSE = 20.8, MAE = 14.4) at stations 4, 5, 6, and 7. At station 7, the LASSO algorithm was also able to exhibit the same performance as ANN. At stations 1, 2, and 3, the RF and kNN algorithms were observed to be the algorithms producing the best results. When all the results were evaluated, the highest performance values were observed for station 5, and the lowest performance values were observed for station 3. According to these results, it is observed that using the PM_{10} values of other regions is useful for predicting PM_{10} values in one region in Ankara, except for 3, 6, and 7 station regions.

The monthly comparison of PM_{10} concentrations obtained as a result of the studied algorithms with the actual values for the year 2018 is given in Fig. 2. Accordingly, stations 1 and 5 are the stations that gave results closest to the actual values. At all stations, PM_{10} concentration in the 3rd and 10th months are observed to exceed 50 $\mu g/m^3$, which is determined as the limit value according to the WHO and Air Pollution Control Regulation. It is predicted that these high concentration levels may be caused by dust transport originating from anthropogenic activities in the third month and by heating due to the transition to winter months in the 10th month.

Fig. 3 presents the execution times of machine learning algorithms for PM10 prediction for all of the stations. As can be seen in Fig. 3, the highest execution time is observed for Random Forest Regression algorithm. xGBoost, ANN and SVR are other algorithms that have higher execution times. LASSO and kNN algorithms show the best performance among all algorithms in terms of execution times. In terms of stations-level execution times of algorithms, the algorithms show similar trend with small variations.

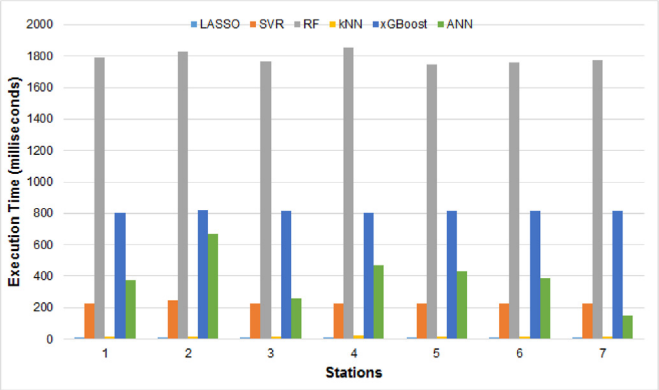

Fig. 3. Execution time comparison of machine learning algorithms.

3.2. Spatial analysis results

The investigation of the spatial distribution of the prediction results of PM_{10} concentrations for the year 2018 made by machine learning algorithms is vital in terms of taking measures for future land use planning scenarios. For this reason, spatial analyses, covering the stations and the central districts in their vicinity, were carried out. The data used in the spatial analysis are the concentration prediction values of each algorithm for the year 2018. The method used is the Kriging function of the ArcGIS 10.6 software.

The spatial distribution map for each algorithm prediction made was created separately (Fig. 4). It was determined that the best result was produced by the RF and kNN algorithms at stations 1, 2, and 3 and by ANN at stations 4, 5, 6, and 7. Accordingly, the spatial distributions formed according to the RF and kNN results for stations 1, 2, and 3, and according to the ANN results for stations 4, 5, 6, and 7 were monitored, and the level of change of air quality was interpreted. It can be said that in the central districts, where stations 5 and 7 are located, the air quality values are below the values determined by the legislation, in the central districts where stations 1 and 2 are located, the air quality values are at the limit of the values determined by the legislation, and in the central districts where stations 3, 4, and 6 are located, the air quality values are above the values determined by the legislation (Fig. 5).

The spatial distribution of the weather prediction values made with the data collected from the stations should be examined in terms of the relationship with the current development of the city. Therefore, the current status of land use characteristics (population density, income status, the amount of green areas, transportation density, natural gas usage, and the location of industrial facilities) related to air pollution in the central districts was analyzed. These analyses were performed with the data received from the Ankara Province Environmental Master Plan Report and from the Turkish Statistical Institute and mapped thematically with the help of ArcGIS 10.6 software (Fig. 4).

In regions where stations 1, 4, and 6 are located, population, transportation density, and income level are high. Furthermore, there is high-density industrial use around station 1 and low-density industrial use around station 6. Accordingly, it can be said that parameters related to air quality rise above the limits determined by legislation depending on the level of urbanization and industrialization in these regions. At the same time, when the land use structure in these station regions is examined, green area usage and natural gas consumption in these regions are found to be much

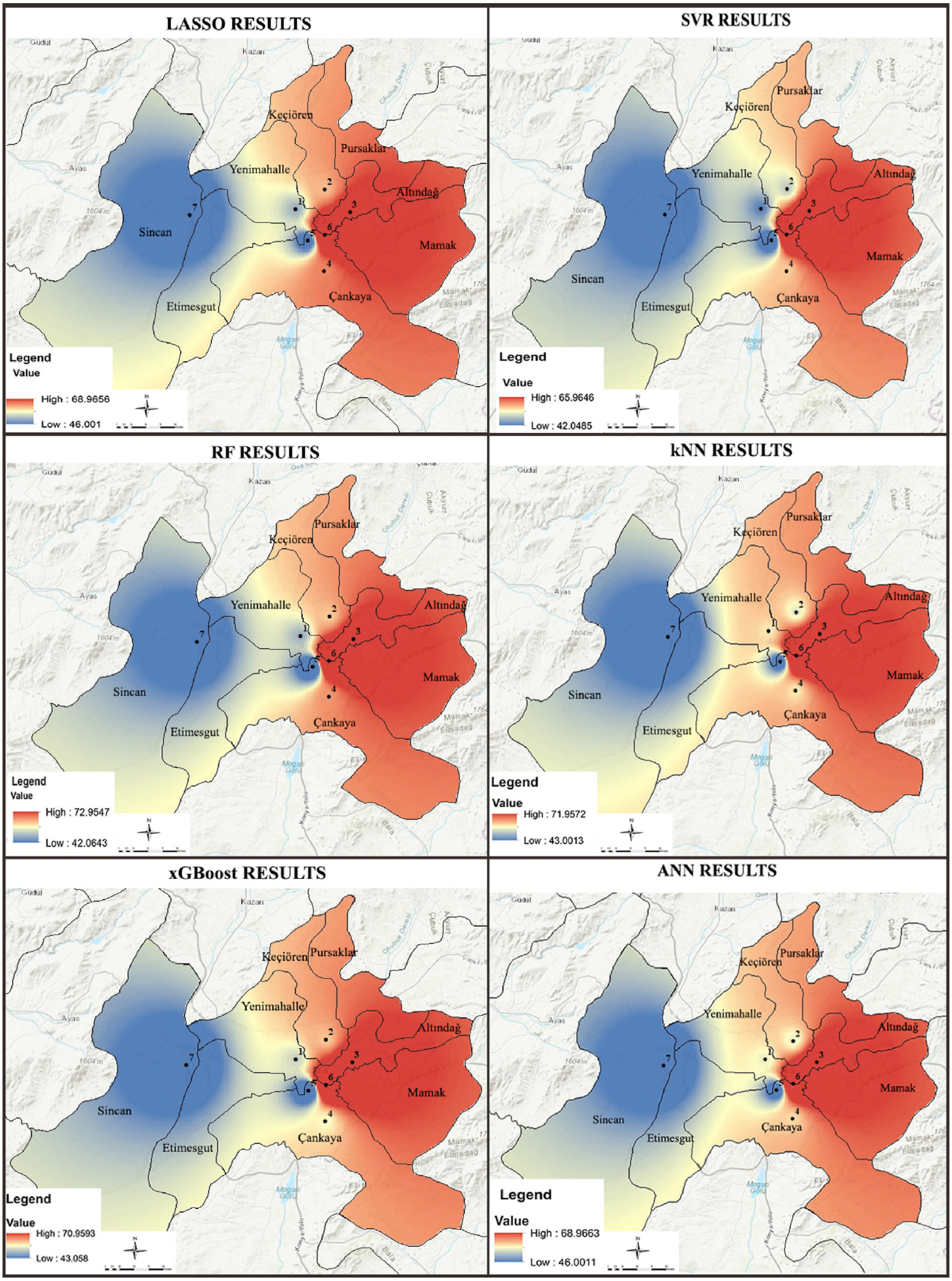

Fig. 4. Spatial distribution maps related to each algorithm prediction.

higher than in other regions. However, this situation shows that the use of green areas here is insufficient in spite of intensive urbanization and industrialization. Accordingly, industries with low density remaining between high housing density around station 6 should be removed from the city center. By creating green area corridors for high-density industrial areas remaining around station 1, air quality may be improved.

Here, the region where station 3 is located is different from the

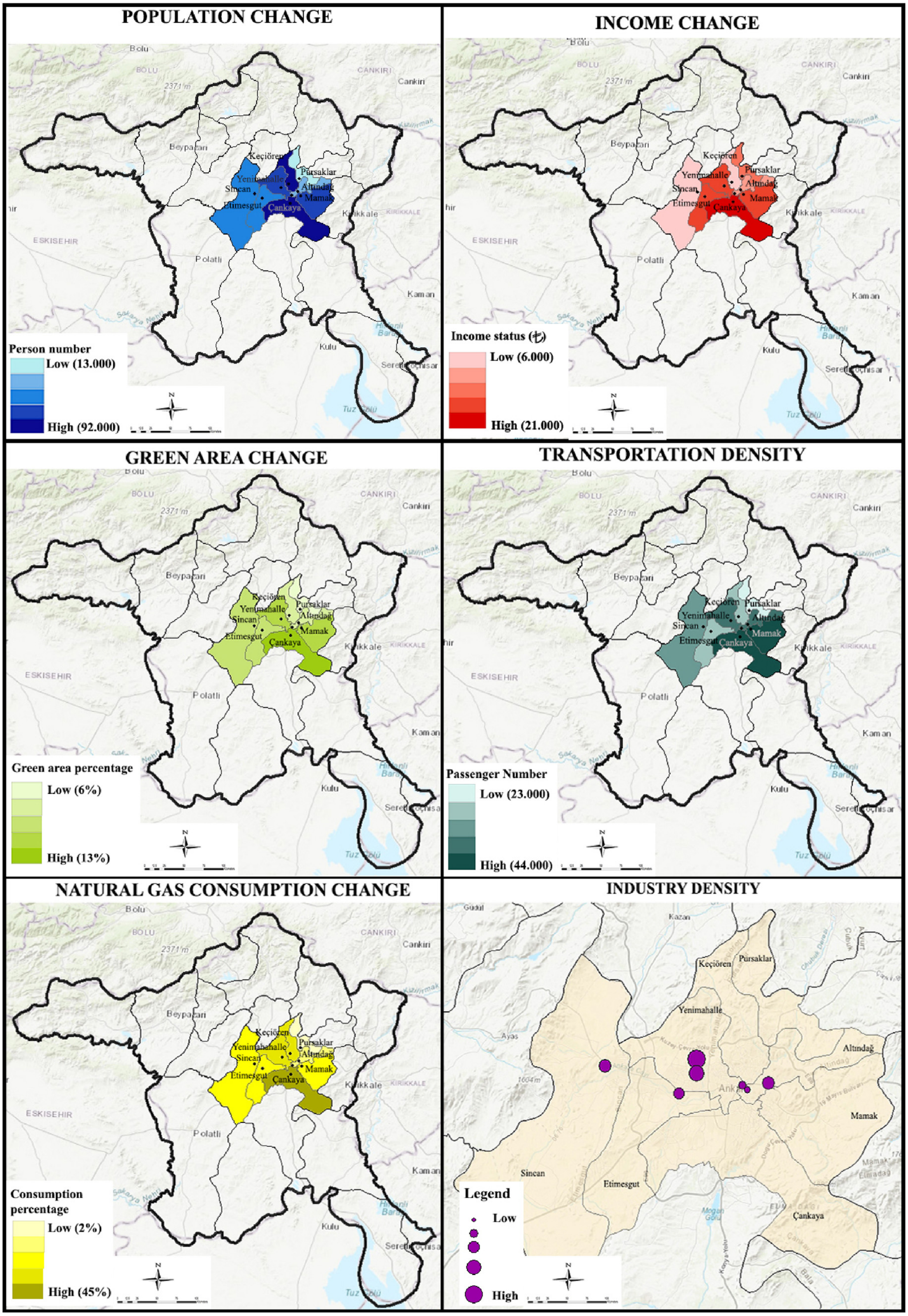

Fig. 5. Land use characteristics related to air pollution in the central districts.

others. Although urbanization in terms of population, transportation density, and income level is low in this region, the air quality concentration value was predicted to be above the limit determined by legislation. The reason for this can be explained by the presence of the woodworking industry in this region. In this region, air quality can be improved by increasing green areas and natural gas consumption.

Although the population density in the region where station 2 is located is high, the other uses examined are at a moderate-level density. Furthermore, the absence of a use that adversely affects air quality, such as industry, caused the air quality level to remain at the limit for the region where station 2 is located. Besides increasing green areas and natural gas consumption, policies supporting public transportation should be followed in this region.

Finally, when the regions where stations 5 and 7 are located are examined, station 5 is observed to be located in a region where housing density is gradually decreasing, which is close to the end of the urban border, and away from industrial use. Station 7 is in the region where industrial use takes place, but urbanization is quite low. Therefore, the prediction of air quality was low in this region.

4. Discussion

In recent years, a lot of research has been done on air pollution, which has direct effects on human health on cities. Researches conducted for predicting air quality by using machine learning techniques have addressed the problem of air quality in different frameworks. Accordingly, it has been determined that the studies regarding the air quality prediction are examined under the following topics:

- measurement of threat levels related to the pollution level of air quality parameters (Chen et al., 2018; Irmak and Aydilek, 2019),
- determination of the method suitable for the number of stations and increasing data volüme (Huang and Kuo, 2018; Ma et al., 2019),
- determination of the effects of meterological variables changing with climatic effects on air quality parameters (Kleine Deters et al., 2017; Li and Zhang, 2019; Masih, 2019; Rybarczyk; Zalakeviciute, 2016; Zhong et al., 2019) and
- investigation of spatial distribution of parameters (Lim et al., 2019; Liu and Sullivan, 2019; Reid et al., 2015; Requia et al., 2019; Zhan et al., 2017)

Unlike the studies in the literature, in this study, the prediction of PM_{10} concentration values for the 2018 year was carried out by performing regional analysis with machine learning techniques. The spatial distribution of the differences created by the estimated concentration values in the regional analysis was evaluated together with the land use structure of the city.

This study is important in terms of determining the sustainable spatial planning strategies for the air quality of the city by examining the results of the regional analysis together with the land use change.

5. Conclusion

In this study, PM_{10} parameter data is used between the dates (01.01.2009–31.12.2017) taken from 7 stations located in the city of Ankara in Turkey. In accordance with the aim of the study, the following studies were carried out:

- The prediction of the regional variation of air pollution with high accuracy according to machine learning algorithms (LASSO, SVM, RF, kNN, xGBoost, ANN),
- The analysis of the spatial distribution according to the predicted values obtained with the help of GIS,
- The establishment of environmental measures and land use strategies for improving air pollution to be taken in the space.

When the performance metrics of all stations were evaluated, the best performance was determined to be obtained with the ANN algorithm. The highest performance value was obtained for station 5 ($R^2 = 0.58$, RMSE = 20.8, MAE = 14.43) among these stations.

In the examination of the spatial distribution of air quality, besides the results of the algorithms, the importance of the effect of land use was revealed. When the spatial distribution of the prediction results is analyzed, it is observed that while the air quality decreases at the station points (3, 4, and 6) where urbanization and industrialization increase, the air quality improves at the station points (1, 2, 5, and 7) where urbanization and industrialization decrease.

The study contributes to the literature by modeling the air quality of high-density cities by machine learning algorithms and providing suggestions for the sustainable planning of land use affecting air quality.

For the future studies, deep learning techniques, such as RNN and LSTM, will be considered for prediction of different air quality indexes.

CRediT authorship contribution statement

Aslı Bozdağ: Software, Writing - review & editing, Investigation. **Yeşim Dokuz:** Visualization, Writing - review & editing, Investigation. **Öznur Begüm Gökçek:** Conceptualization, Writing - review & editing, Investigation.

References

Ataol, M., 2010. Level changes in lake burdur. Coğrafi Bilim. Dergisi. 8, 77–92.

Athira, V., Geetha, P., Vinayakumar, R., Soman, K.P., 2018. DeepAirNet: applying recurrent networks for air quality prediction. Procedia Comput. Sci. 132, 1394–1403. https://doi.org/10.1016/j.procs.2018.05.068.

Ayturan, Y.A., 2019. Estimation of Particulate Matter Concentration in the Air with Deep Learning. KTO Karatay University.

Breiman, L.E.O., 2001. Random forests. Mach. Learn. 45, 5–32.

Chen, G., Li, S., Knibbs, L.D., Hamm, N.A.S., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M.J., Guo, Y., 2018. A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information. Sci. Total Environ. 636, 52–60. https://doi.org/10.1016/j.scitotenv.2018.04.251.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

Choubin, B., Abdolshahnejad, M., Moradi, E., Querol, X., Mosavi, A., Shamshirband, S., Ghamisi, P., 2020. Spatial hazard assessment of the PM10 using machine learning models in Barcelona , Spain. Sci. Total Environ. 701, 134474. https://doi.org/10.1016/j.scitotenv.2019.134474.

Chu, H., Wei, J., Wu, W., 2019. Stream flow prediction using LASSO-FCM-DBN approach based on hydro-meteorological condition classification. https://doi.org/10.1016/j.jhydrol.2019.124253.

Cujia, A., Agudelo-castañeda, D., Pacheco-bustos, C., Teixeira, E.C., 2019. Forecast of PM 10 time-series data : a study case in Caribbean cities. Atmos. Pollut. Res. 10, 2053–2062. https://doi.org/10.1016/j.apr.2019.09.013.

Debry, E., Mallet, V., 2014. Ensemble forecasting with machine learning algorithms for ozone , nitrogen dioxide and PM 10 on the Prev ' Air platform. Atmos. Environ. 91, 71–84. https://doi.org/10.1016/j.atmosenv.2014.03.049.

Drucker, H., Burges, C.J.C., Kaufman, L., Alex, S., Vladimir, V., 1997. Support vector regression machines. Adv. Neural Inf. Process. Syst. 1, 155–161.

Ertürk, A., Ekdal, A., Gurel, M., Yucel, K., Tanik, A., 2004. Use of mathematical models to estimate the effect of nutrient loadings on small streams. Fresenius Environ. Bull. 13, 1361–1370.

Fan, W., Si, F., Ren, S., Yu, C., Cui, Y., Wang, P., 2019. Integration of continuous restricted Boltzmann machine and SVR in NOx emissions prediction of a tangential firing boiler. Chemometr. Intell. Lab. Syst. 195, 103870. https://doi.org/10.1016/j.chemolab.2019.103870.

Fan, Y., Hou, L., Yan, K.X., 2018. On the density estimation of air pollution in Beijing. Econ. Lett. 163, 110–113. https://doi.org/10.1016/j.econlet.2017.12.020.

García Nieto, P.J., Sánchez Lasheras, F., García-Gonzalo, E., de Cos Juez, F.J., 2018.

PM10 concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: a case study. Sci. Total Environ. 621, 753–761. https://doi.org/10.1016/j.scitotenv.2017.11.291.

Hazlewood, W.R., Coyle, L., 2009. On ambient information systems: challenges of design and evaluation. Int. J. Ambient Comput. Intell. 1, 1–12. https://doi.org/10.4018/jaci.2009040101.

Huang, C.J., Kuo, P.H., 2018. A deep cnn-lstm model for particulate matter (Pm2.5) forecasting in smart cities. Sensors (Switzerland) 18. https://doi.org/10.3390/s18072220.

Irmak, M.E., Aydilek, I.B., 2019. Using ensemble regression algorithms for improving the prediction success of air quality index. https://doi.org/10.21541/apjes.478038, 507-514.

Jain, A.K., Mao, J., 1996. Artifical neural networks: a tutorial. Computer (Long. Beach. Calif). 29, 31–44.

Kaminska, 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 217, 164–174. https://doi.org/10.1016/j.jenvman.2018.03.094.

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. https://doi.org/10.1155/2017/5106045, 2017.

Kök, İ., Şimşek, U.M., Özdemir, S., 2017. A deep learning model for air quality prediction in smart cities. In: 2017 IEEE International Conference on Big Data (Big Data), pp. 1983–1990.

Li, X., Peng, L., Hu, Y., Shao, Ji, Chi, T., 2016. Deep learning architecture for air quality predictionsNo Title. Environ. Sci. Pollut. Res. 23, 22408–22417.

Li, X., Zhang, X., 2019. Predicting ground-level PM 2 . 5 concentrations in the Beijing-Tianjin- Hebei region : a hybrid remote sensing and machine learning. Environ. Pollut. 249, 735–749. https://doi.org/10.1016/j.envpol.2019.03.068.

Lim, C.C., Kim, H., Vilcassim, M.J.R., Thurston, G.D., Gordon, T., Chen, L., Lee, K., Heimbinder, M., Kim, S., 2019. Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul , South Korea. Environ. Int. 131, 105022. https://doi.org/10.1016/j.envint.2019.105022.

Liu, Z., Sullivan, C.J., 2019. Prediction of weather induced background radiation fluctuation with recurrent neural networks. Radiat. Phys. Chem. 155, 275–280. https://doi.org/10.1016/j.radphyschem.2018.03.005.

Ma, J., Cheng, J.C.P., Lin, C., Tan, Y., Zhang, J., 2019. Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. Atmos. Environ. 214, 116885. https://doi.org/10.1016/j.atmosenv.2019.116885.

Masih, A., 2019. Machine learning algorithms in air quality modeling. Glob. J. Environ. Sci. Manag. 515–534. https://doi.org/10.22034/gjesm.2019.04.10, 0.

Murillo-escobar, J., Sepulveda-suescun, J.P., Correa, M.A., Orrego-metaute, D., 2019. Urban Climate Forecasting concentrations of air pollutants using support vector regression improved with particle swarm optimization : case study in Aburrá Valley , Colombia. Urban Clim 29, 100473. https://doi.org/10.1016/j.uclim.2019.100473.

Nilanjan, D., Simon, F., Wei, S., Cho, K., 2018. Forecasting energy consumption from smart home sensor network by deep learning. In: International Conference on Smart Trends for Information Technology and Computer Communications, pp. 255–265.

Qiu, X., Ren, Y., Nagaratnam Suganthan, P., Amaratung, G.A.J., 2017. Empirical Mode Decomposition based ensemble deep learning for load demand time series forecasting. Appl. Soft Comput. 54, 246–255.

Reid, C.E., Jerrett, M., Petersen, M.L., Pfister, G.G., Morefield, P.E., Tager, I.B., Raffuse, S.M., Balmes, J.R., 2015. Spatiotemporal prediction of fine particulate matter during the 2008 Northern California wildfires using machine learning. Environ. Sci. Technol. 49, 3887–3896. https://doi.org/10.1021/es505846r.

Ren, Z., Zhu, J., Gao, Y., Yin, Q., Hu, M., Dai, L., Deng, C., Yi, L., Deng, K., Wang, Y., Li, X., Wang, J., 2018. Maternal exposure to ambient PM 10 during pregnancy increases the risk of congenital heart defects : evidence from machine learning models. Sci. Total Environ. 630, 1–10. https://doi.org/10.1016/j.scitotenv.2018.02.181.

Requia, W.J., Coull, B.A., Koutrakis, P., 2019. Evaluation of predictive capabilities of ordinary geostatistical interpolation , hybrid interpolation , and machine learning methods for estimating PM 2 . 5 constituents over space. Environ. Res. 175, 421–433. https://doi.org/10.1016/j.envres.2019.05.025.

Robert, T., 2019. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. . Ser. B (Methodol.) 58, 267–288.

Rubal, D. Kumar, 2018. Evolving differential evolution method with random forest for prediction of air pollution. Procedia Comput. Sci. 132, 824–833. https://doi.org/10.1016/j.procs.2018.05.094.

Rybarczyk, Y., Zalakeviciute, R., 2016. Machine Learning Approach to Forecasting Urban Pollution. 2016 IEEE Ecuador Tech. Chapters Meet. ETCM 2016 1–6. https://doi.org/10.1109/ETCM.2016.7750810.

Saxena, A., Shekhawat, S., 2017. Ambient air quality classification by grey wolf optimizer based support vector machine. J. Environ. Public Health. https://doi.org/10.1155/2017/3131083, 2017.

Sharma, E., Deo, R.C., Prasad, R., Parisi, A.V., 2019. Jo ur l P. Sci. Total Environ. 135934 https://doi.org/10.1016/j.scitotenv.2019.135934.

Son, Y., Osornio-vargas, Á.R., O'Neill, M.S., Hystad, P., Texcalac-sangrador, J.L., Ohman-strickland, P., Meng, Q., Schwander, S., 2018. Land use regression models to assess air pollution exposure in Mexico City using finer spatial and temporal input parameters. Sci. Total Environ. 639, 40–48. https://doi.org/10.1016/j.scitotenv.2018.05.144.

Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., Hoogh, K. De, Donato, F. De, Gariazzo, C., Lyapustin, A., Michelozzi, P., Renzi, M., Scortichini, M., Shtein, A., Viegi, G., Kloog, I., Schwartz, J., 2019. Estimation of daily PM 10 and PM 2 . 5 concentrations in Italy , 2013 – 2015 , using a spatiotemporal land-use random-forest model. Environ. Int. 124, 170–179. https://doi.org/10.1016/j.envint.2019.01.016.

Suleiman, A., Tight, M.R., Quinn, A.D., 2019. Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM 10 and PM 2 . 5). Atmos. Pollut. Res. 10, 134–144. https://doi.org/10.1016/j.apr.2018.07.001.

Sun, H., Gui, D., Yan, B., Liu, Y., Liao, W., Zhu, Y., Lu, C., 2016. Assessing the potential of random forest method for estimating solar radiation using air pollution index. Energy Convers. Manag. 119, 121–129. https://doi.org/10.1016/j.enconman.2016.04.051.

Tamas, W., Notton, G., Paoli, C., Nivet, M.L., Voyant, C., 2016. Hybridization of air quality forecasting models using machine learning and clustering: an original approach to detect pollutant peaks. Aerosol Air Qual. Res. 16, 405–416. https://doi.org/10.4209/aaqr.2015.03.0193.

Taşpınar, F., 2015. Improving artificial neural network model predictions of daily average PM 10 concentrations by applying principle component analysis and implementing seasonal models Improving artificial neural network model predictions of daily average PM 10 concentrat. J. Air Waste Manag. Assoc. 65, 800–809. https://doi.org/10.1080/10962247.2015.1019652.

Tosun, E., 2017. THE EVALUATION OF TURKEY'S AIR QUALITY DATA BETWEEN 2009 AND 2016. Hacettepe University.

Turkish Statistical Institute, 2019.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory, first ed. Springer Science & Business Media.

Wang, Y., Du, Y., Wang, J., Li, T., 2019. Calibration of a low-cost PM 2 . 5 monitor using a random forest model. Environ. Int. 133, 105161. https://doi.org/10.1016/j.envint.2019.105161.

Wen, C., Liu, S., Yao, X., Peng, L., Li, X., Hu, Y., Chi, T., 2019. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. Sci. Total Environ. 654, 1091–1099. https://doi.org/10.1016/j.scitotenv.2018.11.086.

Xu, G., Ren, X., Xiong, K., Li, L., Bi, X., Wu, Q., 2020. Analysis of the driving factors of PM2.5 concentration in the air: a case study of the Yangtze River Delta, China. Ecol. Indicat. 110, 105889. https://doi.org/10.1016/j.ecolind.2019.105889.

Yao, Z., Ruzzo, W.L., 2006. A Regression-Based K Nearest Neighbor Algorithm for Gene Function Prediction from Heterogeneous Data 11, pp. 1–11. https://doi.org/10.1186/1471-2105-7-S1-S11.

Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M.L., Shen, X., Zhu, L., Zhang, M., 2017. Spatiotemporal prediction of continuous daily PM2.5 concentrations across China using a spatially explicit machine learning algorithm. Atmos. Environ. 155, 129–139. https://doi.org/10.1016/j.atmosenv.2017.02.023.

Zhong, J., Zhang, X., Wang, Y., 2019. Relatively weak meteorological feedback effect on PM 2 . 5 mass change in Winter 2017/18 in the Beijing area : observational evidence and machine-learning estimations. Sci. Total Environ. 664, 140–147. https://doi.org/10.1016/j.scitotenv.2019.01.420.

Zhu, S., Lian, X., Liu, H., Hu, J., Wang, Y., Che, J., 2017. Daily air quality index forecasting with hybrid models : a case in. Environ. Pollut. 231, 1232–1244. https://doi.org/10.1016/j.envpol.2017.08.069.

Zhu, S., Lian, X., Wei, L., Che, J., Shen, X., Yang, L., Qiu, X., Liu, X., Gao, W., Ren, X., Li, J., 2018. PM 2 . 5 forecasting using SVR with PSOGSA algorithm based on CEEMD , GRNN and GCA considering meteorological factors. Atmos. Environ. 183, 20–32. https://doi.org/10.1016/j.atmosenv.2018.04.004.