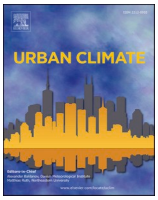
Urban Climate 41 (2022) 101055

Contents lists available at ScienceDirect

Urban Climate

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



Spatio-temporal prediction and factor identification of urban air quality using support vector machine

Chih-Chun Liu, Tzu-Chi Lin, Kuang-Yu Yuan, Pei-Te Chiueh [*]

Graduate Institute of Environmental Engineering, College of Engineering, National Taiwan University, Taipei, Taiwan

ARTICLE INFO

Keywords:
Air quality prediction
Support vector machine
Machine learning
Spatio-temporal features
Geographic information system

ABSTRACT

Accurate air quality prediction can provide better supervision and reference for management policies. Due to difficulties in data acquisition, combined spatio-temporal prediction is still inconclusive. This study utilizes the support vector machine (SVM) method to predict air quality of unknown space and time. Extracted from a geographic information system (GIS), geographic features such as population, land use, economy, pollution sources, and terrain parameters were added to a time series. Temporal prediction was first executed in the reference stations, and the predicted air quality index (AQI) was then used to spatially infer the future AQI of unknown locations. Verification indicated high accuracy for short-term temporal prediction. Various meteorological and climatic effects were observed to be influential in seasonal difference. In the spatial inference stage, urbanization and city types were spatial features that appeared to impact air quality. Agriculture and forest use, transportation use, residential use, and economic factors were clearly correlated to AQIs, whereas population and labor force were not. This study establishes a prediction framework in northern Taipei based on SVM. Other locations can build their own models based on local actual data to achieve better decision-making, urban planning, or other applications.

1. Introduction

Urban air quality has been regarded as a crucial issue for decades due to the rising concerns for human health. Exposure to poor-quality air may cause allergic reactions and even lead to respiratory and circulatory diseases (Dockery et al., 1993; Harrison and Yin, 2000; Hong et al., 2002; Pope III et al., 2002). These diseases can contribute to enormous economic loss due to the demand for medical treatments and the decline in productivity. Therefore, accurately predicting air quality is a development goal for governments and researchers (Feng et al., 2015) to provide valuable information to the public. Compared to air quality monitoring models, air quality forecasting models are considered more complicated, and the prediction methods typically vary from country to country.

Transport models are commonly used to assess the concentration distribution of air pollutants. By inputting certain emission inventories and meteorological data, these deterministic-type models can simulate key chemical and physical processes, including transformation or transfer of air pollutants, to calculate concentrations spatially and temporally. The most widely used atmospheric diffusion model is the Gaussian plume model (Kalhor and Bajoghli, 2017). However, it is mathematically simplified to a certain extent because of the complexity of nature and the insufficiency of the spatial or temporal resolution for other applications. Further, its

[*] Corresponding author at: Graduate Institute of Environmental Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 106, Taiwan, ROC.
E-mail address: ptchueh@ntu.edu.tw (P.-T. Chiueh).

https://doi.org/10.1016/j.uclim.2021.101055
Received 3 January 2021; Accepted 11 December 2021
Available online 20 December 2021

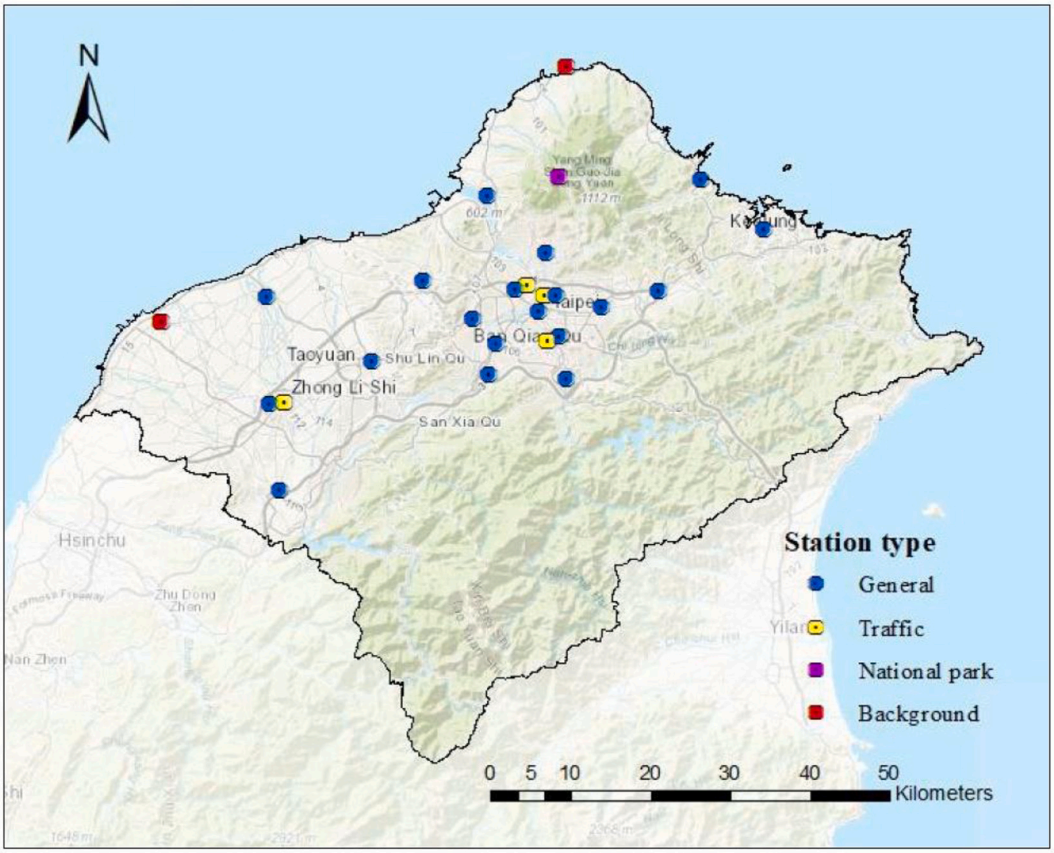


Fig. 1. Scope of the Northern Air Quality Basin.

execution tends to be relatively time-consuming; the time for computation can be several hours in some cases (Isukapalli, 1999). Another cause of uncertainty is difficulty in defining emission sources. Much of the required emission data are unavailable or indirect, which may increase estimation and model operation errors. Due to these limitations of transport models, machine learning based on statistics or mathematical algorithms has received considerable attention in recent years. Data patterns, also known as a training model, can be derived from the input data to make predictions of unknown data. Well-developed supervised learning algorithms include artificial neural network (ANN), decision tree (DT) (Kadiyala and Kumar, 2017), and support vector machine (SVM). The suitability of SVR for analyzing air quality has been proven (Murillo-Escobar et al., 2019). Because of higher accessibility and lower execution time, these methods are widely applied to air quality prediction as alternatives to physical and chemical models.

Most studies that apply machine learning to air quality prediction focus on meteorological effects on air quality and determine the parameters, e.g., temperature, relative humidity, precipitation, wind speed, and air pollutant concentration or air quality index (Dragomir and Oprea, 2014; Lu and Wang, 2005; Singh et al., 2013; Yeganeh et al., 2012). In terms of spatial features, human activities and divergence of land use are key influencing factors of urban air quality (Fenger, 1999; Karagulian et al., 2015; Kundu and Stone, 2014; Thouron et al., 2019; Xian, 2007). Vegetation has also been shown to have an impact on urban air quality (Sheng et al., 2019). In addition, traffic patterns evidently impact air quality (Cope et al., 2008); e.g., the concentration of NO_X has a high correlation with the peak period of urban vehicle activity (Kaminska, 2018). Due to the distinct spatial features, air quality at nearby monitoring stations may be significantly different at the same moment (Faridi et al., 2019; Guan et al., 2019; Leung et al., 2018).

Several studies have considered spatial features and conducted spatial air quality prediction using statistical methods. For example, Geographically Weighted Regression was used to characterize and quantify NO_2 concentrations (Alahmadi et al., 2019); Land Use Regression gave the formula for calculating the concentration of some air pollutants (Ma et al., 2019; Naughton et al., 2018; Van den Bossche et al., 2018). However, there remain obstacles to connecting prediction with temporal features and implementing dynamic prediction. Most spatial data, such as terrain, land use, emission inventory, are incomplete, imprecise, or static. This creates difficulty in combining heterogeneous features. Co-training frameworks were proposed for better integration of spatial and temporal data (Hsieh et al., 2015; Zheng et al., 2013), but the problem of limited spatial features still needs to be addressed (Leung et al., 2019).

Rapid urbanization may have far-reaching effects in a short period (Huang et al., 2019), with air quality reduced owing to the difficulty in pollutant dispersion caused by high buildings (Eeftens et al., 2013). Therefore, statistical methods using restricted spatial data and combinations of temporal and spatial data are worthy of development (Ali et al., 2014; Fang et al., 2017; He et al., 2019).

This study aims to apply an SVM algorithm to air quality prediction with temporal and spatial features to provide a better view of

Table 1
Air quality statistics of different types of air quality monitoring stations.

	Park	Background	Traffic	General
Mean of AQI	44.70	56.42	65.42	65.62
Standard deviation	18.49	22.20	26.01	25.80

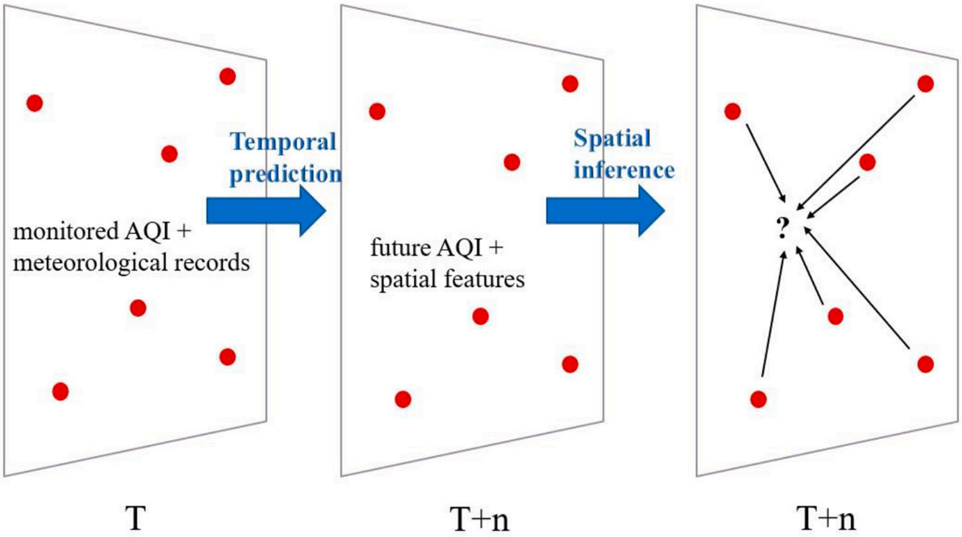


Fig. 2. Data processing framework of temporal prediction and spatial inference.

air quality of arbitrary locations without monitoring stations nearby. Geographic information systems (GIS) are used in this research to process spatial data for necessary features. By discovering the hidden relationships between air quality and spatio-temporal data, prediction of an unknown location with specific weather conditions becomes easy. The concrete objectives of this research are (i) to construct a general SVM-based air quality inference process with temporal and spatial features to increase the capability of unknown area prediction, and (ii) to verify the robustness of this prediction method in practice. A number of applications may be developable under this framework and may offer timelier and lower-cost alternatives to air quality prediction.

2. Methods

2.1. Study area

The study area was located in the Northern Air Basin of Taiwan (the Basin), an area unit of air management, as shown in Fig. 1. The Basin is densely populated and highly developed with approximately 40% of the entire Taiwan population residing in this area, despite the area only being 10% of the total area of Taiwan. Industry and commerce are the major economic patterns, and they contribute to heavy traffic flow, which impacts air quality in the Basin. There are twenty-five air quality monitoring stations distributed in this area, forming a network of air quality monitoring. The monitoring stations were categorized into background, park, traffic, and general stations, and their air quality statistics are summarized in Table 1. The Basin shows characteristics of a sub-tropical monsoon climate with monthly average temperatures of approximately 30 °C in summer and approximately 16 °C in winter. The annual average precipitation is approximately 2000 mm around most of the areas in the Basin. Sources of precipitation in this area are mainly plume rains in spring, convectional rains and typhoons in summer and in autumn, and the northeast monsoon in winter.

The northeast monsoon brings not only high moisture content to northern Taiwan but also air pollution, mostly dust and other particulate matters, to Taiwan in winter and even in spring. The majority of this cross-border pollution originates from China, and it is quite difficult for the Taiwan government to manage. Therefore, it is difficult to estimate air quality by estimating emissions. By contrast, heavy traffic and business activities in the metropolis are critical air pollution sources as well. The point sources, such as factories, incinerators, and power plants, may not be strongly responsible for the air pollution in the central business districts, but they are possible main sources in some nearby specific area. The air quality in winter and spring may be worse due to low diffusivity because of high-pressure systems (Yassin et al., 2018).

2.2. Framework of data processing

This study designed a framework to conduct temporal and spatial tasks, e.g., feature extraction and prediction, to achieve the goal

Table 2
Detailed spatial features.

Category	Item description	Unit
Population	Population	people/buffer zone
	Labor force	people/buffer zone
Road network	Length of freeway	km/buffer zone
	Length of expressway	km/buffer zone
	Length of highway	km/buffer zone
	Length of countyway	km/buffer zone
	Length of city road	km/buffer zone
Land use	Area of agricultural & forest use	km^2/buffer zone
	Area of transportation use	km^2/buffer zone
	Area of residential use	km^2/buffer zone
	Area of industrial use	km^2/buffer zone
	Area of business use	km^2/buffer zone
	Area of school use	km^2/buffer zone
	Area of other use	km^2/buffer zone
	Number of schools	schools/buffer zone
	Number of factories	factories/buffer zone
Economic	Energy consumption per month	kwh/(buffer zone.month)
	Households with energy	households/buffer zone
	Income of factories	NTD*/buffer zone
	Employees of factories	people/buffer zone
Point source	Minimum distance to incinerators	km
	Average distance to incinerators	km
	Minimum distance to fire plants	km
	Average distance to fire plants	km
Elevation	Elevation	m

* New Taiwan Dollar.

of future air quality prediction of unknown locations, as shown in Fig. 2. Both temporal prediction and spatial inference were executed via support vector regression (SVR). Here, a data instance is composed of a label and several attributes. The label is the dependent variable and the desired prediction. The attributes, as independent variables, can be any features relevant to the label. A training model, formed after inputting training dataset into the SVR computation, is the learned relation between labels and attributes. Predictions can be therefore implemented by acting on the training model.

- Temporal prediction

In this article, temporal prediction refers to a process that uses temporal attributes to predict future air quality. Various air pollutant concentrations were monitored and hourly air quality index (AQI) could be calculated. However, the AQI data could only be obtained from locations with monitoring stations. Each data instance, including the current AQI and meteorological records, contains temporal attributes. Future AQI (e.g., AQI of the next hour) was regarded as a label of a data instance. As a dataset usually works with labels and temporal attributes, 80% of the data were selected as training data, and the remainder were used for prediction and verification. The purpose of temporal prediction is to learn the future AQIs of existing monitoring stations (or reference stations) according to the relationships between labels and temporal attributes constructed from the training data. In addition, air quality in the near past could be influential to future air quality; abrupt considerable change of air quality seldom occurred in the monitoring records.

- Spatial inference

Spatial inference was conducted to acquire future AQIs of unknown locations from known locations. Note that the term “inference” is used instead of “prediction”. The key factor supporting effective inference of the AQIs of unknown locations is the connection of spatial features among locations. Locations at the same timestamp with similar spatial patterns tend to have similar air quality. These spatial features were inserted as attributes into the SVR process by GIS tools. Future AQIs obtained from temporal prediction were regarded as labels. Each inference was allocated to the training set (80%) or the inferring set (20%), with corresponding labels and spatial attributes attached. Note that all the data in each inference must share the same timestamp. Training set data came from the known reference stations while inferring set data were extracted from any unknown location. To measure the performance of temporal prediction and spatial inference concretely, the predicted/inferred results were compared to actual values in predicting/inferring sets.

2.3. Data extraction

To carry out temporal prediction and spatial inference, raw data was converted into temporal and spatial features as attributes to be inputted in the SVR operation.

- Temporal features

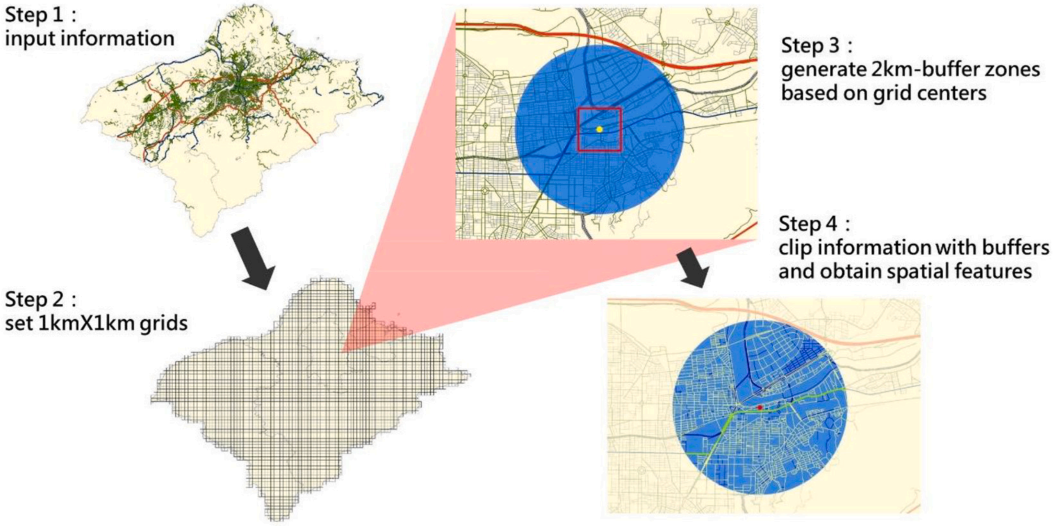


Fig. 3. Schematic flow of the spatial feature data extraction.

Monitoring data, varied with time, was selected to be transformed into temporal features, such as the current AQI and meteorological parameters. According to USEPA guidelines (USEPA, 2006), the AQI of a certain air pollutant can be calculated (Eq. (1)) as follows:

$$I_p = \frac{I_{Hi} - I_{Lo}}{BP_{Hi} - BP_{Lo}} \left(C_p - BP_{Lo}\right) + I_{Lo} \tag{1}$$

where I_p is the AQI for the certain pollutant p; C_p is the concentration of the pollutant p; BP_{Hi} is the break point for the pollutant that is greater than or equal to C_p; BP_{Lo} is the break point for the pollutant that is less than or equal to C_p; I_{Hi} is the sub-index value corresponding to BP_{Hi}; and I_{Lo} is the sub-index value corresponding to BP_{Lo}. When there are multiple pollutants, the AQI is set as the largest number. Hourly AQIs from 2013 and 2014 were calculated from monitoring air pollutant (O_3, $PM_{2.5}$, PM_{10}, CO, SO_2, NO_2) concentration derived from the Taiwan Air Quality Monitoring Network (TAQMN) of Environmental Protection Administration (EPA). The current AQI of a data instance was used as a temporal attribute to represent the temporal continuity of air quality. Other temporal features extracted were meteorological parameters of a adjacent location temporally corresponding to the air quality obtained from the nearest weather station of the Central Weather Bureau, Taiwan. The parameters include pressure (hPa), ambient temperature (°C), relative humidity (%), precipitation (mm), wind speed (m/s), and wind direction (360°).

- Spatial features

Spatial features may suggest different human activity patterns or terrain. Most of the spatial data were surveyed or computed for long-term scales. The spatial features can be viewed as static factors that cannot change with the time resolution (hour) of this study. Detailed spatial features selected in our study are provided in Table 2, including various categories mentioned in previous studies that may have an impact on urban air quality. Population and economic activities directly reveal the development of an area and level of urbanization, and labor force, defined as total population from age 15 to 64, may be capable of implying the developing type and human activity patterns of an area (Marlier et al., 2016). We selected data from village-scale census, accomplished by Ministry of Interior of Taiwan in June 2014 as raw data for population features, and economic activities were expressed in terms of the number of households, employees and financial information, etc. (Karagulian et al., 2015). For traffic impact parameters, we choosed all types of roads as variables, which may have different vehicle composition and affect the prediction results. Similarly, we used each land use category to ensure that there are no overlooked impact factors. Connection with major point sources can also reveal the effects of pollution emission. Furthermore, elevation not only represents the correlation with terrain, but may also demonstrate the effects of meteorological conditions.

Spatial data can be processed into valuable numbers and be spatially assigned to any grid if necessary, as shown in Fig. 3. ESRI ArcGIS 10.1 was used in this study, and the tool ModelBuilder was utilized to iteratively and automatically operate mass layers and features with the same processing procedures. The entire the Basin was divided into 1 km × 1 km grids. The center points of the grids were processed to their own spatial features, representing the characteristics of the grids. The 2 km buffer was used to represent the characteristics of the grid as much as possible, which is summarized as the most representative distance (Beelen et al., 2013; Eeftens et al., 2012) instead of using more redundant information as training data.

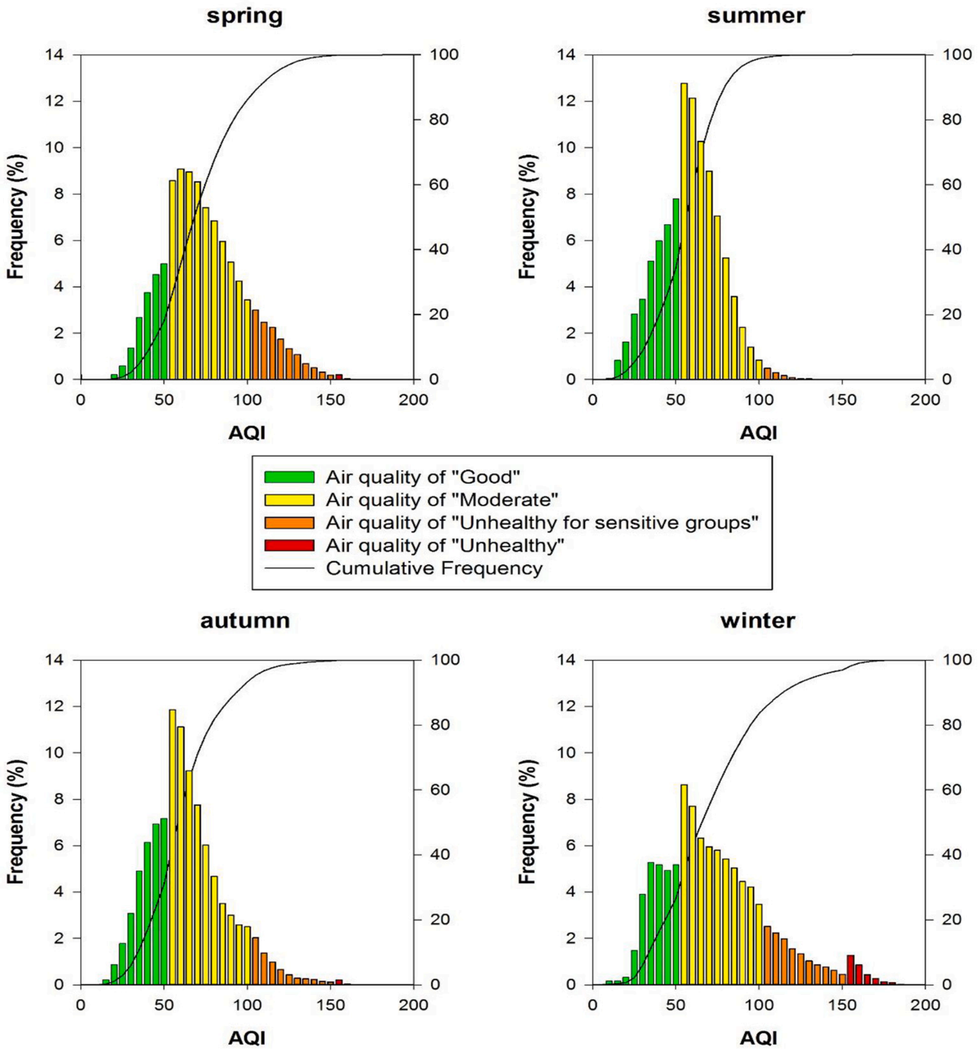


Fig. 4. AQI frequency of four seasons.

2.4. Support vector regression (SVM)

SVM, developed by Vapnik (1995), is popular in many fields. As it can efficiently deal with linear and nonlinear behavior in heterogeneous time series data set, it has sometimes been used in air pollution-related predictions (García Nieto et al., 2018; Saxena and Shekhawat, 2017). As a supervised learning method, the training model obtained from input data can be used for classification and regression. In a linear regression (Eq. (2)), given a data set

$$\mathrm{D} = \left\{ \left(x^1, y^1\right), \left(x^2, y^2\right), \ldots \left(x^n, y^n\right) \right\}, x \in \mathbb{R}^m, y \in \mathbb{R}, \tag{2}$$

where x is the attribute and y is the label with a linear function (Eq. (3)),

$$\mathrm{f(x)} = \langle w, x \rangle + b, \tag{3}$$

and the optimal regression function (Eq. (4)) can be expressed as:

$$min\Phi(w,\xi) = min\left(\frac{1}{2}\|w\|^2 + C\sum_i \xi_i^- + \xi_i^+\right), \tag{4}$$

where w is the support vector, C is the cost for adjusting the training model, and ξ_i^-, ξ_i^+ are the slack variables that enable errors to exist. A loss function can be introduced for SVR, suggesting an acceptable difference between the labels and the prediction. In this paper, the following ε-insensitive loss function is used (Eq. (5)):

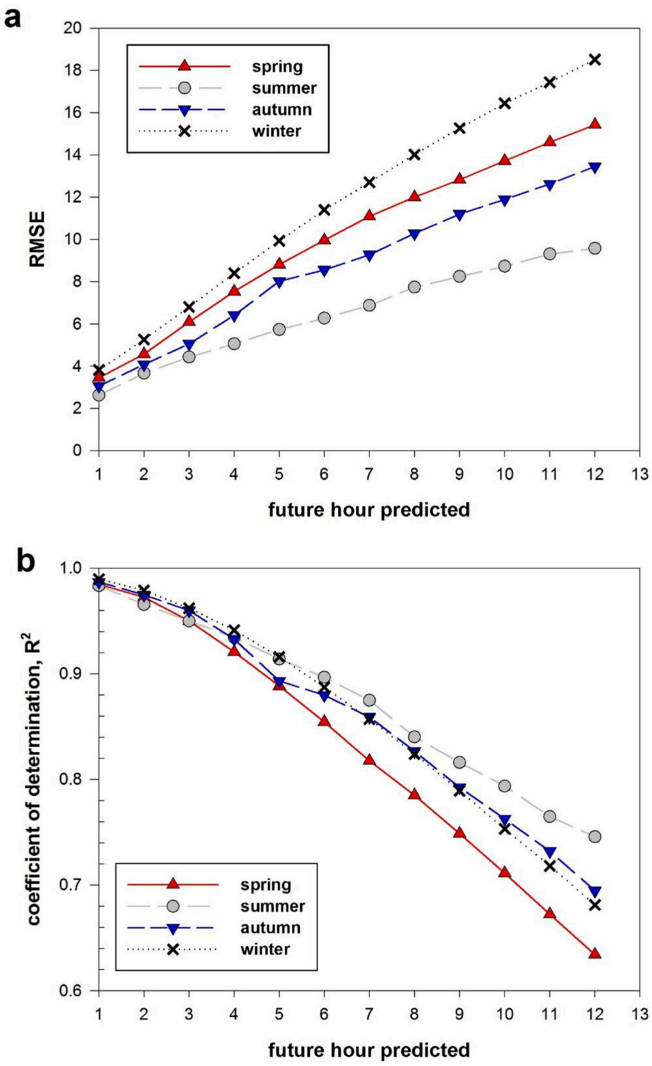


Fig. 5. Temporal effectiveness of prediction: (a) RMSE (b) R^2.

$$L_\varepsilon = \begin{cases} 0\,for\ |f(x) - y|. < \varepsilon \\ |f(x) - y| - \varepsilon, \text{otherwise} \end{cases}. \tag{5}$$

Combine the objective function and the loss function leads to (Eq. (6)):

$$minimize\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\left(\xi_i^- + \xi_i^+\right)\right)$$

$$\text{subject to}\left\{ y_i - w\bullet x_i - b \le \varepsilon + \xi_i^- w\bullet x_i + b - y_i \le \varepsilon + \xi_i^+ \xi_i^-, \xi_i^+ \ge 0. \right. \tag{6}$$

The solution of this problem can be transformed into the following dual form (Eq. (7)):

$$max\left(\begin{array}{c} -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right)\langle x_i, x_j\rangle + \\ \sum_{i=1}^{l}\alpha_i(y_i - \varepsilon) - \alpha_i^*(y_i + \varepsilon) \end{array}\right), \tag{7}$$

Table 3
Average Pearson correlation coefficient of attributes and predicted AQI.

	Spring	Summer	Autumn	Winter
Current AQI	0.925233	0.930283	0.921797	0.922146
Pressure	0.205537	0.245747	0.214595	0.147327
Ambient temperature	0.031149	0.24357	−0.24497	0.174066
Relative humidity	−0.17364	−0.18967	−0.16596	−0.30643
Wind speed	−0.17037	−0.21594	−0.12323	−0.15774
Wind direction	−0.0249	0.029415	−0.02341	−0.03129
Precipitation	−0.06445	−0.07592	−0.08725	−0.18459

Hour and month are not included.

where α_i, α_i* are Lagrange multipliers with constraints (Eq. (8)),

$$\sum_{i}^{l}\left(\alpha_i-\alpha_i^*\right)=0,$$

$$0\le\alpha_i,\alpha_i^*\le C, i=1,\dots,l. \tag{8}$$

The dual form can be solved under the Karush-Kuhn-Tucker (KKT) conditions.

In nonlinear situations, the mapping function (ψ) enables nonlinear data to be mapped into a feature space for better classification or regression. The kernel function is defined as the inner product in a feature space as follows:

$$\mathrm{K}\left(x_i,x_j\right)=\left\langle\psi\left(x_i\right),\psi\left(x_j\right)\right\rangle. \tag{9}$$

The software LibSVM used in this study was developed by Chang and Lin (2011). The chosen kernel function is RBF (radial basis function) and the parameters are set by default, where C and γ are 1 and 1/numbers of attributes, respectively.

3. Results and discussion

The air quality pattern of different seasons varied widely in the Basin as shown in Fig. 4. The air quality in winter is found to be worse than in other season with obvious higher frequency of AQI categorized into "Unhealthy." Air quality in spring is slightly better than air quality in winter despite the fact that the general air quality performance is not good enough as a consequence of similar climatic behavior to winter. Summer and autumn have a resembling pattern of better air quality, which corresponds to their climatic characteristics as well.

3.1. Performance of temporal prediction

To verify the model's performance, 5000 instances were randomly selected to compose a dataset, which was split into training data and testing data. All stations were selectable in this verification. Seasonal differences for temporal prediction in terms of the Root Mean Square Error (RMSE) and the coefficient of determination (R^2) are illustrated in Fig. 5. The predicted AQIs were compared with the observed ones. When the current AQI was used as an attribute to predict the AQIs for the next hour, the prediction showed reasonable results of low RMSE and high R^2. The R^2 value for each season reached quite low RMSE, suggesting that air quality in the near future was strongly connected to the current air quality.

With each increment of future hour predicted, the predictability declined noticeably for all seasons. The temporal prediction performed best in summer for both statistical indicators (RMSE and R^2). Even for the next twelve hours prediction, RMSE of summer prediction remained under ten, and R^2 remained over 0.7. By contrast, winter and spring predictions were poorer. The predictability deteriorated when prediction time was prolonged.

Explanation of these phenomena is shown in Table 3 with discussions for the Pearson correlation coefficients of each attribute and the AQI predicted in the training dataset. Excluding the current AQI, which was the strongest attribute, the following attributes contributed more in summer than in other seasons: pressure, ambient temperature, relative humidity, and wind speed. Note that the Pearson correlation coefficient can only explain linear relations between dependent variables and independent variables. Relations between variables may be nonlinear; however, the Pearson correlation coefficient can still be representative to a certain extent.

To conclude, short-term predictability in the four seasons was proven robust. Prediction in summer performs the best. Even though the predictive target was a half day after, the robustness can remain at a satisfying level. This may be attributed to the relatively strong contribution of meteorological factors.

3.2. Performance of spatial inference

With the evidenced temporal predictability, the predicted AQI of air quality monitoring stations was utilized to infer the future air quality of unknown locations. The twenty-five air quality monitoring stations in the Basin were randomly separated into training stations (twenty stations) and testing stations (five stations). To infer the air quality of these five stations by using the data of the

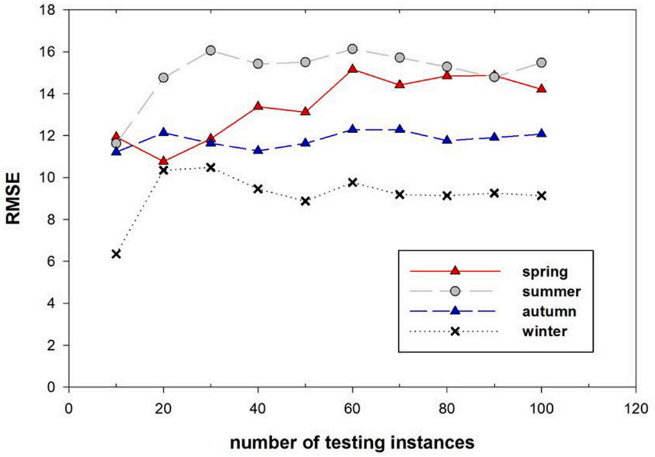


Fig. 6. Performance of spatial inference.

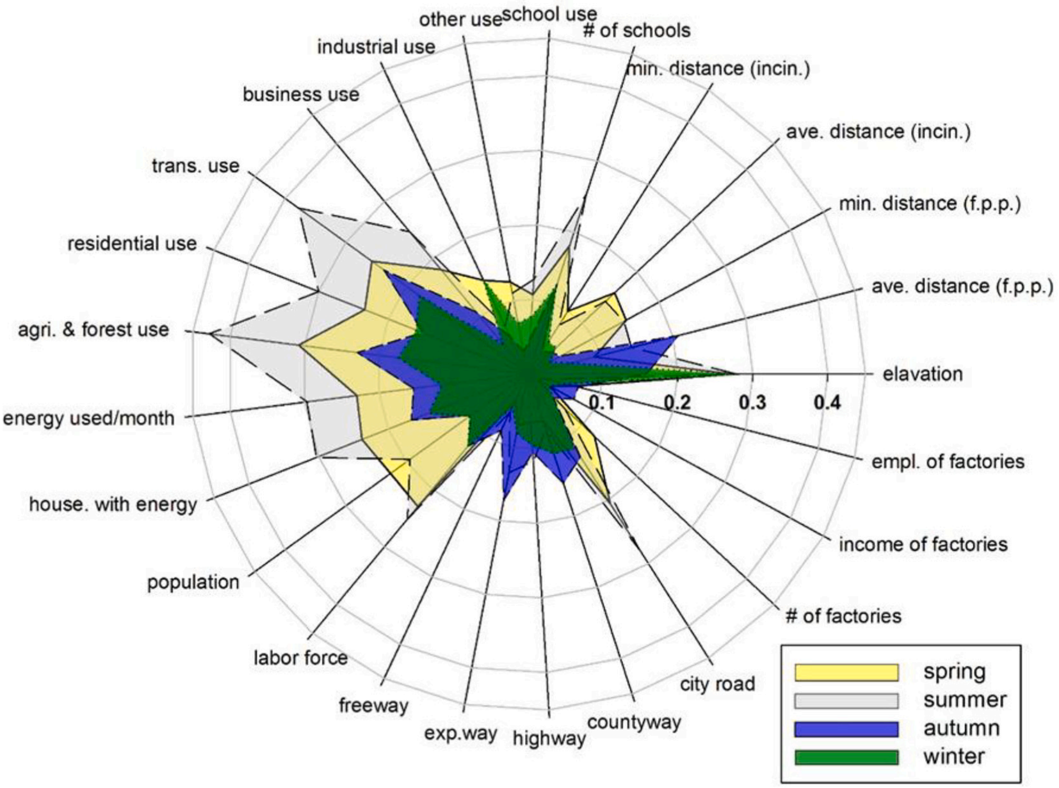


Fig. 7. Coefficient of determination (R^2) of spatial attributes and AQI.

twenty stations, the timestamp of the air quality in the training set must be the same in each inference. Twenty training sets (i.e., four hundred instances in total) were randomly sampled. Each instance was attached with the next hour AQI produced in the temporal prediction stage as the label.

3.2.1. *Seasonal difference in prediction performance*

The RMSEs of the four seasons with different amounts of datasets are shown in Fig. 6. Compared to the temporal prediction, an increase in the RMSE was observed in this stage. This might be because the spatial attributes were all static, which probably caused inaccuracy in the inference of the complex spatial relationship. The RMSE was used to keep the prediction realistic when inferring the AQI of arbitrary unknown locations. To ensure the RMSE was representative, Fig. 6 also shows the performance against the number of datasets. The peaks mostly occurred with fewer samples; after about sixty instances, the RMSE became stable. Under this condition, the RMSE could be considered consistent.

Interestingly, the performance of distinct seasons was in contrast to the results of the temporal prediction. Spatial inference worked the best in winter and the worst in summer. The spatial air quality varied with seasons. The standard deviation of the AQIs of all twenty-five stations in each dataset (i.e., twenty-five instances share the same timestamp) was calculated and averaged seasonally, giving values of 12.77, 18.82, 10.88, and 10.54 from spring to winter, respectively. Standard deviations in winter and autumn were lower than in summer. This phenomenon might be associated with climate characteristics of northern Taiwan. In spring and winter, relative high pressure occurs, and the air tends to be stable and unlikely to diffuse. In addition, mass cross-border air pollution (mostly particulate matters) is transported via the steady northeast monsoon. The air quality in northern Taiwan is generally dominated by larger-scale climate factors, which consequently causes small spatial divergence. As shown in Table 3, weather factors are more influential in summer, which may lead to higher spatial divergence of air quality. The R^2 values of the standard deviation of air quality in each training set (twenty instances) and the RMSE of each testing set (five instances) were calculated (0.4694), showing that the higher spatial standard deviation resulted in higher RMSE. In other words, less spatial diversity enabled the inference model to have better performance. This issue may be related to the scope determination. The Basin is defined as a combination of administrative districts with high spatial heterogeneity. Thus, it diminished the utility of the training model involving insufficient attributes.

3.2.2. Effects of spatial attributes

Effects of spatial attributes on AQIs are shown in Fig. 7. Coefficients of determination were calculated for spatial attributes versus observed AQIs in each training set. Similar patterns were found in all seasons. Land use, especially for agriculture and forest use, transportation use, and residential use, were clearly correlated to AQIs. Economic factors related to energy (electricity) use were also relevant to AQIs. These attributes were strongly connected to the strength of human activities and urbanization. The relatively smaller effects of population and labor force on AQIs might be explained by underestimation of the real population as floating population was excluded in household registered data. Industrial use and other factory-related factors were approximately negligible since there are only a few industrial districts in Northern Taiwan. By contrast, commercial activities played a major role in the prediction results. Most road network density information was neglected except for city roads (lanes included). The high density of city roads often caused slower vehicle speed, which results in major production of pollution. Elevation, relating to terrain, was another essential factor of AQIs as expected which was influential in all seasons. The correlation between other attributes and air quality in winter is less significant than in other seasons.

3.3. Advantages and application of spatio-temporal air quality prediction

3.3.1. Advantages compared to current forecasts and other methods

The first and most significant advantage of spatio-temporal air quality prediction is to provide a more detailed air quality forecast in both spatial and temporal scales compared to current models. Human judgments play an important role in the current forecast methods in Taiwan. Experienced air quality forecasters summarize several results from different prediction methods to produce their final forecasts. The spatial resolution of forecasted air quality obeys the allocated air quality basins in which more than three counties may be included. That is, the spatial scale is too large for the more specific application. Regarding the temporal scale, air quality forecasts in Taiwan offers three-day forecasts, which makes citizens unable to react to abrupt changes in air quality.

Spatio-temporal air quality prediction proposed in this study was expected to offer a better interpretation of air quality fluctuation by using detailed information on the time and space scales. In fact, the scale could be arbitrary if matching data are inserted. The output characteristics (e.g., time scale, size of grids, or unit) of the machine learning process were identical to the input data. Hourly air quality forecasts were presented on a 1 km × 1 km spatial scale as a product of the spatio-temporal prediction in this study.

The second advantage of the developed prediction model is the facilitation of dynamic visualization. Statistical prediction methods normally consider predicted air quality in an entire scope, rather than in spatial variation. Transport models can simulate spatial dispersion of air pollution, but have limitations in visualization, e.g., the model may be unable to collect immediate emission data, and the operation time may take too long for real-time display. By contrast, the procedure of spatio-temporal air quality prediction was less complicated and the operation time was relatively short. Automation of the entire process of prediction was feasible. Spatial feature extraction was first established, and then the air quality monitoring records and meteorological measurements were captured. Overall, dynamic visualization of predicted air quality can be realized more accurately with this approach.

3.3.2. Management and application of spatio-temporal air quality prediction

As discussed, the spatio-temporal air quality prediction proposed in this study offers air quality prediction with smaller space and time resolution. The spatial resolution of a 1 km × 1 km grid provides a "community-scale" forecast, and the temporal resolution of 1 h enables comparison with real-time air quality monitoring records. Given that this approach outputs air quality forecasts with more detailed information, the government should be able to use these forecasts for more precise policies and decision-making.

Accurate air quality forecasting is essential for specific groups (e.g., children, students, and the elders) to take actions of pollution prevention. Sensitive locations may include hospital, schools, nurseries, and caring centers. With the information in their hands, administrations should be able to regulate the activities in these places for necessary protection, such as avoiding unnecessary outdoor activities, provision of masks, requesting air purification operations, or even cancelling school activities, based on the forecast report. Improvement in spatio-temporal air quality forecast may also benefit other people. For example, air quality management for gathering crowds (e.g., attendants for concerts or ball games) can be carried out with the support of these air quality forecasts. Detailed

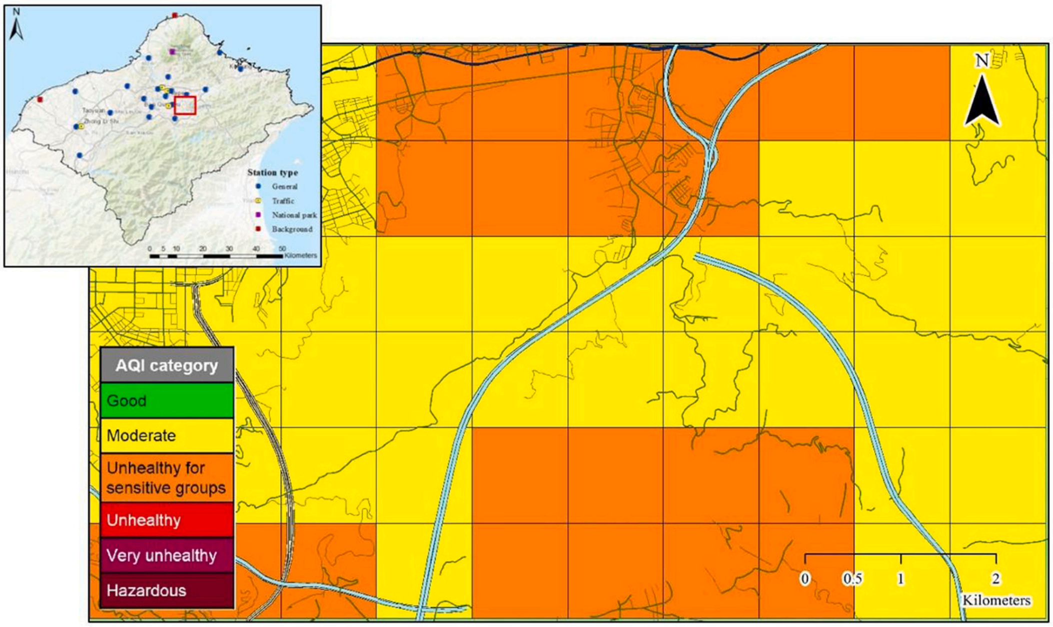


Fig. 8. Spatial display of authentic one-hour prediction at 3:00 am on Dec. 24, 2014.

information of air quality at regional or community levels also allows the public to take precautionary actions (e.g., wearing masks or avoiding outdoor activities). Although taking measurements may be time-consuming, visualization of the air quality can enable rapid response.

Regarding automation of spatio-temporal air quality prediction, real-time forecasting applications may be developed for mobile devices. Current air quality applications mostly focus on monitoring records at general scales. When dynamic visualization is achieved, the public will be able to use their own mobile devices to access future air quality in their locations of choice. A conceptual demonstration of the visualization is presented in Fig. 8, using a one-hour prediction of 3:00 am December 24, 2014 as an example. Spatial data of each grid were processed to obtain the spatial features. Air quality could be spatially inferred by inputting real-time air quality monitoring records into the training model, which was renewed by the input data of air quality and was used for spatial inference. The demonstration analyzed the spatial data of the grids and acquired their spatial features. The air quality records of the reference stations were used as the basis of the training model, which was subsequently used to infer the air quality of the grids.

4. Conclusions

In the spatio-temporal prediction system of air quality proposed by this study, high accuracy was observed in temporal prediction, especially in short-term prediction, due to the continuity of air quality. Prediction performance in summer was better than that in other seasons, because of the strong meteorological effects on air quality. By contrast, spatial inference only achieved acceptable accuracy, indicating that spatial inference with various spatial features might not be able to perfectly interpret the complexity of spatial heterogeneity of air quality. Using dynamic spatial data, such as monitoring city traffic volume or hourly electricity consumption, might strengthen the connection between time and space. In contrast to temporal prediction, the inference performances in winter were better than those in other seasons, which may be attributed to the climatic characteristics in Taiwan. In addition, the spatial features of urbanization and city types were found to correlate with air quality. Similar patterns were found for all seasons regarding correlation between the AQI and agriculture, transportation, and economic factors. Commercial activities also played a major role in the prediction results. However, there was no significant correlation between population and the AQI. This might indicate that in the prediction of urban air quality, the parameter of floating population and human activity is more important than household registered data. The proposed framework combining spatio-temporal features with an SVM algorithm achieved practically accurate air quality prediction for unknown time and space. The utilization of a GIS tool enhanced the capacity to process complicated spatial data. Despite the spatial inference only achieving acceptable accuracy and maintaining some obstacles, the framework is feasible in practice with controlled errors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Ministry of Science and Technology, R.O·C (Taiwan) [109-2621-M-002-009-], project of Measuring the Environmental Co-benefits of Strategies for Urban Climate Change Adaptation.

References

Alahmadi, S., Al-Ahmadi, K., Almeshari, M., 2019. Spatial variation in the association between NO_2 concentrations and shipping emissions in the Red Sea. Sci. Total Environ. 676, 131–143.

Ali, S., Tirumala, S.S., Sarrafzadeh, A., 2014. SVM aggregation modelling for spatio-temporal air pollution analysis. IEEE 249–254.

Beelen, R., Hoek, G., Vienneau, D., et al., 2013. Development of NO_2 and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe – the ESCAPE project. Atmos. Environ. 72, 10–23.

Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) 2 (3), 27.

Cope, M., et al., 2008. Traffic and meteorological impacts on near-road air quality: summary of methods and trends from the Raleigh near-road study. J. Air Waste Manage. Assoc. 58 (7), 865–878.

Dockery, D.W., et al., 1993. An association between air pollution and mortality in six US cities. N. Engl. J. Med. 329 (24), 1753–1759.

Dragomir, E.G., Oprea, M., 2014. Nonlinear Dynamics of Electronic Systems. Springer, pp. 387–394.

Eeftens, M., Beelen, R., de Hoogh, K., et al., 2012. Development of land use regression models for $PM_{(2.5)}$, $PM_{(2.5)}$ absorbance, $PM_{(10)}$ and $PM_{(coarse)}$ in 20 European study areas; results of the ESCAPE project. Environ. Sci. Technol. 46, 11195–11205.

Eeftens, M., et al., 2013. Quantifying urban street configuration for improvements in air pollution models. Atmos. Environ. 72, 1–9.

Fang, Z., et al., 2017. Spatiotemporal model for assessing the stability of urban human convergence and divergence patterns. Int. J. Geogr. Inf. Sci. 31 (11), 2119–2141.

Faridi, S., et al., 2019. Spatial homogeneity and heterogeneity of ambient air pollutants in Tehran. Sci. Total Environ. 697, 134123.

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.

Fenger, J., 1999. Urban air quality. Atmos. Environ. 33 (29), 4877–4900.

García Nieto, P.J., et al., 2018. PM_{10} 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.

Guan, Y., et al., 2019. Measurement of air-pollution inequality through a three-perspective accounting model. Sci. Total Environ. 696, 133937.

Harrison, R.M., Yin, J., 2000. Particulate matter in the atmosphere: which particle properties are important for its effects on health? Sci. Total Environ. 249 (1), 85–101.

He, Z., et al., 2019. Mining spatiotemporal association patterns from complex geographic phenomena. Int. J. Geogr. Inf. Sci. 1–26.

Hong, Y.-C., et al., 2002. Effects of air pollutants on acute stroke mortality. Environ. Health Perspect. 110 (2), 187.

Hsieh, H.-P., Lin, S.-D., Zheng, Y., 2015. Inferring air quality for station location recommendation based on urban big data. ACM 437–446.

Huang, K., et al., 2019. Estimating daily $PM_{2.5}$ concentrations in New York City at the neighborhood-scale: implications for integrating non-regulatory measurements. Sci. Total Environ. 697.

Isukapalli, S.S., 1999. Uncertainty Analysis of Transport-Transformation Models. The State University of New Jersey. PhD thesis.

Kadiyala, A., Kumar, A., 2017. Applications of Python to evaluate environmental data science problems. Environ. Prog. Sustain. Energy 36 (6), 1580–1586.

Kalhor, M., Bajoghli, M., 2017. Comparison of AERMOD, ADMS and ISC3 for incomplete upper air meteorological data (case study: steel plant). Atmosph. Pollut. Res. 8 (6), 1203–1208.

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 Wroclaw. J. Environ. Manag. 217, 164–174.

Karagulian, F., et al., 2015. Contributions to cities' ambient particulate matter (PM): a systematic review of local source contributions at global level. Atmos. Environ. 120, 475–483.

Kundu, S., Stone, E.A., 2014. Composition and sources of fine particulate matter across urban and rural sites in the Midwestern United States. Environ Sci Process Impacts 16 (6), 1360–1370.

Leung, Y., et al., 2018. An integrated web-based air pollution decision support system – a prototype. Int. J. Geogr. Inf. Sci. 32 (9), 1787–1814.

Leung, Y., et al., 2019. Integration of air pollution data collected by mobile sensors and ground-based stations to derive a spatiotemporal air pollution profile of a city. Int. J. Geogr. Inf. Sci. 33 (11), 2218–2240.

Lu, W.Z., Wang, W.J., 2005. Potential assessment of the "support vector machine" method in forecasting ambient air pollutant trends. Chemosphere 59 (5), 693–701.

Ma, X., et al., 2019. A site-optimised multi-scale GIS based land use regression model for simulating local scale patterns in air pollution. Sci. Total Environ. 685, 134–149.

Marlier, M.E., et al., 2016. Extreme air pollution in global megacities. Curr. Climate Change Rep. 15–27.

Murillo-Escobar, J., Sepulveda-Suescun, J.P., Correa, M.A., Orrego-Metaute, D., 2019. Forecasting concentrations of air pollutants using support vector regression improved with particle swarm optimization: case study in Aburrá Valley, Colombia. Urban Clim. 29, 100473.

Naughton, O., et al., 2018. A land use regression model for explaining spatial variation in air pollution levels using a wind sector based approach. Sci. Total Environ. 630, 1324–1334.

Pope III, C.A., et al., 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Jama 287 (9), 1132–1141.

Saxena, A., Shekhawat, S., 2017. Ambient air quality classification by Grey Wolf optimizer based support vector machine. J. Environ. Public Health 3131083.

Sheng, Q., et al., 2019. An experimental study to quantify road greenbelts and their association with $PM_{2.5}$ concentration along city main roads in Nanjing, China. Sci. Total Environ. 667, 710–717.

Singh, K.P., Gupta, S., Rai, P., 2013. Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmos. Environ. 80, 426–437.

Thouron, L., Kim, Y., Carissimo, B., et al., 2019. Intercomparison of two modeling approaches for traffic air pollution in street canyons. Urban Clim. 27, 163–178.

USEPA, 2006. Guidelines for the Reporting of Daily Air Quality –The Air Quality Index (AQI).

Van den Bossche, J., et al., 2018. Development and evaluation of land use regression models for black carbon based on bicycle and pedestrian measurements in the urban environment. Environ. Model. Softw. 99, 58–69.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory.

Xian, G., 2007. Analysis of impacts of urban land use and land cover on air quality in the Las Vegas region using remote sensing information and ground observations. Int. J. Remote Sens. 28 (24), 5427–5445.

Yassin, M.F., Al-Shatti, L.A., Al Rashidi, M.S., 2018. Assessment of the atmospheric mixing layer height and its effects on pollutant dispersion. Environ. Monit. Assess. 190 (7), 372.

Yeganeh, B., et al., 2012. Prediction of CO concentrations based on a hybrid Partial Least Square and Support Vector Machine model. Atmos. Environ. 55, 357–365.

Zheng, Y., Liu, F., Hsieh, H.-P., 2013. U-air: when urban air quality inference meets big data. ACM 1436–1444.