ESKİŞEHİR TECHNICAL UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY
A- APPLIED SCIENCES AND ENGINEERING

2019, 20(3), pp. 365 - 372, DOI: 10.18038/estubtda.545386

AIR QUALITY FORECASTING FOR ALL SEASONS IN LARGE GEOGRAPHICAL AREAS

Efnan ŞORA GÜNAL[1,*]

[1] Department of Computer Engineering, Eskişehir Osmangazi University, Eskişehir, Türkiye

ABSTRACT

Today, air quality monitoring plays a vital role due to increasing number of pollutants that threaten human health. Importance of providing accurate information on air quality for forthcoming times is therefore very high. For this purpose, many studies have been carried out to develop air quality forecasting models. However, most of these studies focus on a particular season and relatively small geographical areas. In this paper, unlike the previous ones, an air quality forecasting model is proposed for all seasons in large geographical areas. Turkiye that is a pretty large country, where there are seven distinct regions with different geographical and meteorological characteristics, is selected to apply the forecasting model. The proposed model categorizes the upcoming 6-hour air quality level as "healthy", "moderate" and "unhealthy". The model utilizes low and high order statistical features extracted from the measurements of air quality monitoring stations covering most parts of the geographical regions of Turkiye. The features are then fed into both linear and non-linear classifiers including artificial neural networks, Fisher's linear discriminant analysis, *k*-nearest neighbor and Bayes classifier. The results of the experimental study indicate that the proposed forecasting model is a promising candidate to predict air quality through all seasons at relatively large geographical areas with varying characteristics.

Keywords: Air pollution, Air quality index, Forecasting model

1. INTRODUCTION

Air is an essential source of life. Consequently, air quality has a direct influence on life quality. Nowadays, many technological advances aiming to facilitate human life unfortunately have negative effect on the air quality. Industrial wastes, vehicle emissions and specific cosmetic products are just a few examples that pollute the air and reduce the air quality as well. Air pollution may cause serious health problems on particularly old people and children. It is therefore one of the major tasks of local authorities to observe air quality and warn people when air pollution is at risky levels. For this purpose, air quality monitoring stations are established in most of the countries all over the world. These stations periodically measure levels of major pollutants such as coarse particulate matter (PM_{10}), fine particulate matter ($PM_{2.5}$), carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen dioxide (NO_2) and ozone (O_3). Air quality index (AQI) values are then calculated using these measurements and announced to people so that they are aware of air quality level and take preventive measures accordingly.

Due to abovementioned influence of air quality on human health, providing precise predictions on air quality for future is a significant research topic in recent years. In the literature, there exist numerous studies related to this topic. A few examples to these studies are as follows: A novel analysis-forecast system is developed for air quality index forecasting by considering eight major cities of China in [1]. A hybrid system for air quality early-warning is proposed and validated for three cities of China in [2]. An attempt is made to estimate the air quality index of an urban station Kolkata, India in [3]. Another study [4] estimates ground-level respirable particulate matter by the combined use of satellite remote sensing and meteorological measurements over Jaipur, semi-arid region in North-western, India. Authors propose an air quality forecasting module for two cities of China in [5]. A hybrid air quality forecasting method is developed for predicting pollutant concentrations in two cities of China, namely Chengdu and Hangzhou in [6]. Trend of air quality index is evaluated for Tehran, Iran in [7]. Air quality

[*] Corresponding Author: esora@ogu.edu.tr

Received: 28.03.2019 Published: 26.09.2019

parameters are predicted using historical observations at Mexico City in [8]. Next day O_3 concentrations are predicted with a forecasting model in the eastern US for a particular month in [9]. Authors evaluate the performance of an air quality forecasting system for summer and winter season at an urban and rural site in [10]. A seasonal $PM_{2.5}$ forecasting model is presented and tested for Santiago, Chile in [11]. A seasonal (April - August) PM_{10} forecasting scheme is proposed and tested at Santiago, Chile again in [12]. Several models are developed to forecast daily averages of $PM_{2.5}$ on the US-Mexico border in [13].

As in the case of above examples, most of the related studies have been previously focused on providing either seasonal and/or local predictions. Nevertheless, there is no significant work that dealt with proposing an air quality forecasting model for all seasons in a relatively large region with different geographical and meteorological characteristics inside. This study, therefore, proposes an air quality forecasting model for all seasons in a pretty large country, Turkiye, on which there are seven distinct regions with different geographical and meteorological characteristics. In the forecasting model, prediction of upcoming 6-hour air quality level is carried out by extracting features from statistical analysis of PM_{10}, SO_2 and meteorological parameters of the last 24 hours. Extracted features are then fed into both linear and non-linear classifiers including artificial neural network (ANN), Fisher's linear discriminant analysis (FLDA), *k*-nearest neighbor (*k*-NN) and Bayes classifier [14]. Air quality levels are classified as Healthy (0 < AQI < 50), Moderate (51 < AQI < 100) and Unhealthy (101 < AQI). Although there are additional levels of AQI above these values such as "Very Unhealthy" and "Hazardous", these situations are not usually encountered in Turkiye. Hence, the classification is carried out by considering the regarding three levels. Experimental data for the regarding regions cover the year of 2018, and is retrieved from National Air Quality Monitoring Network of Turkiye (www.havaizleme.gov.tr). The results of forecasting experiments verify that the proposed model shows great potential to predict air quality through all seasons at relatively large geographical areas with varying characteristics.

Organization of the rest of the paper is as follows: In Section 2, the proposed air quality forecasting model is described in detail. In Section 3, experimental study is explained and the results of the study are provided. Finally, in Section 4, conclusions are given.

2. PROPOSED MODEL

This section introduces the proposed air quality forecasting model. Raw data retrieval process, feature extraction from the retrieved raw data, and classification of the features are described, respectively.

2.1. Data Retrieval

The proposed forecasting model is applied to a large country, Turkiye. The country, which is surrounded by Mediterranean, Aegean, Marmara and Black Sea, spans pretty large vicinity in Asia and Europe [15]. The real area of Turkiye is approximately 814.000 km^2 whereas the projected area is around 780.000 km^2. According to Turkish Statistical Institute (http://www.tuik.gov.tr), overall population of the country is around 82.000.000 as of 2018. The First Geography Congress, which was held at Ankara in 1941, divided the country into seven geographical regions by considering various factors such as climate, topography, vegetation and agriculture. Name and area of these regions are listed in Table 1. Also, borders of the regions are indicated on a map of the country in Fig. 1 [16].

Eastern Anatolian is the largest region of Turkiye; however, it has the lowest population. Average altitude is around 2200 meters. Main geographic features are high mountains and plateaus. Since the region is away from sea and has relatively high altitude, winters are long and snowy whereas summers are short. The region has the lowest average temperature (-25°C) among all seven regions. The summer average is around 20°C. The region has also the largest annual temperature difference. Central Anatolia is the second largest region and has the second highest population. The region has mostly plain geographical characteristic. Average altitude is around 1000 meters. Since the region is away from sea and surrounded by mountains, steppe type climate is observed. Summers are hot and dry, winters are cold and snowy. It rains mostly in springs.

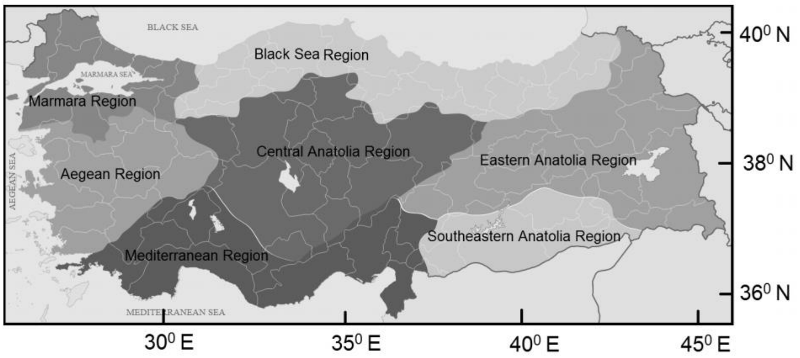
Table 1. Regions of Turkiye

Region	Area (km^2)
Eastern Anatolia	171.000
Central Anatolia	162.000
Black Sea	146.000
Mediterranean	122.000
Aegean	85.000
Marmara	67.000
Southeastern Anatolia	61.000

Black Sea region has a steep and rocky coast. High mountains lie parallel to the coast and limits access to inner parts of the region. In the coastal parts, it rains all seasons due to the influence of Black Sea climate. Nevertheless, inner parts have steppe climate. The region has the lowest annual temperature difference. Mediterranean region is mountainous and rugged. The region has typical Mediterranean climate that is characterized by warm, relatively humid winters and hot, dry summers. Temperatures can go as high as 24°C in winters whereas it is above 30°C in summers.

Aegean region has the longest coastal length among all regions. There exist several mountains lying vertical to the coast. This alignment enables the coastal Mediterranean climate to reach inner parts of the region. Summers are hot and dry whereas winters are warm and rainy due to Mediterranean climate effect. Marmara region contains the largest population although it has one of the smallest areas among seven regions. It is, therefore, the most densely populated region. Average altitude is much lower with respect to other regions and landscapes are mostly plain. Climate of the region is influenced by Black Sea and Mediterranean climate. Southeastern Anatolia is the smallest region of Turkiye. It has also the second lowest population. Savannas and plateaus with moderate heights cover a large area in the region. Summers are very hot, dry and winters are rainy, relatively cold. The lowest monthly temperature is between 1.5°C and 6°C while the highest monthly temperature is around 30°C.

Figure 1. A map of Turkiye



The data required for the forecasting process includes the measurements of PM$_{10}$, SO$_2$, wind speed, wind direction, temperature, relative humidity, and atmospheric pressure values. These values are retrieved from air quality monitoring stations of the cities covering most parts of the abovementioned geographical regions of Turkiye for 2018. At the end of retrieval, 2500 different measurement data are acquired for each of three air quality levels mentioned before.

2.2. Feature Extraction and Classification

After the retrieval of raw data from the regarding air quality monitoring stations for three classes, which are previously mentioned as “healthy”, “moderate”, and “unhealthy”, a statistical feature extraction is carried out. The feature vector contains low and high order statistical information (minimum, maximum, mean, variance, skewness and kurtosis) of PM_{10}, SO_2, wind speed, wind direction, temperature, relative humidity, and atmospheric pressure values for the last 24-hours. The resulting feature vector following the extraction process is 42 dimensional. Thus, 2500 feature vectors are extracted from 2500 measurements so that a dataset with the size of (2500×42) per class is obtained. The layout of a feature vector is given in Table 2.

Table 2. Layout of a feature vector

Feature No	Feature Description
1 – 6	[Min, Max, Mean, Variance, Skewness, Kurtosis] of PM_{10} ($\mu g/m^3$)
7 – 12	[Min, Max, Mean, Variance, Skewness, Kurtosis] of Wind Speed (m/s)
13 – 18	[Min, Max, Mean, Variance, Skewness, Kurtosis] of Wind Direction (Deg)
19 – 24	[Min, Max, Mean, Variance, Skewness, Kurtosis] of Temperature (C°)
25 – 30	[Min, Max, Mean, Variance, Skewness, Kurtosis] of Relative Humidity (%)
31 – 36	[Min, Max, Mean, Variance, Skewness, Kurtosis] of Atmospheric Pressure (mb)
37 – 42	[Min, Max, Mean, Variance, Skewness, Kurtosis] of SO_2 ($\mu g/m^3$)

Here, minimum, and maximum of a parameter indicate the minimum and maximum values of the regarding parameter within the last 24 hours. Mean, which is the first order statistical information, of a parameter is computed as in (1), where *x* indicates the regarding parameter and *N* corresponds to the last 24 hours.

$$m_x = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

Variance, which is the second order statistical information, of a parameter is calculated using (2).

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2 \tag{2}$$

Skewness and kurtosis, the higher order statistical information, are obtained by (3) and (4), respectively.

$$skewness_x = \frac{\sum_{i=1}^{N}\left(x_i - m_x\right)^3}{\left(N-1\right)\sigma_x^3} \tag{3}$$

$$kurtosis_x = \frac{\sum_{i=1}^{N}\left(x_i - m_x\right)^4}{\left(N-1\right)\sigma_x^4} - 3, \tag{4}$$

Following the feature extraction, classification of the extracted features takes place so that forecasting of the air quality level is performed. In our study, celebrated linear and nonlinear classifiers including ANN, FLDA, *k*-NN and Bayes are employed. ANN is already proven to be a successful classifier in most of the related studies mentioned before. However, training time of ANN is considerably long. Therefore, in addition to ANN, efficacies of FLDA, *k*-NN and Bayes classifiers are also investigated due to their relatively shorter training times with respect to ANN.

ANN model used in the study is Levenberg-Marquardt trained feed-forward neural network [14]. This training method is a good exchange between the speed of Newton algorithm and the stability of steepest descent. The second classifier, which is Fisher's Linear Discriminant Analysis (FLDA), utilizes both within- and between-class scatters [14]. In this method, features are projected onto a subspace which minimizes the distance between members of the same class while maximizing the distance to members of the remaining classes. In order to find a transformation for the regarding subspace, the criterion function shown in (5) is to be maximized, where W is the transformation matrix, S_B and S_W denote between-class and within-class scatter, respectively.

$$J(W) = \frac{\left|W^T S_B W\right|}{\left|W^T S_W W\right|} \tag{5}$$

Solution of this maximization problem yields that the columns of W should be constructed from the eigenvectors corresponding to the largest eigenvalues of $S_W^{-1}S_B$. Once the transformation is completed, "nearest mean" decision rule is used for classification in the projected subspace. In the third classifier, *k*-NN, an unknown sample is classified by a majority vote of its neighbors so that the unknown sample is assigned to the class, which is most common amongst its k nearest neighbors [14]. Optimal k value for this study is empirically determined as 21. The fourth and last classifier employed in the study is Bayes. This classifier operates based on the decision rule which minimizes the probability of classification error [14]. A feature vector x is assigned to class c_i among N classes if the a-posteriori error statement in (6) is satisfied. Gaussian (normal) distribution is usually assumed for the probability density functions.

$$p(x \mid c_i)p(c_i) > p(x \mid c_j)p(c_j) \quad j = 1,\ldots,N;\, j \neq i \tag{6}$$

3. EXPERIMENTAL WORK

As mentioned in previous section, a dataset that contains 42-dimensional 2500 feature vectors per class is acquired by a statistical feature extraction process on the measurements from air quality monitoring stations. The dataset is then fed into ANN, FLDA, *k*-NN and Bayes classifiers to measure performance of the proposed forecasting model. Classification is carried out using 5-fold cross validation approach so that overall dataset is fairly evaluated. In this way, confusion matrices of the classifiers that indicate the distribution of the correct predictions, misses, and false alarms are obtained as given in Table 3. Then, average recognition accuracies for each classifier are calculated. The accuracies and corresponding standard deviations are listed in Table 4.

Apparently, ANN classifier is superior to the other classifiers in terms of prediction accuracy. According to Table 3-a, 5918 air quality measurements are predicted correctly out of 7500 measurements for all of healthy, moderate, and unhealthy levels. In other words, overall prediction accuracy achieved by ANN classification is 78.91% as shown in Table 4. FLDA classifier is the runner-up and offers pretty close accuracy to ANN. Bayes and *k*-NN classifiers have however poor prediction performance compared to both ANN and FLDA. As another way of measuring the efficacy of the proposed model, precision rates of the classifiers are also obtained via confusion matrices. Resulting precision rates of each classifier are listed in Table 5. Here, ANN has the highest precision for healthy and moderate levels of air quality whereas FLDA is slightly better than ANN for unhealthy level. Precision of Bayes and *k*-NN classifiers are not good at all compared to ANN and FLDA as in the case of accuracy. Considering both accuracy and precision of the classifiers, ANN is superior to all. In the meantime, FLDA offers a pretty close performance to ANN; moreover, its training time is much shorter. Besides, FLDA is particularly more successful in classifying unhealthy air quality, which has a greater importance.

Table 3. Confusion matrix for a) ANN b) FLDA c) *k*-NN d) Bayes classifier

		Predicted		
		Healthy	**Moderate**	**Unhealthy**
Actual	**Healthy**	2051	395	54
	Moderate	301	1858	341
	Unhealthy	86	405	2009

(a)

		Predicted		
		Healthy	**Moderate**	**Unhealthy**
Actual	**Healthy**	2042	414	44
	Moderate	592	1592	316
	Unhealthy	78	528	1894

(b)

		Predicted		
		Healthy	**Moderate**	**Unhealthy**
Actual	**Healthy**	1543	706	251
	Moderate	1032	905	563
	Unhealthy	332	454	1714

(c)

		Predicted		
		Healthy	**Moderate**	**Unhealthy**
Actual	**Healthy**	660	1731	109
	Moderate	304	2029	167
	Unhealthy	74	1232	1194

(d)

Table 4. Recognition accuracies (%) of the classifiers

	ANN	**FLDA**	***k*-NN**	**Bayes**
Healthy	82.04% σ= 11.23	81.68% σ= 9.75	61.72% σ= 1.79	26.40% σ= 37.54
Moderate	74.32% σ= 2.32	63.68% σ= 2.61	36.20% σ= 2.69	81.16% σ= 20.74
Unhealthy	80.36% σ= 6.83	75.76% σ= 9.35	68.56% σ= 6.22	47.76% σ= 7.12
Average	78.91% σ= 1.54	73.71% σ= 1.33	55.49% σ= 1.72	51.77% σ= 6.34

Table 5. Precision rates of the classifiers

	ANN	**FLDA**	***k*-NN**	**Bayes**
Healthy	0.8412	0.7529	0.3550	0.6358
Moderate	0.6990	0.6282	0.4382	0.4064
Unhealthy	0.8356	0.8402	0.6780	0.8122

It is obvious from the results of experimental study that the proposed forecasting model can predict air quality level of upcoming 6 hours with a pretty satisfactory accuracy and even without limitation to a particular season or geographical area of the country.

4. CONCLUSIONS

In the literature, most of the studies on air quality forecasting have aimed to provide either seasonal or local predictions. However, this study offers a generic forecasting model without dependency to a particular season or region. For this purpose, Turkiye is selected as the test site because the country is large and has seven distinct regions with different geographical and meteorological characteristics. In other words, variations of air quality values at each region are different through all seasons in a year.

Hence, only a generic forecasting model rather than local or seasonal one may provide accurate air quality predictions. The raw data required to develop the model is collected from various cities of the abovementioned regions for one year of period covering four seasons. Thus, the dataset contains overall air quality characteristic of the country through all seasons in a considerably long time window.

The dataset includes the measurements of PM_{10}, SO_2, wind speed, wind direction, temperature, relative humidity and atmospheric pressure values. Recent studies on air quality forecasting have usually preferred these parameters and employ them directly or their low order statistics in the forecasting model. However, in this paper, feature extraction process is carried out by computing both low and high order statistics of these parameters rather than employing them directly in the model. The extracted features are then fed into the classification stage to complete the forecasting operation. Majority of the forecasting studies have previously developed ANN-based prediction models. On the other hand, our study employs both linear and nonlinear classifiers including FLDA, *k*-NN and Bayes in addition to ANN. These classifiers have proven to be very effective in many pattern recognition applications. Also, training time of the regarding classifiers is much shorter than of ANN.

Efficacy of the proposed forecasting model is evaluated using confusion matrix, recognition accuracy and precision rate, which are useful metrics to determine how well the model is suited for the forecasting process. Considering the results obtained, the forecasting model using ANN classifier is superior to all. Meanwhile, FLDA provides a pretty close performance to ANN whereas *k*-NN and Bayes could not reach to a satisfactory forecasting level. Since the training time of FLDA is much shorter than of ANN, air quality forecasting model using FLDA classifier might be used as a good alternative to the model with ANN as well.

The bottom line is that the proposed model shows great potential to predict air quality level successfully through all seasons at relatively large geographical areas with varying characteristics. This capability of the proposed model would reduce the need for designing local and/or seasonal forecasting models. Analysis of feature relevancies and combining classifiers to improve prediction accuracy remain as interesting future works.

ACKNOWLEDGEMENTS

The author would like to thank to Republic of Turkiye, Ministry of Environment and Urbanization for the data retrieved from National Air Quality Monitoring Network (www.havaizleme.gov.tr).

REFERENCES

[1] Li H, Wang J, Li R, Lu H. Novel analysis–forecast system based on multi-objective optimization for air quality index. Journal of Cleaner Production 2019; 208: 1365-1383.

[2] Hao Y, Tian C. The study and application of a novel hybrid system for air quality early-warning. Applied Soft Computing 2019; 74: 729-746.

[3] Chaudhuri S, Chowdhury AR. Air quality index assessment prelude to mitigate environmental hazards. Natural Hazards 2018; 91(1): 1–17.

[4] Soni M, Payra S, Verma S. Particulate matter estimation over a semi arid region Jaipur, India using satellite AOD and meteorological parameters. Atmospheric Pollution Research 2018; 9(5): 949-958.

[5] Yang Z, Wang J. A new air quality monitoring and early warning system: air quality assessment and air pollutant concentration prediction. Environmental Research 2017; 158: 105-117.

[6] Wang J, Zhang X, Guo Z, Lu H. Developing an early-warning system for air quality prediction and assessment of cities in China. Expert Systems with Applications 2017; 84:102-116.

[7] Motesaddi S, Nowrouz P, Alizadeh B, Khalili F, Nemati R. Sulfur dioxide AQI modeling by artificial neural network in Tehran between 2007 and 2013. Environmental Health Engineering and Management Journal 2015; 2(4): 173–178.

[8] Carbajal-Hernández JJ, Sánchez-Fernández LP, Carrasco-Ochoa JA, Martínez-Trinidad JF. Assessment and prediction of air quality using fuzzy logic and autoregressive models. Atmospheric Environment 2012; 60: 37–50.

[9] Sahu SK, Yip S, Holland DM. Improved space–time forecasting of next day ozone concentrations in the eastern US. Atmospheric Environment 2009; 43: 494–501.

[10] Cai C, Hogrefe C, Katsafados P, Kallos G, Beauharnois M, Schwab JJ, Ren X, Brune WH, Zhou X, He Y et al. Performance evaluation of an air quality forecast modeling system for a summer and winter season – photochemical oxidants and their precursors. Atmospheric Environment 2008; 42: 8585–99.

[11] Perez P, Salini G. PM2.5 forecasting in a large city: comparison of three methods. Atmospheric Environment 2008; 42: 8219–24.

[12] Perez P, Reyes J. An integrated neural network model for PM10 forecasting. Atmospheric Environment 2006; 40: 2845–51.

[13] Ordieres JB, Vergara EP, Capuz RS, Salazar RE. Neural network prediction model for fine particulate matter (PM2.5) on the US-Mexico border in El Paso (Texas) and Ciudad Juarez (Chihuahua). Environmental Modelling & Software 2005; 20(5): 547–59.

[14] Duda RO, Hart PE, Stork DG. Pattern Classification: John Wiley & Sons, Inc.; 2001.

[15] De Blij HJ, Downs R. College Atlas of the World: Wiley/National Geographic; 2007.

[16] Ansari K, Corumluoglu O, Sharma S. Numerical simulation of crustal strain in Turkey from continuous GNSS measurements in the interval 2009–2017. Journal of Geodetic Science 2017; 7(1): 113–129.