Aerosol and Air Quality Research

ORIGINAL RESEARCH
https://doi.org/10.4209/aaqr.210315

Optimizing the $PM_{2.5}$ Tradeoffs: The Case of Taiwan

Shihping Kevin Huang[1*], Sin-Yao Chen[1], Kuei-Lan Chou[2], Wei Chung Hsu[2], Kang-Hua Lai[2], Tung-Hung Chueh[2], Lopin Kuo[3], William Lu[1]

[1] Institute of Management of Technology, National Yang Ming Chiao Tung University, Hsinchu 300093, Taiwan
[2] Green Energy and Environment Research Laboratories, Industrial Technology Research Institute, Hsinchu 310401, Taiwan
[3] TamKang University, New Taipei 251301, Taiwan

ABSTRACT

The causes of $PM_{2.5}$ is dynamic and systematic. However, many studies approach the $PM_{2.5}$ problem by focusing only on either socioeconomic factors or geo-meteorological factors in isolation such data insufficiency might undermine the effort to control $PM_{2.5}$. We propose a LSTM-XGBoost model composing both socioeconomic and geo-meteorological factors together to improve the $PM_{2.5}$ monitoring system. We forecast the weekly $PM_{2.5}$ concentrations in five regions in Taiwan based on machine learning training data. The results indicate that overall small trucks usage should be reduced by 80% while maintaining semi-trucks and passenger cars at current level. In addition, coal and IPP Gas power have no impact on $PM_{2.5}$ concentrations in central Taiwan while usage in passenger cars, small tracks and tractor trailers should be reduced by 80% in central Taiwan. Overall, central Taiwan and Chiayi regions have the highest $PM_{2.5}$ projections at XGBoost output of 68.5 and 59.1 level. Finally, our model indicates that the use of fossil fuel based small tracks and tractor trailers should be reduced by 80% to maintain a reasonable level of $PM_{2.5}$.

Keywords: Air pollution, Machine learning, $PM_{2.5}$, Forecasting

OPEN ACCESS

Received: November 28, 2021
Revised: June 7, 2022
Accepted: June 18, 2022

* **Corresponding Author:**
ksph@nycu.edu.tw

Publisher:
Taiwan Association for Aerosol Research
ISSN: 1680-8584 print
ISSN: 2071-1409 online

1 INTRODUCTION

Air pollution is a serious problem in Asia. Sun *et al.* (2019) reveal that the annual $PM_{2.5}$ (fine particulate matter) concentration in China is 57.75 μg m^{-3}, compared with 8.47 μg m^{-3} in the USA. High concentrations of $PM_{2.5}$ have altered weather patterns and created health hazards (Etchie *et al.*, 2017; Forbes *et al.*, 2009; Maione *et al.*, 2016). Consequently, many countries began to establish air quality standards and $PM_{2.5}$ monitoring systems as an initial step to mitigate the impact of $PM_{2.5}$ (Chen *et al.*, 2001). The causes of $PM_{2.5}$ can primarily be attributed to socioeconomic and geo-meteorological factors (Yun *et al.*, 2019). However, many studies approach the $PM_{2.5}$ problems by focusing only on either socioeconomic factors or geo-meteorological factors in isolation. The causes of $PM_{2.5}$ is dynamic and systematic. Treating the causes of $PM_{2.5}$ in isolation might fail to reveal many aspects of the $PM_{2.5}$ problems (Yang *et al.*, 2019; Zhou *et al.*, 2018; Ji *et al.*, 2018). We propose a model composing both socioeconomic and geo-meteorological factors together, to improve the $PM_{2.5}$ monitoring system.

This study categorizes relevant factors that influence $PM_{2.5}$ into socioeconomic and geo-meteorological factors. We use the machine learning method to forecast and analyze the short-term controllable factors on the premise of real geo-meteorological scenarios. We forecast weekly $PM_{2.5}$ concentrations for five regions in Taiwan using machine learning training data. Furthermore, we use the LSTM (Long Short-Term Memory) model and the XGBoost (eXtreme Gradient Boosting) model to analyze 31 kinds of variables to obtain the optimal recommended usage amount for regional transportation and power generation. Finally, we estimate the optimal balance between energy and transportation factors in these five districts that will minimize $PM_{2.5}$ concentrations. The results will assist stakeholders and policy makers in their $PM_{2.5}$-related decision-making process.

Both socioeconomic and geo-meteorological factors exhibit strong correlation to the density of $PM_{2.5}$. In terms of socioeconomic factors such as industrialization and urbanization in many countries is accompanied by air pollution. Yang *et al.* (2019), Zhou *et al.* (2018), and Ji *et al.* (2018) have analyzed the relationship between China's socioeconomic factors and $PM_{2.5}$ problems. Yang *et al.* (2019) first divided 31 Chinese provinces into seven major areas. They then analyzed and compared the correlation between $PM_{2.5}$ concentrations and four socioeconomic factors (GDP, industrialization, family car ownership and urban population density) from 1998 to 2016. The results demonstrate that there is a significant correlation between socioeconomic factors and $PM_{2.5}$ concentrations in China (Yang *et al.*, 2019). urbanization plays a major role in $PM_{2.5}$ formations (Lu *et al.*, 2017). Zhou *et al.* (2018) divided social economy into seven factors: gross domestic product (GDP), population density, industrial structure, industrial dust emission, road density, trade openness, electricity consumption, and analyzed the correlation between $PM_{2.5}$ concentrations in 190 Chinese cities and these seven factors. The spatial model results show that population density, industrial structure, industrial dust emission and road density have significant correlations ($R^2 = 0.634$) on $PM_{2.5}$ concentrations, while only GDP is negatively correlated. Industrial dust is produced in large quantities during the manufacturing and transportation process of large-scale industries, which leads to the large increases in $PM_{2.5}$ concentrations (Zhou *et al.*, 2018). Chinese Academy of Social Sciences indicates that the haze disasters in China are caused by coal combustion, industrial pollution, and vehicle exhaust emissions (Wang and Liu, 2014). In fact, a few large sources are responsible for most of the emissions such as the case in the Hebei Province, 21 power plants emit more than 90% of the emissions (Zhang *et al.*, 2014). Based on data from 79 developing countries from 2001 to 2010, Ji *et al.* (2018) also proposed that industrialization and urbanization are the main socioeconomic factors influencing $PM_{2.5}$ concentrations. For Hebei, Inner Mongolia, Anhui, Henan and Hunan, the economic development of these provinces relies heavily on secondary industries, which results in the emission of pollutants. In contrast, for Beijing, Tianjin, Shanghai and Guangdong, where industries have greater scale and GDP level is also higher, urbanization makes a much greater contribution to $PM_{2.5}$ concentrations than industrialization (Zhao *et al.*, 2019). Higher population density increases energy consumption and pollutant emissions because of the increased level of human activity. Lin *et al.* (2014) revealed that areas with high $PM_{2.5}$ concentrations have high levels of population densities, GDP and urbanization, such as North China, East China (including Shandong, Anhui and Jiangsu Province) and the Beijing-Tianjin-Hebei region (Lin *et al.*, 2014).

In terms of the geo-meteorological factors, several studies have revealed that terrain and climate, such as temperature, atmospheric pressure, relative humidity, and rainfall, have an important influence on the accumulation and diffusion of $PM_{2.5}$ (Liu *et al.*, 2017). Yun *et al.* (2019) also indicated that in addition to socioeconomic factors, the natural environment factors such as terrain and climate also contribute to $PM_{2.5}$ concentration. The accumulation and diffusion of $PM_{2.5}$ at different altitudes are affected by airflow, atmospheric pressure, temperature, and rainfall (Alvarez *et al.*, 2013). $PM_{2.5}$ can be suspended in the air more easily at low altitudes. Temperatures gradually decrease with the increases in altitude, while increases in air convection increases mobility of $PM_{2.5}$ particles. Therefore, the concentration of $PM_{2.5}$ in low altitude areas is higher than the high-altitude areas. The most influential climate factors on $PM_{2.5}$ are temperature and rainfall. The main diffusion way of suspended particles like $PM_{2.5}$ in the air is Brownian motion, whose intensity is related to the temperature. Therefore, when the atmospheric temperature is higher, the Brownian motion of these suspended particles becomes more intense, and the particles diffuse better. In terms of rainfall, raindrops will grab aerosol particles in the air through Brownian diffusion and inertial collision, so rainfall will decrease the concentration of $PM_{2.5}$ (Yun *et al.*, 2019). Previous studies have also found a positive correlation between temperature and $PM_{2.5}$ concentrations in Northwest China in the summer months (He *et al.*, 2019). The average concentration of $PM_{2.5}$ in summer and autumn is much lower than winter and spring; and air quality in summer is substantially better (Cheng and Li, 2010). Several studies indicate that most areas in China suffer from severe $PM_{2.5}$ pollutions in winter, especially in the Beijing-Tianjin-Hebei Region (Zhang *et al.*, 2016; Chen *et al.*, 2017). In these studies, the authors analyzed the correlation between geo-meteorological factors and $PM_{2.5}$ concentrations by seasons, rather than years. $PM_{2.5}$ concentrations was also found to be lowest in the summer, which may be because there are fewer geo-meteorological factors affecting $PM_{2.5}$ concentrations. Additional investigations on

these geo-meteorological factors affecting $PM_{2.5}$ in winter, such as wind direction, humidity, and solar radiation, found that these factors have mutual influence relationships with $PM_{2.5}$. Chen *et al.* (2017) point out that solar radiation induces atmospheric photolysis on organic carbon to reduce $PM_{2.5}$ concentration (Yang *et al.*, 2019; Chen *et al.*, 2017; Zhang and Cao, 2015). Several studies have found that relative humidity is negatively correlated with $PM_{2.5}$ as the season changes (Yun *et al.*, 2019). However, the research results on relative humidity made by Chen *et al.* (2017) and Huang *et al.* (2015) are contrary to other studies which found that relative humidity is positively correlated with $PM_{2.5}$. The main reason is that the higher relative humidity will cause more water vapors to attach to the particulate matters and increase the size and mass, resulting in a rise of $PM_{2.5}$ concentrations. The study conducted by Huang *et al.* (2015) noted that the wind speed, sunshine hours, and precipitation on the previous day are typically negatively correlated with $PM_{2.5}$, while the relative humidity and atmospheric pressure three days ago are positively correlated with $PM_{2.5}$ (Huang *et al.*, 2015). Yang *et al.* (2017) also analyze the factors of season, year, city and district scale and spatial and season changes. In their study, the spatial and temporal heterogeneity was identified as one of the important factors of $PM_{2.5}$ pollutions.

In the case of Taiwan, the researchers found that the main metal pollution sources of $PM_{2.5}$ were coal combustion (34.7%), traffic related emissions (24.2%), secondary aluminum smelting (22.3%) and heavy oil combustion (18.8%), and the concentrations of PM_{10} and $PM_{2.5}$ both rise in winter (Hsu *et al.*, 2016). Several studies on $PM_{2.5}$ have targeted traffic congestion emissions (Kuo *et al.*, 2009; Lin *et al.*, 2015) and detection of metallic composition in urban areas (Chen *et al.*, 2013). The main pollution sources of $PM_{2.5}$ in Kaohsiung are traffic exhaust gas, secondary aerosol, and outdoor combustion (agricultural burning and garbage incineration), etc. Numerous vehicles and factories have made air quality in Kaohsiung exceptionally poor (Chen *et al.*, 2001). The main sources of $PM_{2.5}$ in Tainan are traffic emissions and industrial activities (Lu *et al.*, 2016). There are approximately 1.94 million vehicles of various types, such as scooters, automobiles, buses, and trucks in the urban area of Tainan, among which the traffic volume of scooters is the highest, as well as more than 9000 factories. Insofar, current research on Taiwan's $PM_{2.5}$ problems are mostly correlational based without offering any dynamic model that realistically monitor the $PM_{2.5}$ densities in Taiwan. This research is an attempt to fill in this gap through proposing a dynamic model which incorporates both socioeconomic and geo-meteorological data. In addition, $PM_{2.5}$ studies in Taiwan fail to provide a complete data analysis from wider data sources. We believe Taiwan's dynamic geo-meteorological factors, full range of transportation modes as well as a wide mixture of power generation methods make Taiwan a good study case for our research purpose.

2 METHODOLOGY AND DATA

We divided Taiwan into five urban regions (Fig. 1): northern Taiwan (Taipei, New Taipei and Keelung), Hsinchu and nearby areas (Taoyuan, Hsinchu and Miaoli), Central Taiwan (Taichung, Changhua, and Nantou), Chiayi and nearby areas (Yunlin, Chiayi, and Tainan) and southern Taiwan (Kaohsiung and Pingtung). We adopted power generation, traffic, and meteorological conditions as influencing factors in the prediction of future values, and sensitivity analysis was carried out to assist the economic decision-making process regarding power generation and traffic. We measured traffic data reflected the traffic flow of 5 types of vehicles at roadside detection stations, namely, passenger cars, minivans, buses, trucks, and trailers every five minutes. Power generation data from the Ministry of Economic Affairs which included the hourly data from 8 energy sources of coal, diesel, natural gas, hydraulic power, nuclear energy, fuel, solar energy, wind power and biogas, while hourly the meteorological data from the weather bureau contained 9 kinds of information, namely, ultraviolet index, atmospheric temperature, rainfall, relative humidity, wind speed, wind direction, acid-base value (acid rain), conductivity (acid rain) and $PM_{2.5}$. In addition, we also collected 9 kinds of pollutants such as sulfur dioxide, carbon monoxide, ozone, nitrogen oxides, nitric oxide, nitrogen dioxide, total hydrocarbons, non-methane hydrocarbons, and methane. Moreover, changing seasons, counties and cities were regarded as variables, and the study period extended from 2016 to 2018.

We also adopt deep learning model to forecast $PM_{2.5}$ concentration based on a combination model of LSTM-XGBoost which provides better results than a single forecasting model (Li *et al.*, 2019;

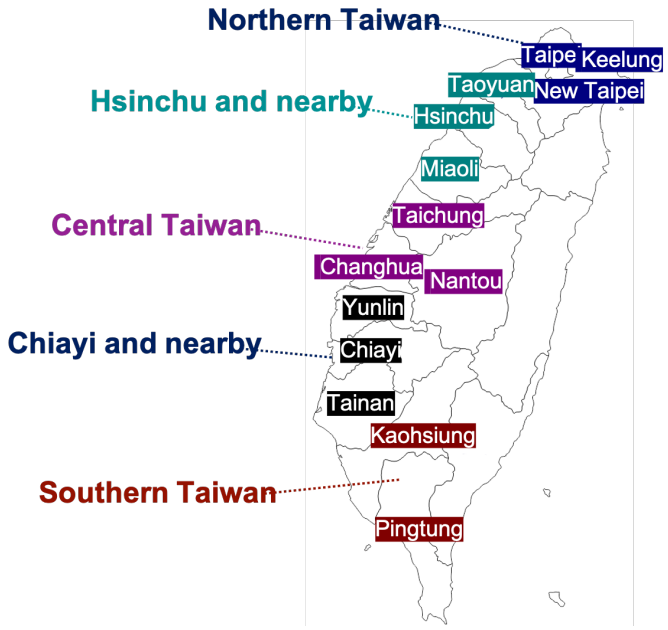

Fig. 1. The map of Taiwan.

Zhang and Zhang, 2020; Wang and Lu, 2020; Wei and Zeng, 2021). The proposed LSTM-XGBoost model is illustrated in Fig. 2. We first use LSTM to project future data. We then loaded the projected data into XGBoost for sensitivity analysis. On the other hand, we also try to determine the appropriate model using error comparison. We will explain the details in the following sections.

Our objective is to select an appropriate forecasting algorithm based on the sensitivity analysis of the factors influencing air pollution. Air pollution data are very complicated and dynamic. An appropriate algorithm was selected by means of the data features. The linear hypothesis of multiple linear regression is unsuitable for this purpose. When samples are imbalanced, the forecast deviation produced by k-NNs (k-Nearest Neighbor) is much larger. BPN (Back Propagation Neural Network) training data may experience the problems of convergence failure and local minimum values. Regression trees are simple decision trees without bagging or boosting, but their effects are poorer than those obtained with the XGBoost and random forest models. The SVR (support vector regression) cannot process data sets with excessive noise and more easily produces overfitting. Regarding the kernel, the Gaussian kernel feature is the best choice among the three kernels. Compared to bagging, boosting reduces errors and results in a lower variance, while bagging reduces the variance and slightly increases errors. In terms of the expected generalization error, the effect of boosting is better, so the expected effect of XGBoost is superior to that of a random forecast.

2.1 Statistical Measure of the Errors

When the same physical quantity is measured several times, the value and absolute error of each measurement are not the same. The absolute error of each measurement is determined as the absolute value and then averaged.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - y_i'\right| \tag{1}$$

Additionally, referred to as the standard error, the RMSE (Root Mean Square Error) is the square root of the ratio of the square of the deviation between the observed value and real value to the observation times n. The RMSE is very sensitive to extreme measurement errors. It only considers the average error instead of positive and negative values. The RMSE assesses the measurement precision well.

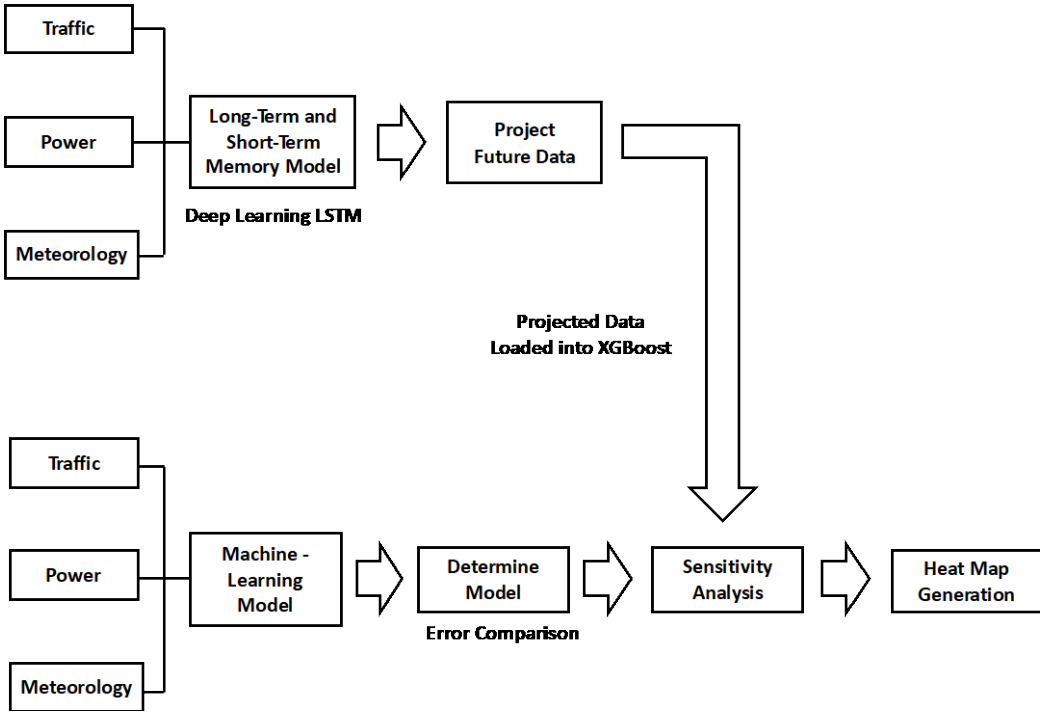

Fig. 2. LSTM-XGBoost model flow chart.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|y_i - y_i'\right|^2} \tag{2}$$

Compared to the standard error, the MAPE (Mean absolute percentage error) not only considers the error between the forecast value and the real value but also the proportion between the error and the real value, which objectively determines the difference between the estimated value and the evaluation value. The error in the MAE (Mean Absolute Error) and RMSE depends on the value of the test item. In certain cases, the error proportion between the forecast value and actual value yields a greater reference value than does the absolute error, while the MAPE considers not only the absolute error but also the dimension of the proportion.

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - y_i'}{y_i}\right| \tag{3}$$

According to Lewis (1982), the MAPE is the most effective evaluation index, and the MAPE is divided into four levels: highly accurate forecasts at an MAPE less than 10, excellent forecasts at an MAPE between 10 and 20, reasonable forecasts at an MAPE between 20 and 50 and inaccurate forecasts at an MAPE higher than 50.

2.2 Sensitivity Analysis

We gradually decreased the number of influencing factors of power generation and traffic by 10%. We first selected a certain factor to determine the value percentage decrease and denoted it with an 11 × 1 matrix, and the same change was then made to the next factor, with an 11 × 1 matrix denoted again. Each column of the second factor represented 11 columns of type 1, namely, the column was divided into 11 equal parts of type 1. Therefore, the second matrix actually contained 11^2 columns. The third factor was subjected to the same change. Each column of the third factor was a matrix of type 2 in the same way as mentioned above, so the third matrix contained 11^3 columns, and the process was repeated until all the factors were selected. Assuming that there are N factors, there should be 11^N combinations in total. In this way, we

clearly understand all combinations of the different factors matched with different percentage reduction values, and the $PM_{2.5}$ values yielding the minimum value can be determined. Otherwise, a reference source cannot be identified.

Assuming X is a column where the minimum $PM_{2.5}$ value is X ÷ 11 = A … B, and B is the numerical value represented by the column number of type 1 from top to bottom.

$X \div 11^2 = C \ldots D$, and D ÷ 11 = E, where E enters the integer unconditionally, which is the numerical value represented by the column number of type 2 from top to bottom.

$X \div 11^3 = F \ldots G$, and $G \div 11^2 = H$, where H enters the integer unconditionally, which is the numerical value represented by the column number of type 3 from top to bottom.

The calculation method is as follows: the dividend is divided by 11, and the remainder is the percentage of the value represented by the column number of type 1 from top to bottom. The dividend is divided by 11^2, and the remainder is again divided by 11 and enters the integer unconditionally, which is the percentage of the value represented by the column number of type 2 from top to bottom. The dividend is divided by 11^3, and the remainder is again divided by 11^2. The remainder is again divided and enters the integer unconditionally, which is the percentage of the value represented by the column number of type 3 from top to bottom, and the process is repeated until the last factor is selected. Then, the conditions pertaining to all factors can be summarized, as shown in Fig. 3.

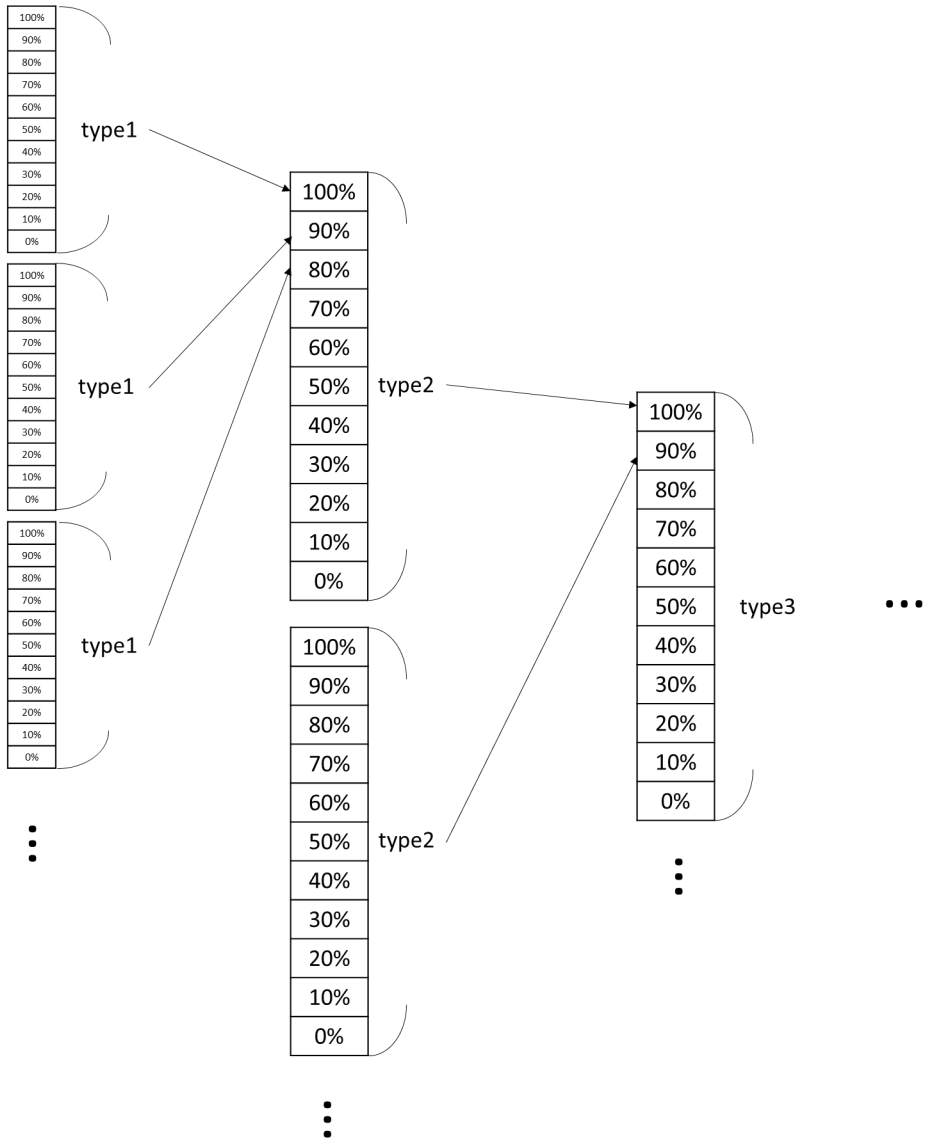

Fig. 3. The calculation method.

2.3 Data Collection

We first devised an hourly schedule from 2016 to 2018 and then populated the table with data. We also marked data with missing values as well as hidden missing data without any records. The time points of all the measurement stations with more than half missing data were deleted. If less than half the data was missing, all the measurement stations were consolidated, the air pollution values were averaged, and the power-generation amount and traffic volume were calculated. The deleted time points included May 19, May 27, June 16, June 20, June 21, June 25, August 29, August 31, September 22, September 23 and September 29 in 2016, 11 days altogether, which accounted for 1% of the total period. The data were consolidated into a 26016 × 47 matrix.

Not all variables in the data yielded a major impact on $PM_{2.5}$. We screened the data variables via stepwise regression and then established the machine learning model. However, the model screening results of the northern and southern regions were determined to be subject to the same variable. Therefore, the model from northern and southern Taiwan were consolidated. Five variables were selected. The first one is "District", including Hsinchu and nearby regions, Central Taiwan, Chiayi and nearby areas, and northern and southern Taiwan (combined into one variable). The second one is "Traffic", including passenger cars, minivans, buses, trucks, and trailers. The third one is "power generation", including fuel coal, fuel coal (privately operated), light oil (diesel), natural gas, natural gas (privately operated electricity), hydraulic power (pumped storage power generation), nuclear energy, fuel, solar energy and wind power. The fourth kind including atmospheric temperature, methane, carbon monoxide, non-methane hydrocarbons, nitric oxide, nitrogen dioxide, ozone, $PM_{2.5}$, rainfall, relative humidity, sulfur dioxide, total hydrocarbons, wind direction hourly value (average vector over the whole hour) and wind speed hourly value (arithmetic average over the whole hour). The fifth type is "Season", including spring, summer, autumn and winter.

2.4 Selection of the Machine Learning Model

The functionality of the LSTM model lies in the forecasting of sequence-related values. Although this study is related to time sequences, the ultimate purpose is to analyze the sensitivity of the selected factors. The operation principle of the LSTM model does not conform to the forecasting objective of the $PM_{2.5}$ value with a single set of parameters. XGBoost relies on loss function adjustment to optimize the model. Moreover, to objectively compare the advantages, disadvantages, and accuracy of these algorithms, three verification and evaluation indicators were compared among several models, as listed in Table 1, which all performed well. Therefore, XGBoost was selected as the model for sensitivity analysis.

The LSTM model was adopted to forecast the data of the coming week (168 groups of conditions by the hour over 7 days).

2.5 Sensitivity Analysis of the Influencing Factor Intensity

Among the data to forecast the following week, the numerical value of $PM_{2.5}$, which ranks the third highest among these districts, is chosen as an example.

Traffic is divided into four categories: passenger cars (type 31), minivans (type 32), buses, trucks (type 42) and trailers (type 5). The original value is 100%, which gradually decreases to 0%

Table 1. Model verification and evaluation.

	RMSE	MAE	MAPE
MLR	9	5.82	37%
SVR (kernel = linear)	8.03	5.73	34%
SVR (kernel = radial)	5.67	3.9	21.3%
SVR (kernel = polynomial)	6.08	4.23	23.6%
Regression tree	9.86	7.24	45.7%
k-NN regression	7.25	5.1	29%
NN (one hidden layer and two nodes)	13.4	10.5	74.8%
Random forest	5.67	4	22.8%
XGBoost	4.87	3.42	18.42%

at intervals of 10%. There is a total of 11^4 condition combinations. After the sensitivity analysis, a minimum of 110 groups of $PM_{2.5}$ forecast conditions is selected to generate a heat map.

The selection of the power-generation mode varies from region to region. There are three kinds of main power-generation modes namely gas, coal, nuclear, in northern and southern Taiwan, and there is a total of 11^3 conditions with the value decreasing from 100% to 0% at intervals of 10%. A minimum of 110 groups of $PM_{2.5}$ forecast conditions is selected to produce a heat map (approximately the top 8%). There are 2 kinds of main power-generation modes in Hsinchu and nearby areas, Central Taiwan and Chiayi and nearby areas, and a total of 21^2 conditions with the value decreasing from 100% to 0% at intervals of 5%. A minimum of 35 groups of $PM_{2.5}$ forecast conditions is selected to generate a heat map (approximately the top 8%). Regarding the considered condition analysis improvement in the selection of forecast data, the third priority is selected as a reference to prevent extreme values.

The future parameters forecast with the LSTM model determine the $PM_{2.5}$ value. In regard to sensitivity analysis, the XGBoost model is applied to forecast $PM_{2.5}$ values based on forecast parameters that do not include $PM_{2.5}$, and two sets of $PM_{2.5}$ values are generated. In addition to the respective loss functions of the two models, loss function analysis of $PM_{2.5}$ between the two groups is also performed to confirm the difference between their values.

3 RESULTS AND DISCUSSION

According to the LSTM model, the 3rd highest $PM_{2.5}$ projected concentrations are selected for the purpose of consistency and rough average from the top 10 highest $PM_{2.5}$ concentrations of the following week to serve as a base for further XGBoost analysis. The projections of $PM_{2.5}$ density using LSTM and XGBoost are shown in Table 2. The first box in the upper column indicates Northern Taiwan's LSTM output as 28.65 and lower column shows XGBoost output as 23.1. We applied the same process to the rest of the regions. Overall, we observed that both LSTM and XGBoost outputs are closely correlated. We applied additional analysis in Table 3 using RMAE, MAE, and MAPE to further support the accuracy. The results revealed that only northern Taiwan attains a large MAPE value of 30% (due to the smaller value resulting in a higher percentage), while the other 4 regions all attain MAPE values within 20% which indicate the accuracy is within the acceptable range.

With these data, heat map analysis is performed. Among the 5 regions in Table 2, the projected $PM_{2.5}$ concentrations in Central Taiwan (XGBoost output of 68.5) and Chiayi regions (XGBoost output of 59.1) is the highest, that in Hsinchu and southern Taiwan is moderate, and northern Taiwan exhibits the lowest projected $PM_{2.5}$ concentrations. A heat map of each region is then generated by applying the traffic and power-generation variables to the XGBoost model. In

Table 2. Projections of the $PM_{2.5}$ concentration with the LTSM and XGBoost models.

Northern Taiwan		Hsinchu and Nearby		Central Taiwan	
LSTM	28.65	LSTM	48.83	LSTM	61.06
XGBoost	23.1	XGBoost	46.8	XGBoost	68.5
Chiayi and Nearby		Southern Taiwan			
LSTM	58.45	LSTM	43.02		
XGBoost	59.1	XGBoost	38.5		

Table 3. MAPE (Mean Absolute Percentage Error) analysis between the LTSM and XGBoost models in each region.

	RMSE	MAE	MAPE
Northern Taiwan	5.73	4.63	30.3%
Hsinchu and Nearby	6.08	4.75	19.4%
Central Taiwan	6.08	4.75	19.4%
Chiayi and Nearby	7.42	5.92	19%
Southern Taiwan	5.9	4.49	12.9%

northern Taiwan, passenger cars and small trucks are major means of transportation, with values ranging from 0.4–0.7, and nuclear power is the main contributor to power generation, with values up to 0.824. In Hsinchu and its nearby regions, major traffic factors include passenger cars and small trucks, with values of approximately 0.7, and IPP (independent power producers) Gas is the main power-generation source. In Central Taiwan, small truck usage is relatively high at 0.57 among the 4 traffic factors, and coal is the main source of power with a value of 0.667. In the Chiayi area, all 4 traffic factors are calculated at approximately 0.3, and the usage of IPP Coal as a power source yields a value of 0.644, which is slightly higher than that of IPP Gas at 0.498. Finally, passenger cars and tractor trailers are the main traffic factors, and nuclear power is the main power-generation source, at a value of 0.824.

The 3rd highest $PM_{2.5}$ value is selected in sensitivity and heat map analysis of the contributing factors (please refer to supplementary data). Based on the heat map analysis results, the darker areas demonstrate that the data points pertaining to these areas adhere to the condition of the minimum $PM_{2.5}$ concentrations, thus indicating the ideal level of the corresponding factors.

According to the initial condition of each factor and the obtained heat maps, to attain the minimum $PM_{2.5}$ concentrations, the small truck category is the traffic factor that requires the most reduction. Among the power-generation factors, nuclear power usage yields the highest value at 0.824. However, the resulting heat map suggests that nuclear power should be improved to only 70–80% of its original value. In contrast, with a relatively low original usage level and impact, the optimal level of coal consumption is 70–90% or below 10% of its original usage level. Finally, despite its relatively low impact, the reduction in oil consumption to 40–60% of its original value may still slightly improve the $PM_{2.5}$ concentrations. In the Hsinchu region for example, based on Table 4, if the $PM_{2.5}$ concentration occurs at its minimum level, small trucks and tractor trailers are the traffic factors that required the most reduction, with both requiring an improvement from 10–20% of their initial usage levels. Regarding the other 2 traffic factors, usage reduction is not suggested. In contrast, gas is the power-generation factor that should be the most dramatically

Table 4. Traffic and power-generation variables in the 5 selected regions.

Northern Taiwan (Taipei city, New Taipei city, and Keelung)				
Traffic	Passenger Cars	Small Trucks	Semi-Trucks	Tractor Trailers
	0.492	0.63	0.219	0.286
Power Generation	Coal	Nuclear	Oil	
	0.107	0.824	0.187	
Hsinchu and Nearby Regions (Taoyuan, Hsinchu, and Miaoli)				
Traffic	Passenger Cars	Small Trucks	Semi-Trucks	Tractor Trailers
	0.757	0.783	0.369	0.277
Power Generation	Gas	IPP Gas		
	0.385	0.726		
Central Taiwan (Taichung, Changhua, and Nantou)				
Traffic	Passenger Cars	Small Trucks	Semi-Trucks	Tractor Trailers
	0.411	0.57	0.347	0.285
Power Generation	Coal	IPP Gas		
	0.667	0.193		
Chiayi and Nearby Regions (Yunlin, Chiayi, and Tainan)				
Traffic	Passenger Cars	Small Trucks	Semi-Trucks	Tractor Trailers
	0.291	0.358	0.307	0.37
Power Generation	IPP Coal	IPP Gas		
	0.644	0.498		
Southern Taiwan (Kaohsiung and Pingtung)				
Traffic	Passenger Cars	Small Trucks	Semi-Trucks	Tractor Trailers
	0.176	0.132	0.077	0.198
Power Generation	Coal	Nuclear	Gas	
	0.332	0.824	0.413	

Traffic: Highway traffic volume/hour; Power Generation: Power generated/hour.

improved, with a suggested consumption reduction target of more than 80%. Comparing IPP Gas to gas, even though the former exhibits approximately twice the usage level of that of the latter, gas remains the more important power-generation factor to improve the $PM_{2.5}$ level.

Based on Table 4 and the heat maps, to achieve the minimum $PM_{2.5}$ level, passenger cars, small trucks, and tractor trailers are all key traffic factors and should be reduced below their initial state. Regarding the power-generation factors, both major power-generation factors do not require adjustment. In the Chiayi region, to best improve the $PM_{2.5}$ level, tractor trailers and small trucks are the key traffic factors that should be reduced to below 10% and 20–10%, respectively. In terms of the power factors, IPP Coal is the most important factor and should be reduced to 70–55% or 90% of its initial usage level. The IPP Gas usage level, however, should remain at 80%, 90% and 100%. Even though IPP Gas achieves a slightly higher usage, IPP Coal remains the more important power factor in terms of improving the $PM_{2.5}$ level.

To achieve the minimum $PM_{2.5}$ level in southern Taiwan, tractor trailers and small trucks are the traffic factors that should be reduced the most regarding their usage levels. Tractor trailer usage should be reduced to 20% of its original usage level. Small truck usage, however, should be reduced to 30–40%. In terms of the power-generation factors, the most notable suggested reduction pertains to gas, which should be reduced to below 60%, whereas the coal and nuclear power usage levels may remain at 100%. To achieve $PM_{2.5}$ improvements, gas is the key factor among the power-generation factors.

The impact of the above traffic and power-generation factors on the $PM_{2.5}$ concentrations in the selected regions in Taiwan is summarized in Table 5 based on the heat map analysis. The Table 5 is the suggested traffic level for each region for maintaining a minimum $PM_{2.5}$ concentration.

Heat map analysis suggests that the small truck and tractor trailer usage levels should be reduced to 50% of their original usage levels in the 5 regions, which indicates that these 2 traffic factors are the common major contributing factors. In contrast, the passenger car usage level does not greatly impact the $PM_{2.5}$ concentrations in the selected regions, except for Central Taiwan, where the passenger car usage level should be reduced by more than 80%. In contrast, except in northern Taiwan, the tractor trailer usage level should be reduced by 50–90%, while it is suggested to increase this factor by 10% or 20% in the other regions. Comparing Tables 5 and 6, each region shares a similar usage level of tractor trailers. The causes underlying these findings should be further investigated to serve as references for future policy making.

Regarding the power factors, except in Central Taiwan, the considered power sources exert a meaningful degree of influence on the $PM_{2.5}$ levels in the other 4 regions. Gas is applied as a power source only in the Hsinchu region and southern Taiwan, and IPP Gas is consumed in the Hsinchu region, central Taiwan, and Chiayi region for power-generation purposes. Table 5 suggests that gas exerts a greater impact than does IPP Gas on the $PM_{2.5}$ concentrations. However, nuclear power is applied only in northern and southern Taiwan, and the heat map analysis results suggest that the nuclear power usage level should be reduced by 70–80%, implying that nuclear power yields a relatively low impact on the $PM_{2.5}$ concentrations. The results also reveal that in Central Taiwan, the 4 traffic factors impose greater effects on the $PM_{2.5}$ concentrations than do the 2 power factors, while in the other 4 regions, there is no significant difference between the impacts of the traffic and power-generation factors.

Table 5. Summary of the suggested levels of the traffic factors in each region based on the heat map.

Traffic				
	Passenger Cars	Small Trucks	Semi-Trucks	Tractor Trailers
Northern Taiwan	Maintain at 100% or reduce to 80 90%	Reduce to 10–20%	Maintain at 100% or reduce to 80–90%	Maintain at 100% or reduce to 50–90%
Hsinchu	Maintain at 100%	Reduce to 10%	Maintain at 100%	Reduce to 20%
Central Taiwan	Reduce to under 20%	Reduce to under 20%	Reduce to 70% or 90%	Reduce to under 20%
Chiayi	Reduce to 70–90%	Reduce to 10–20%	Maintain at 100% or reduce to 90%	Reduce to below 10%
Southern Taiwan	Reduce to 70–90%	Reduce to 30–50%	Maintain at 100% or reduce to 50%	Reduce to under 50%

*Percentage: Compared to the original usage level.

Table 6. Summary of the suggested levels of the power-generation factors in each region.

Power Generation			
Northern Taiwan	Coal	Nuclear	Oil
	Reduce to 70–90% or below 10%	Reduce to 70–80%	Reduce to 40–60%
Hsinchu	Gas	IPP Gas	
	Reduce to under 20%	Maintain at 100%, reduce to 80–95%, or reduce to 55%	
Central Taiwan	Coal	IPP Gas	
	Maintain at 100%	Maintain at 100%	
Chiayi	IPP Coal	IPP Gas	
	Reduce to 65–70%	Maintain at 100% or reduce to 90% or 80%	
Southern Taiwan	Coal	Nuclear	Gas
	Maintain at 100% or reduce to 70%	Maintain at 100% or reduce to 80% or 60%	Reduce to 40–60%

*Percentage: Compared to the original usage level.

4 CONCLUSIONS

In our research, we forecast the $PM_{2.5}$ concentration based on big data and machine learning methods. We suggest that decision makers consider controllable factors such as traffic and power-generation factors in their effort to reduce the $PM_{2.5}$ concentrations. The results of this study, based on actual geo-meteorological conditions, suggest that we need to reduce using fossil- fuel-based semi-trucks to ranging from 50% to 100% of current level depending on the regions. Consequently, replacing fossil- fuel-based semi-trucks to electric-based trucks could be a viable alternative. The government should encourage more conversation to non-fossil fuel-based vehicles in order to meet $PM_{2.5}$ standards. Furthermore, we found that coal and IPP Gas power have no impact on $PM_{2.5}$ concentrations in central Taiwan while usage in passenger cars, small tracks and tractor trailers should be reduced by 80% in Central Taiwan. In addition, we also revealed that a certain variability occurs in the $PM_{2.5}$ compositions among the different regions in Taiwan. Central Taiwan and Chiayi regions have the highest $PM_{2.5}$ projections at XGBoost output of 68.5 and 59.1 level. While Northern Taiwan's XGBoost output is only 23.1. Overall, machine-learning-based analysis is more dynamic than traditional analysis since the frequencies of our data ranged from 5 minutes to an hour. As such, the $PM_{2.5}$ concentrations was projected more accurately, and more effective proposals could be formulated to improve the traffic and power-generation characteristics of the above regions.

ACKNOWLEDGEMENTS

The author would like to express appreciation to the Ministry of Science and Technology's funding support (MOST 109-2410-H-009-040).

SUPPLEMENTARY MATERIAL

Supplementary material for this article can be found in the online version at https://doi.org/10.4209/aaqr.210315

REFERENCES

Alvarez, H.B., Echeverria, R.S., Alvarez, P.S., Krupa, S. (2013). Air quality standards for Particulate Matter (PM) at high altitude cities. Environ. Pollut. 173, 255–256. https://doi.org/10.1016/j.envpol.2012.09.025

Chen, H.W., Chen, W.Y., Chang, C.N., Chuang, Y.H. (2013). Characterization of particles in the ambience of the high-tech industrial park of central Taiwan. Aerosol Air Qual. Res. 13, 699–708. https://doi.org/10.4209/aaqr.2012.06.0155

Chen, K.S., Lin, C.F., Chou, Y.M. (2001). Determination of source contributions to ambient $PM_{2.5}$in Kaohsiung, Taiwan, using a receptor model. J. Air Waste Manage. Assoc. 51, 489–498. https://doi.org/10.1080/10473289.2001.10464287

Chen, Z., Cai, J., Gao, B., Xu, B., Dai, S., He, B., Xie, X. (2017). Detecting the causality influence of individual meteorological factors on local $PM_{2.5}$ concentration in the Jing-Jin-Ji region. Sci. Rep. 7, 40735. https://doi.org/10.1038/srep40735

Cheng, Y.H., Li, Y.S. (2010). Influences of traffic emissions and meteorological conditions on ambient PM_{10} and $PM_{2.5}$ levels at a highway toll station. Aerosol Air Qual. Res. 10, 456–462. https://doi.org/10.4209/aaqr.2010.04.0025

Etchie, T.O., Sivanesan, S., Adewuyi, G.O., Krishnamurthi, K., Rao, P.S., Etchie, A.T., Pillarisetti, A., Arora, N.K., Smith, K.R. (2017). The health burden and economic costs averted by ambient $PM_{2.5}$ pollution reductions in Nagpur, India. Environ. Int. 102, 145–156. https://doi.org/10.1016/j.envint.2017.02.010

Forbes, L.J.L., Patel, M.D., Rudnicka, A.R., Cook, D.G., Bush, T., Stedman, J.R., Whincup, P.H., Strachan, D.P., Anderson, H.R. (2009). Chronic exposure to outdoor air pollution and diagnosed cardiovascular disease: Meta-analysis of three large cross-sectional surveys. Environ. Health 8, 30. https://doi.org/10.1186/1476-069x-8-30

He, J., Ding, S., Liu, D. (2019). Exploring the spatiotemporal pattern of $PM_{2.5}$ distribution and its determinants in Chinese cities based on a multilevel analysis approach. Sci. Total Environ. 659, 1513–1525. https://doi.org/10.1016/j.scitotenv.2018.12.402

Hsu, C.Y., Chiang, H.C., Lin, S.L., Chen, M.J., Lin, T.Y., Chen, Y.C. (2016). Elemental characterization and source apportionment of PM_{10} and $PM_{2.5}$ in the western coastal area of central Taiwan. Sci. Total Environ. 541, 1139–1150. https://doi.org/10.1016/j.scitotenv.2015.09.122

Huang, F., Li, X., Wang, C., Xu, Q., Wang, W., Luo, Y., Tao, L., Gao, Q., Guo, J., Chen, S., Cao, K., Liu, L., Gao, N., Liu, X., Yang, K., Yan, A., Guo, X. (2015). $PM_{2.5}$ spatiotemporal variations and the relationship with meteorological factors during 2013-2014 in Beijing, China. PLoS One 10, e0141642. https://doi.org/10.1371/journal.pone.0141642

Ji, X., Yao, Y., Long, X. (2018). What causes $PM_{2.5}$ pollution? Cross-economy empirical analysis from socioeconomic perspective. Energy Policy 119, 458–472. https://doi.org/10.1016/j.enpol.2018.04.040

Kuo, C.Y., Wang, J.Y., Chang, S.H., Chen, M.C. (2009). Study of metal concentrations in the environment near diesel transport routes. Atmos. Environ. 43, 3070–3076. https://doi.org/10.1016/j.atmosenv.2009.03.028

Lewis, E.B. (1982). Control of Body Segment Differentiation in Drosophila by the Bithorax Gene Complex, in: Lipshitz, H.D. (Ed.), Genes, Development and Cancer, Springer US, Boston, MA, pp. 239–253. https://doi.org/10.1007/978-1-4419-8981-9_15

Li, C., Chen, Z., Liu, J., Li, D., Gao, X., Di, F., Li, L., Ji, X. (2019). Power Load Forecasting Based on the Combined Model of LSTM and XGBoost, in: Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence - PRAI '19, Presented at the the 2019 the International Conference, ACM Press, Wenzhou, China, pp. 46–51. https://doi.org/10.1145/3357777.3357792

Lin, G., Fu, J., Jiang, D., Hu, W., Dong, D., Huang, Y., Zhao, M. (2014). Spatio-temporal variation of $PM_{2.5}$ concentrations and their relationship with geographic and socioeconomic factors in China. Int. J. Environ. Res. Public Health 11, 173–186. https://doi.org/10.3390/ijerph110100173

Lin, Y.C., Tsai, C.J., Wu, Y.C., Zhang, R., Chi, K.H., Huang, Y.T., Lin, S.H., Hsu, S.C. (2015). Characteristics of trace metals in traffic-derived particles in Hsuehshan Tunnel, Taiwan: Size distribution, potential source, and fingerprinting metal ratio. Atmos. Chem. Phys. 15, 4117–4130. https://doi.org/10.5194/acp-15-4117-2015

Liu, H., Fang, C., Zhang, X., Wang, Z., Bao, C., Li, F. (2017). The effect of natural and anthropogenic factors on haze pollution in Chinese cities: A spatial econometrics approach. J. Cleaner Prod. 165, 323–333. https://doi.org/10.1016/j.jclepro.2017.07.127

Lu, D., Xu, J., Yang, D., Zhao, J. (2017). Spatio-temporal variation and influence factors of $PM_{2.5}$ concentrations in China from 1998 to 2014. Atmos. Pollut. Res. 8, 1151–1159. https://doi.org/10.1016/j.apr.2017.05.005

Lu, H.Y., Lin, S.L., Mwangi, J.K., Wang, L.C., Lin, H.Y. (2016). Characteristics and source

apportionment of atmospheric $PM_{2.5}$ at a coastal city in Southern Taiwan. Aerosol Air Qual. Res. 16, 1022–1034. https://doi.org/10.4209/aaqr.2016.01.0008

Maione, M., Fowler, D., Monks, P.S., Reis, S., Rudich, Y., Williams, M.L., Fuzzi, S. (2016). Air quality and climate change: Designing new win-win policies for Europe. Environ. Sci. Policy 65, 48–57. https://doi.org/10.1016/j.envsci.2016.03.011

Sun, R., Zhou, Y., Wu, J., Gong, Z. (2019). Influencing factors of $PM_{2.5}$ pollution: Disaster points of meteorological factors. Int. J. Environ. Res. Public Health 16, 3891. https://doi.org/10.3390/ijerph16203891

Wang, K., Liu, Y. (2014). Can Beijing fight with haze? Lessons can be learned from London and Los Angeles. Nat. Hazard. 72, 1265–1274. https://doi.org/10.1007/s11069-014-1069-8

Wang, X., Lu, X. (2020). A host-based anomaly detection framework using XGBoost and LSTM for IoT devices. Wireless Commun. Mobile Comput. 2020, e8838571. https://doi.org/10.1155/2020/8838571

Wei, H., Zeng, Q. (2021). Research on sales Forecast based on XGBoost-LSTM algorithm Model. J. Phys. Conf. Ser. 1754, 012191. https://doi.org/10.1088/1742-6596/1754/1/012191

Yang, Q., Yuan, Q., Li, T., Shen, H., Zhang, L. (2017). The relationships between $PM_{2.5}$ and meteorological factors in China: Seasonal and regional variations. Int. J. Environ. Res. Public Health 14, 1510. https://doi.org/10.3390/ijerph14121510

Yang, Y., Li, J., Zhu, G., Yuan, Q. (2019). Spatio-temporal relationship and evolvement of socioeconomic factors and $PM_{2.5}$ in China during 1998–2016. Int. J. Environ. Res. Public Health 16, 1149. https://doi.org/10.3390/ijerph16071149

Yun, G., He, Y., Jiang, Y., Dou, P., Dai, S. (2019). $PM_{2.5}$ spatiotemporal evolution and drivers in the Yangtze River Delta between 2005 and 2015. Atmosphere 10, 55. https://doi.org/10.3390/atmos10020055

Zhang, D., Liu, J., Li, B. (2014). Tackling air pollution in China—what do we learn from the great smog of 1950s in LONDON. Sustainability 6, 5322–5338. https://doi.org/10.3390/su6085322

Zhang, T., Liu, G., Zhu, Z., Gong, W., Ji, Y., Huang, Y. (2016). Real-time estimation of satellite-derived $PM_{2.5}$ based on a semi-physical geographically Weighted Regression Model. Int. J. Environ. Res. Public Health 13, 974. https://doi.org/10.3390/ijerph13100974

Zhang, X., Zhang, Q. (2020). Short-term traffic flow prediction based on LSTM-XGBoost combination model. CMES 125, 95–109. https://doi.org/10.32604/cmes.2020.011013

Zhang, Y.L., Cao, F. (2015). Fine particulate matter ($PM_{2.5}$) in China at a city level. Sci. Rep. 5, 14884. https://doi.org/10.1038/srep14884

Zhao, H., Guo, S., Zhao, H. (2019). Quantifying the impacts of economic progress, economic structure, urbanization process, and number of vehicles on $PM_{2.5}$ concentration: A provincial panel data model analysis of China. Int. J. Environ. Res. Public Health 16, 2926. https://doi.org/10.3390/ijerph16162926

Zhou, C., Chen, J., Wang, S. (2018). Examining the effects of socioeconomic development on fine particulate matter ($PM_{2.5}$) in China's cities using spatial regression and the geographical detector technique. Sci. Total Environ. 619–620, 436–445. https://doi.org/10.1016/j.scitotenv.2017.11.124