Science of the Total Environment 736 (2020) 139656

Contents lists available at ScienceDirect

Science of the Total Environment

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

Explore spatio-temporal PM2.5 features in northern Taiwan using machine learning techniques

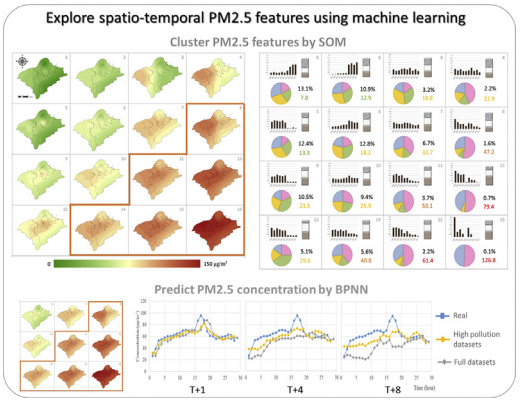
Fi-John Chang [a,*], Li-Chiu Chang [b], Che-Chia Kang [a], Yi-Shin Wang [a], Angela Huang [a]

[a] Department of Bioenvironmental Systems Engineering, National Taiwan University, Taipei 10617, Taiwan
[b] Department of Water Resources and Environmental Engineering, Tamkang University, New Taipei City 25137, Taiwan

HIGHLIGHTS

- The machine learning model well extracts spatio-temporal features of PM2.5 concentration.
- Change of seasons brings obvious effects on PM2.5 concentration variation.
- PM2.5 concentration variations have a direct relationship with human activities.
- High population density and heavy traffic load usually causes high PM2.5 concentrations.
- Multi-step-ahead PM2.5 prediction can be achieved accurately by machine learning.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 16 January 2020
Received in revised form 19 May 2020
Accepted 22 May 2020
Available online 23 May 2020

Editor: Jianmin Chen

Keywords:
PM2.5
Spatio-temporal variation
Multi-step-ahead prediction
Self-organizing map (SOM)
Back propagation neural network (BPNN)
Gamma Test

ABSTRACT

The complex mixtures of local emission sources and regional transportations of air pollutants make accurate PM2.5 prediction a very challenging yet crucial task, especially under high pollution conditions. A symbolic representation of spatio-temporal PM2.5 features is the key to effective air pollution regulatory plans that notify the public to take necessary precautions against air pollution. The self-organizing map (SOM) can cluster high-dimensional datasets to form a meaningful topological map. This study implements the SOM to effectively extract and clearly distinguish the spatio-temporal features of long-term regional PM2.5 concentrations in a visible two-dimensional topological map. The spatial distribution of the configured topological map spans the long-term datasets of 25 monitoring stations in northern Taiwan using the Kriging method, and the temporal behavior of PM2.5 concentrations at various time scales (i.e., yearly, seasonal, and hourly) are explored in detail. Finally, we establish a machine learning model to predict PM2.5 concentrations for high pollution events. The analytical results indicate that: (1) high population density and heavy traffic load correspond to high PM2.5 concentrations; (2) the change of seasons brings obvious effects on PM2.5 concentration variation; and (3) the key input variables of the prediction model identified by the Gamma Test can improve model's reliability and accuracy for multi-step-ahead PM2.5 prediction. The results demonstrated that machine learning techniques can skillfully summarize and visibly present the clusted spatio-temporal PM2.5 features as well as improve air quality prediction accuracy.

* Corresponding author.
E-mail address: changfj@ntu.edu.tw (F.-J. Chang).

https://doi.org/10.1016/j.scitotenv.2020.139656
0048-9697/

1. Introduction

Due to urbanization, industrialization, human activities and climate change in recent years, a large amount of suspended chemicals have directly or indirectly moved around Taiwan through atmospheric circulation, causing poor air quality (Zhang et al., 2018; Zhou et al., 2019a, 2019b). There are many types of air pollutants, among which PM2.5 characterized by small particle size, large surface area, and strong activity could easily adsorb toxic substances (e.g., heavy metals and microorganisms). PM2.5 not only has a long residence time in the atmosphere but also transports over a long distance, causing great impacts on human health and atmospheric environmental quality (Zheng et al., 2016; Sosa et al., 2017; Yu et al., 2018). For instance, Nowak et al. (2013) modeled tree effects on PM2.5 concentrations and human health for 10 U.S. cities. Callén et al. (2014) carried out a source apportionment of total polycyclic aromatic hydrocarbons by positive matrix factorization in order to quantify potential pollution sources of polycyclic aromatic hydrocarbons. He et al. (2016) estimated the spatial distribution of indicators addressing the humidity effect on East China using an observation-based algorithm. Stingone et al. (2017) explored a data-driven method to identify the relationship between air pollutant exposure profiles and children's cognitive skills. Ji et al. (2018) analyzed the socioeconomic factors of PM2.5 through quantitative assessment on stochastic impacts by regression on population, affluence and technology. PM2.5 concentrations greatly vary at a regional scale, depending on the local emission sources as well as climatic and geographic variables (Perrone et al., 2013). Moreover, they are highly associated with seasonal changes. For instance, Marchetti et al. (2019) reported seasonal-dependent biological effects coupled with regional climate variables and emission sources would finally result a high variability in the physico-chemical features of the particulate matter pollution and make the estimation and management of air quality a very challenging work. Therefore, investigating the emission sources and the proliferation mechanisms that cause serious air pollution is essential for effectively implementing air pollution mitigation strategies (Timmermans et al., 2017; Salavati et al., 2018).

Air pollution exhibits a high degree of uncertainty because the source of its generation and the mechanism of the proliferation process are dynamic and complex. A precise classification and prediction of PM2.5 concentrations is notably crucial to regulatory plans, which inform the public and restrain social activities when harmful events are foreseen. For tackling nonlinear problems, machine learning techniques, such as artificial neural networks (ANNs), can effectively extract and learn the spatio-temporal features from big datasets with complex relation between highly dimensional variables (Feng et al., 2015; Xu et al., 2018; Park et al., 2019). Li et al. (2018) introduced a self-adaptive neuro-fuzzy weighted extreme learning machine to predict air pollution concentration. Zhou et al. (2019a & b) explored multi-step-ahead PM2.5 forecasting models using deep long short-term memory (LSTM) and support vector machine (SVM) separately. Machine learning models can evaluate the characteristics of input data based on the results of model output, saving time in computation and scenario simulation. Zaman et al. (2017) and Alimissis et al. (2018) indicated that ANN models performed better than multiple linear regression models for air quality forecasting. Mishra et al. (2015) used a neuro-fuzzy model to forecast PM2.5 during haze episodes. Foehn et al. (2018) developed a regression co-kriging approach to efficiently combine weather radar data with rain gauge data. The complex nonlinear features between variables can be explored and reserved if the results of the machine learning model are used properly (Pisoni et al., 2009; Wu and Li, 2013; Elangasinghe et al., 2014). Therefore, many studies have been devoted to analyzing and exploring the main components of air pollution (Aristodemou et al., 2018) and the consequences at areas suffering from serious air pollution (Lanzaco et al., 2017; Orun et al., 2018). For instance, Brokamp et al. (2015) indicated that personal exposure to PM2.5 and most of its elemental constituents were correlated with both indoor and outdoor measurements significantly. Derwent et al. (2018) utilized Monte Carlo sampling to quantify model output uncertainties raised from global tropospheric ozone precursor emissions and from ozone production as well as destruction in a global Lagrangian chemistry-transport model. Zhao et al. (2018) suggested gross domestic product (GDP), private cars and energy consumption were significant positive factors for PM2.5 at five hotspots in China.

PM2.5 was legislated as an air pollutant in Taiwan in 2012. As known, the spatio-temporal distribution and characteristics of PM2.5 involve complex natural and anthropogenic sources. The mixture of local emission sources and regional transportation makes the control and accurate prediction of PM2.5 a very challenging work. The purpose of this study is to explore the features of urban air pollution and conduct the spatio-temporal analysis based on the long term datasets collected from a number of air quality monitoring stations in northern Taiwan. Besides, a machine learning technique is used to make accurate multi-step-ahead PM2.5 prediction. The results of this study can contribute to extracting the spatio-temporal distributions and transport mechanisms of PM2.5 features as well as providing decision makers with valuable information including air quality control strategies and their risks to human health. This paper is organized into five parts. The research background is given at first, follow by the study area and materials, and then the machine learning methods used in this study. In the fourth part, results are presented in a visual map and a discussion is given. Finally, we summarize the results and draw the important findings of the spatio-temporal analysis on regional PM2.5 concentrations.

2. Study area and materials

The study area spanning 3678 km^2 contains Keelung, Taipei, New Taipei City, and Taoyuan in northern Taiwan. The population density of northern Taiwan remains high, and Taipei, in particular, has the highest population density in Taiwan. Daily transportation and commuting choices are quite diverse. The most common choices for public transportation are mass rapid transit (MRT), bus, train, and high-speed rail. Alternatively, motorcycles and automobiles are two major vehicles contributing to a significant amount of urban air pollution. Air pollution becomes more serious as the number of motorcycles and automobiles increases (Basagaña et al., 2018).

Fig. 1 shows the study area and the locations of 25 air quality monitoring stations established by the Taiwan Environmental Protection Administration (TW_EPA). This study extracted from the open data source of the TW_EPA a total of 87,674 hourly datasets collected at 25 air quality monitoring stations during 2008 and 2018. According to the types of air quality monitoring stations defined by the TW_EPA, this study investigates 18 general stations, 4 traffic stations, 1 national park station, 1 background station, and 1 background and general station. Station information is given in Table 1. As shown, the national park station (#3) has the lowest values of mean and standard deviation among all the stations. Stations (e.g. #8, #9, #10, and #11) located in the centers of Taipei and New Taipei City along the Danshui River, in general, have the highest mean values. Station Dayuan (#21) located within the industrial zone in Taoyuan has the maximum PM2.5 concentration (459 $\mu g/m^3$). Table 1 shows the 18 air quality/meteorological variables monitored at the 25 stations, which are used for modeling PM2.5 concentrations in this study.

3. Methodology

This study uses the SOM to configure the topological map of regional long-term PM2.5 concentrations for exploring the spatio-temporal relationship through a two-dimensional topological map. We carry out the mining of the configured SOM topological map for exploring in-depth interrelation of air quality, climate, and traffic conditions with PM2.5 concentration in the study area. Then, a non-linear method (i.e., Gamma Test) is used to identify the key factors affecting PM2.5

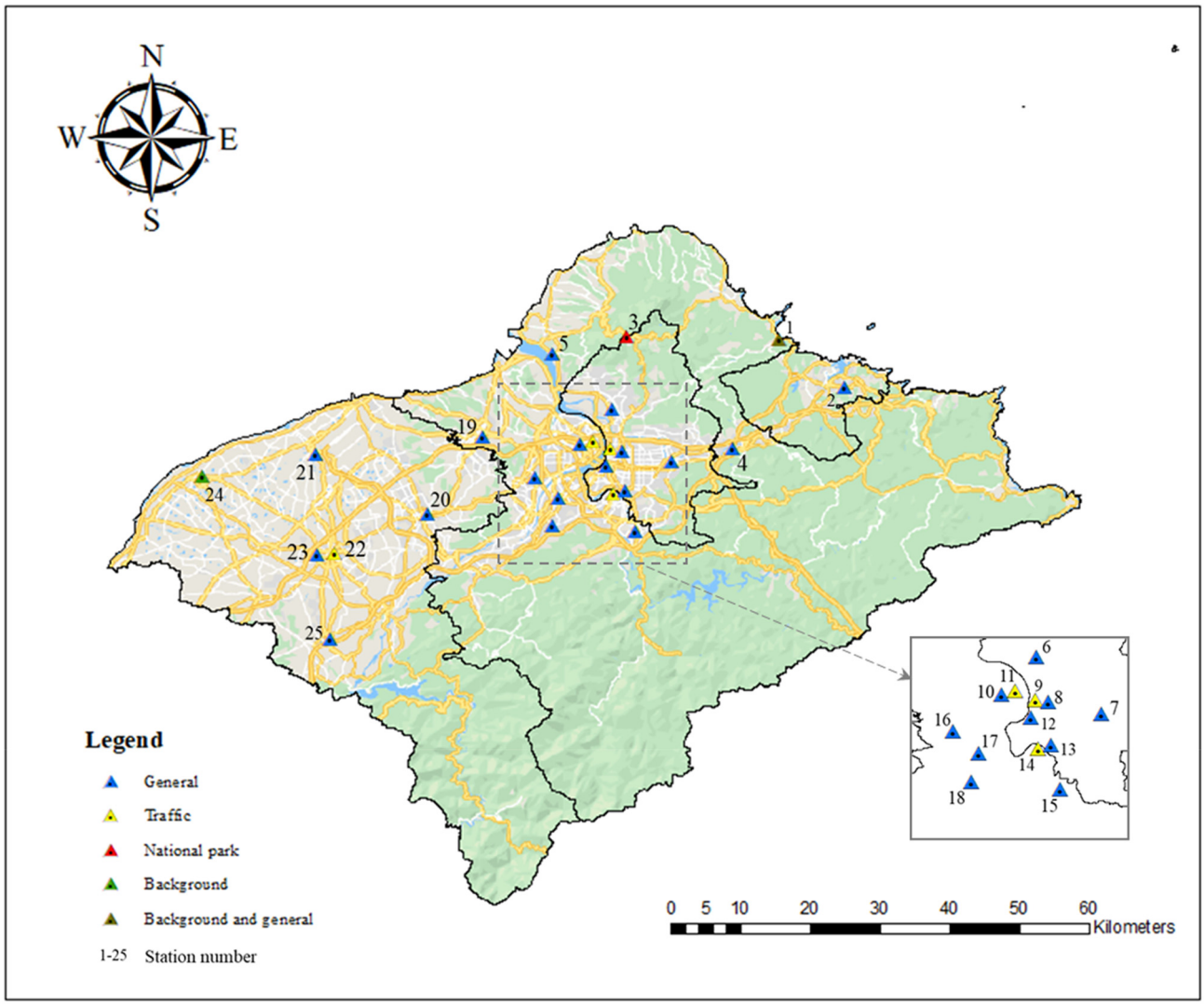


Fig. 1. Study area and the locations of 25 air quality monitoring stations established by the Taiwan Environmental Protection Administration (TW_EPA).

Table 1
Basic information and statistics of PM2.5 concentrations at 25 air quality monitoring stations (2008–2018).

#	Station	Type	Mean ($\mu g/m^3$)	Maximum ($\mu g/m^3$)	Standard deviation ($\mu g/m^3$)
1	Wanli	Background and general	18.14	192	13.26
2	Keelung	General	18.88	243	13.37
3	Yangming	National park	12.64	199	11.51
4	Xizhi	General	21.35	205	14.74
5	Danshui	General	20.57	173	14.68
6	Shilin	General	21.17	207	15.12
7	Songshan	General	24.42	226	15.81
8	Zhongshan	General	28.09	201	16.75
9	Datong	Traffic	27.82	375	19.44
10	Cailiao	General	23.34	310	16.44
11	Sanchong	Traffic	28.26	201	17.79
12	Wanhua	General	24.17	210	16.38
13	Guting	General	23.70	194	16.41
14	Yonghe	Traffic	23.88	184	16.60
15	Xindian	General	20.60	185	15.03
16	Xinzhuang	General	24.16	183	17.48
17	Banqiao	General	24.18	189	17.07
18	Linkou	General	22.90	203	15.81
19	Tucheng	General	23.98	227	16.99
20	Taoyuan	General	24.44	194	17.11
21	Dayuan	General	24.73	459	17.54
22	Zhongli	Traffic	26.21	212	17.29
23	Pingzhen	General	22.78	208	16.09
24	Guanyin	Background	23.65	241	16.30
25	Longtan	General	22.47	178	15.34

variation as inputs to a prediction model. Finally, an ANN model is built to predict multi-step-ahead PM2.5 concentrations for high pollution events, and the model's reliability and accuracy are evaluated. The research framework is shown in Fig. 2, and related methods are presented as follows.

3.1. Self-organizing map (SOM)

Kohonen (1982) first proposed the SOM to explore the interrelationships of high-dimensional multivariate systems. It is similar to the concept of feature mapping in the biological cerebral cortex, where similar information is integrated into similar clusters so that the information can be processed efficiently (Faigl et al., 2011; Serrien et al., 2018). The SOM that has been widely used in a broad range of disciplines can effectively reduce the complexity of high-

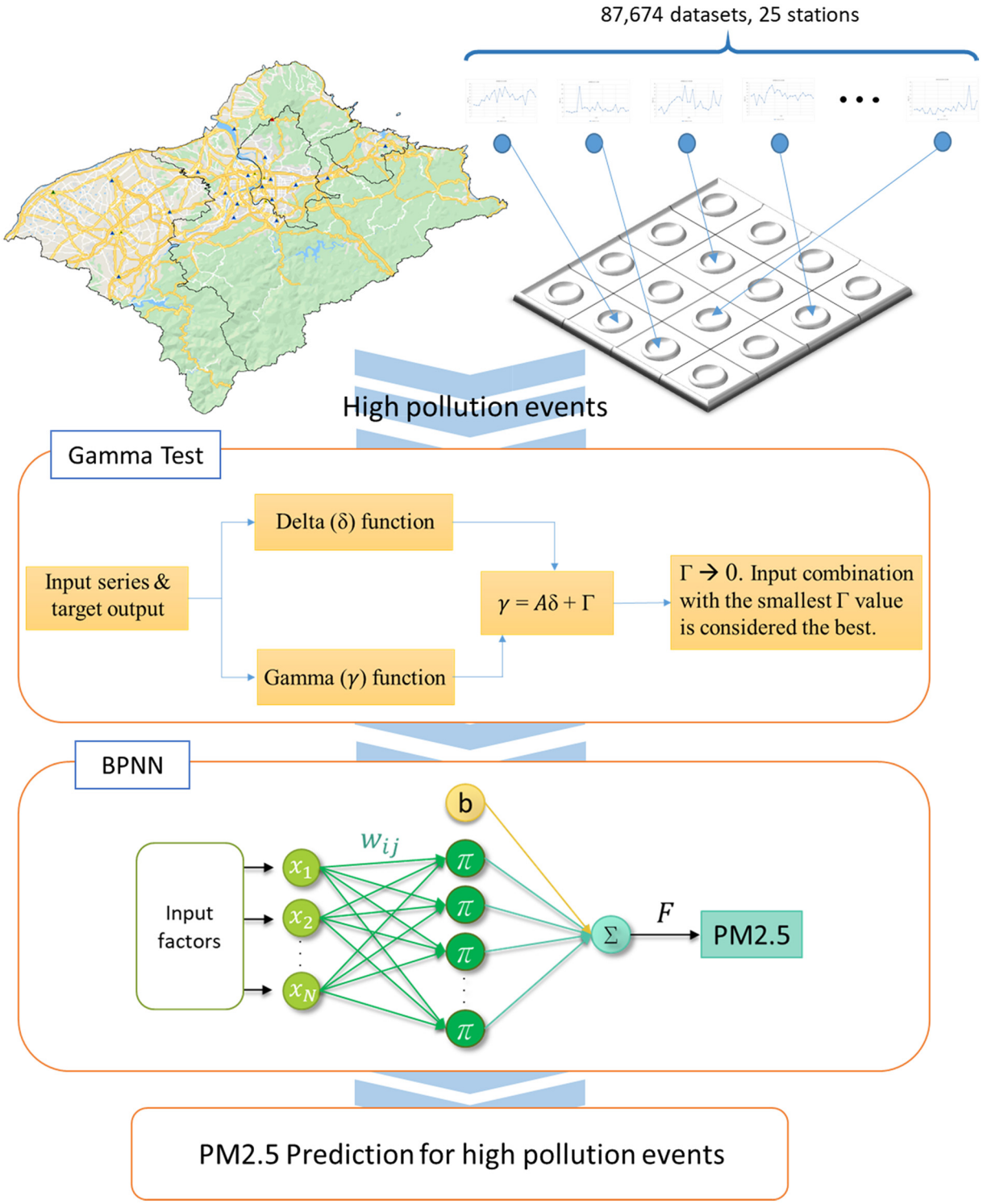


Fig. 2. Research framework of this study.

dimensional systems and map high-dimensional input vectors onto low-dimensional maps (Raza and Kim, 2008; Karaca and Camci, 2010; Newman and Cooper, 2010; Chang et al., 2014, 2016, 2020; Han et al., 2016; Wu et al., 2017). Therefore, a large amount of information can be stored in the weight values of the SOM's neurons, where similar characteristics in input vectors can be found (Chang et al., 2010; Heikkinen et al., 2011).

To establish an SOM topology, the size of the network is crucial and must be determined at first because different map sizes represent different degrees of data deviation. An SOM with a small map size would fail to effectively detect important features of data for clustering purpose while an SOM with a big map size would fail to adequately differentiate the features between neurons (clusters). Nevertheless, to the best of our knowledge, there is neither general theoretical principle to determine the optimal map size nor common evaluation indicator to evaluate the proximity between neurons. The shorter the distances of datasets within each cluster and the longer the distances between cluster centers, the more distinct the clusters. Therefore, this study tried different network sizes coupled with different numbers of iterations for presenting the most adequate topographical map of air pollution. The results revealed that the network became stable after 2000 iterations, and therefore 2000 iterations were determined for implementing the SOM. The network size was next determined among 3*3, 4*4 and 5*5 by trial and error in this study.

There are three schemes established by the SOM to efficiently analyze and visually present the spatio-temporal PM2.5 concentrations in this study, introduced below.

3.1.1. Scheme 1: SOM configuration

An SOM network (e.g., 4*4) was configured based on large datasets (i.e. 87,674) of PM2.5 concentrations at 25 monitoring stations in the study area (Fig. 3(a)). For each neuron of the topological map shown in Fig. 3(b), the pie chart displays the number and the ratio of data clustered in that neuron while the bar chart at the bottom presents the PM2.5 concentration at each station.

3.1.2. Scheme 2: spatial analysis

The configured SOM topological map illustrates the two-dimensional visualization of the spatial distribution of PM2.5 concentrations, where the spatial distribution shown in each neuron spans the point data at 25 stations using the Kriging method (Fig. 3(c)).

3.1.3. Scheme 3: spatio-temporal analysis

To clearly distinguish the temporal features of PM2.5 concentrations of the configured SOM topological map, the temporal behaviors of PM2.5 are extracted and presented at various time scales (Fig. 3(d)). In each neuron, the bar chart, the pie chart and the three-tier bar display PM2.5 concentrations at yearly, seasonal and daily scales, respectively. Besides, the mean of PM2.5 concentrations in each neuron is indicated with color according to the color indication of air quality standard defined by the TW_EPA.

A detailed presentation and description about the three schemes is given in the Results and discussion section.

3.2. Gamma Test

The Gamma Test first proposed by Koncar (1997) is an input selection technique that evaluates the extent to which a given input-output dataset can be modeled by an unknown smooth nonlinear function (Jones et al., 2007). Thus it has been frequently used to select the best combination of inputs for the corresponding outputs (e.g., Noori et al., 2010; Chang et al., 2014, 2015). For each subset of input variables, the Gamma Test uses a smooth function to calculate the noise estimate (Γ value) of the variance of model output that cannot be accounted for.

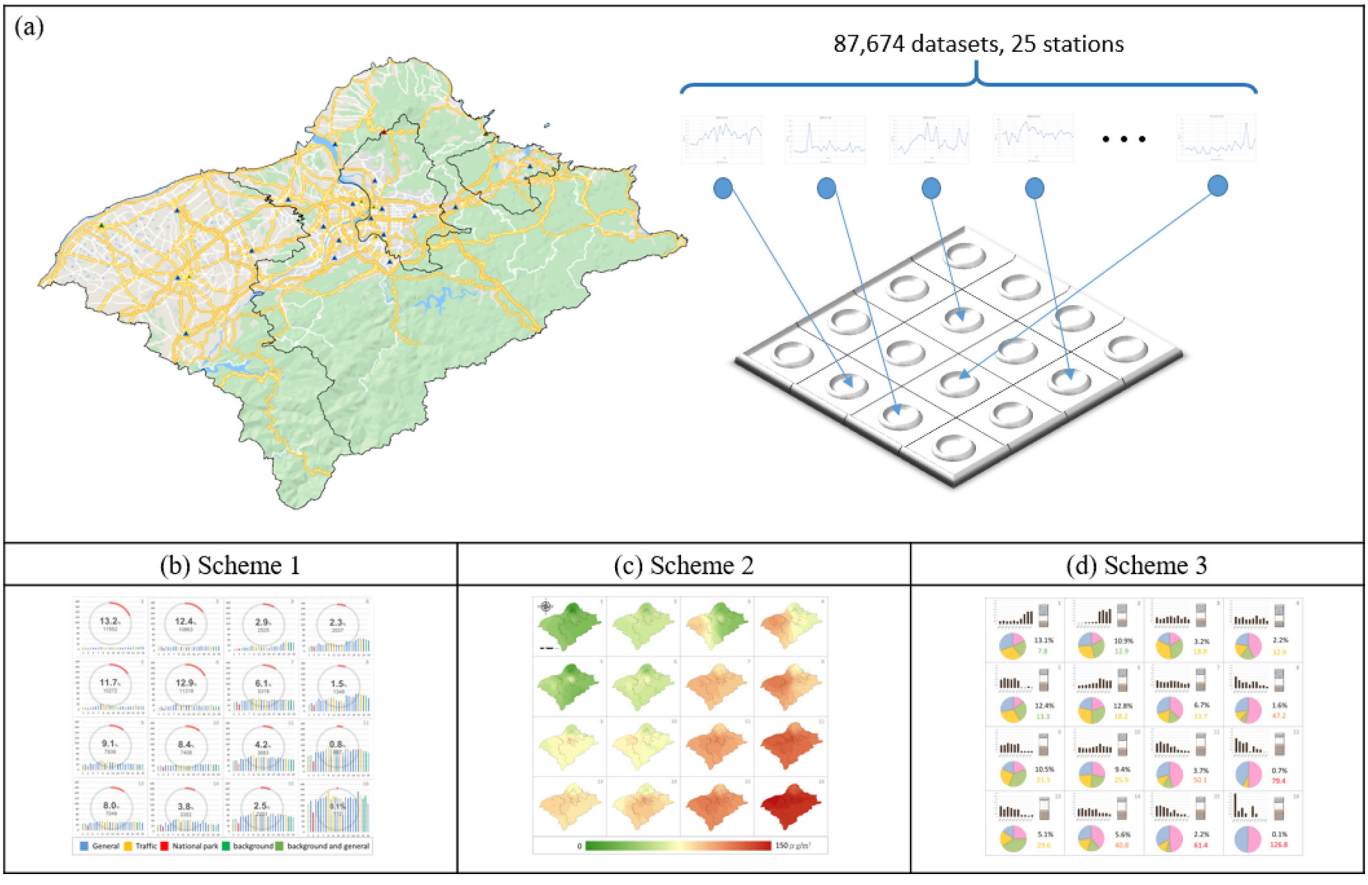


Fig. 3. Spatio-temporal analysis of PM2.5 concentrations at 25 monitoring stations in the northern Taiwan using the SOM.

The subset with its Γ value the closest to zero is determined as "the best combination" of input variables. The Gamma Test is implemented to identify non-trivial input variables for producing accurate outputs of ANN-based models in this study.

3.3. Back propagation neural network (BPNN)

The purpose of an ANN is to build a prediction model based on inputs. A neural network is trained by the patterns between input and output values. The BPNN can achieve predictions by mapping from an n-dimensional space to an m-dimensional space, and its simple structure, good accuracy and high operability makes it the most popular and commonly (frequently) used ANN in many fields. The BPNN with sufficient hidden neurons is capable of producing an accurate approximation of any continuous function through learning from the samples fed to it (Kow et al., 2020). Besides, the BPNN can generalize correct responses that widely resemble the data at the learning stage. Therefore, the BPNN is utilized to construct the prediction model in this study.

3.4. Evaluation metrics

For PM2.5 prediction, it is very essential to know the performance of a model when predicting high-magnitude data. Consequently, this study utilizes three metrics to evaluate model accuracy and predictability of PM2.5 concentrations, which are the Root Mean Square Error (RMSE), the coefficient of determination (R^2), and the Nash-Sutcliffe Efficiency coefficient (NSE, Nash, 1970). The formulae of the three metrics are expressed as follows.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(o_i - p_i)^2}{N}} \tag{1}$$

$$R^2 = \frac{N\sum_{i=1}^{N} o_i p_i - \sum_{i=1}^{N} o_i \sum_{i=1}^{N} p_i}{\sqrt{N\sum_{i=1}^{N} o_i^2 - \left(\sum_{i=1}^{N} o_i\right)^2}\sqrt{N\sum_{i=1}^{N} p_i^2 - \left(\sum_{i=1}^{N} p_i\right)^2}} \tag{2}$$

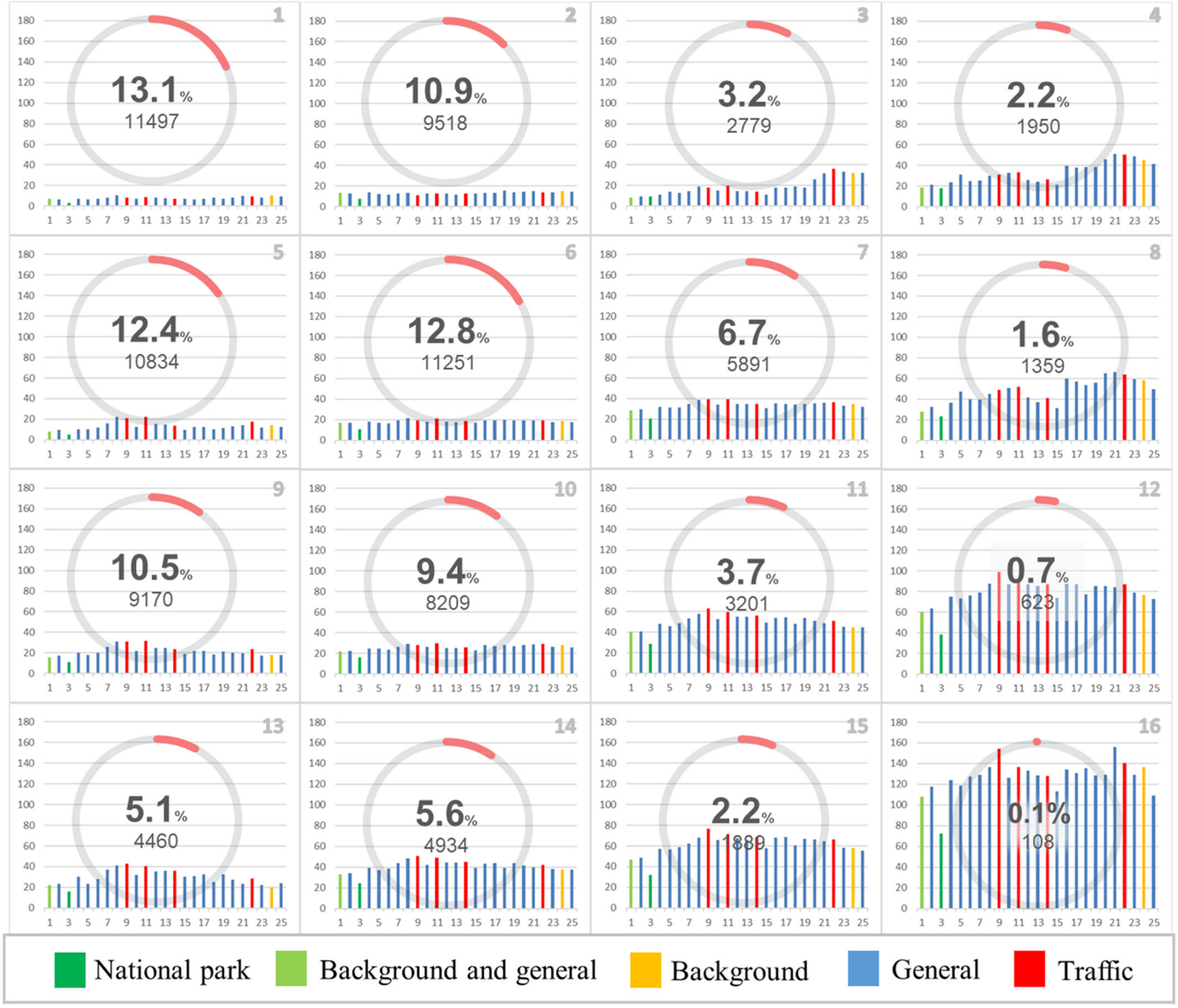


Fig. 4. Main features (number & ratio of datasets, PM2.5 at 25 stations) in the configured 4*4 SOM topological map.

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(o_i - p_i)^2}{\sum_{i=1}^{N}(o_i - \overline{o})^2}, \mathrm{NSE} \le 1 \quad (3)$$

where o_i is the observed data, p_i is the predicted value, $\overline{o}$ is the average of the observed data, and N is the number of the observed data.

It is clear from the definitions of these metrics that a model is considered to perform better if it has higher NSE and R^2 values while lower RMSE values than the other comparative model(s).

4. Results and discussion

We collected a total of 87,674 datasets of hourly PM2.5 concentrations at 25 monitoring stations in northern Taiwan. The spatio-temporal analysis of PM2.5 concentrations is conducted using the SOM to classify the datasets into a visible topological map. The high pollution events are further investigated to make multi-step-ahead prediction of PM2.5 concentration using the BPNN model.

4.1. Spatio-temporal analysis of long-term regional PM2.5 concentrations

The SOM was used to classify the datasets into a visible topological map. Various topological maps (e.g., 3*3, 4*4, and 5*5) of the SOM were explored, and their suitability was evaluated. The main common points of these three topological maps were two-fold. Firstly, it was apparent that the lowest concentration occurred in the upper-left neuron while the highest concentration occurred in the bottom-right neuron, indicating all the three topological maps were well trained and meaningful. Secondly, the comparative analysis showed that the 4*4 network not only could explain the extrema of air quality distribution in more detail than the 3*3 network but also could produce more distinct clusters than the 5*5 network. The reasons were that air quality distributions in the 9 neurons of the 3*3 network failed to fully present the deviation of total datasets whereas air quality distributions associated with some of the neurons in the 5*5 network were not distinct from one another. As a result, the 4*4 topology had the best performance, in terms of suitability and interpretability of the constructed topology. Therefore, 4*4 was determined as the map size the most suitable to meaningfully interpret the PM2.5 data collected in this study. Fig. 4 shows the analytical results of the 4*4 topological map, including (1) the number of datasets clustered in each neuron, (2) the ratio of the datasets in each neuron to the total datasets (red part of the outer circle), and (3) a bar chart of mean PM2.5 concentrations at 25 stations arranged from left to right in representative of the station distribution from northeast to southwest. Each neuron contains the information of clustered datasets (e.g., the 1st neuron has 11,497 datasets while the 16th neuron contains 108 datasets). The topological map nicely illustrates the patterns of PM2.5 concentrations from the top-left corner to the bottom-right corner, which expresses a significant increasing trend of PM2.5 concentrations along the diagonal. The regional air quality, in general, is good because a large number of datasets with low PM2.5 concentrations are clustered in upper-left neurons (for instance, 49.2% of total datasets are clustered into the1st, 2nd, 5th, and 6th neurons), whereas only a few datasets with high PM2.5 concentration are clustered into lower-right neurons (for instance, 0.1%, i.e., 108 datasets, of total datasets are clustered in the 16th neuron). The results also present that the national park station always has the lowest PM2.5 concentration while traffic stations usually show high PM2.5 concentration in each neuron.

To catch a picture of regional PM2.5 distribution, the Kriging spatial interpolation method was applied to drawing a map of PM2.5 distribution in each neuron of the SOM. Fig. 5 shows the two-dimensional topological map of the configured SOM, where the weight values of each neuron are transformed to obtain the spatial characteristic map of the neuron using the Kriging method. It appears that the lowest and the highest regional PM2.5 concentrations occur in the 1st and the 16th neurons, respectively. Besides, regional PM2.5 concentrations gradually increase from upper-left neurons (green) to lower-right neurons (red). It is easy to tell that PM2.5 concentration is more serious (higher) in the southwestern area (Taoyuan) than in the northwestern area (Keelung, Taipei). Besides, PM2.5 concentration also appears more serious in lower right neurons than the others.

Fig. 6 summaries the temporal features of regional PM2.5 concentrations during 2007–2017 in the study area, including 1) the clustered datasets in each neuron at three temporal scales (annual, seasonal, and daily), 2) the ratio of data in each neuron to the total data, and 3) averaged PM2.5 concentrations in each neuron. According to the air quality index of the TW_EPA, it appears that the averaged PM2.5 concentrations (weights) of the 1st, 2nd, 3rd, 5th, and 6th neurons (accounting for 52.4% of the entire datasets) are <20 μg/m^3, leading to good air quality (color indication: green, and yellow). In contrast, the weights of the11th, 12th, 15th, and 16th neurons (accounting for 8% of the entire datasets) exceed 50 μg/m^3, showing unhealthy air quality to human health.

4.1.1. Annual perspective

According to the bar charts (yearly scale, 2008–2017) in each neuron shown in Fig. 6, it is easy to tell that PM2.5 concentrations were much better (lower) during 2013–2017 than during 2008–2012. For those neurons contained very high PM2.5 concentrations, e.g., the 8th, 11th, 12th, 15th, and 16th neurons, most of the data were recorded before 2012. On the other hand, most of datasets in the 1st and 2nd neurons, which had the lowest averaged PM2.5 concentrations, were recorded after 2012. The reason for such pollution mitigation could be the issuance of an air quality control policy by the TW_EPA in 2012, which did affect (mitigate) PM2.5 concentrations apparently.

4.1.2. Seasonal perspective

The results of seasonal variations clearly indicate that PM2.5 concentrations are significantly worse in spring and winter than in summer and autumn, where spring and winter dominate lower-right neurons (i.e., 11th, 12th, 15th, and 16th) while summer and autumn dominate upper-left neurons (i.e., 1st, 2nd, 5th, and 6th). PM2.5 concentrations in spring and winter could much exceed the PM2.5 standard (35 μg/m^3) set by the TW_EPA. Taking the 16th neuron that has the highest averaged PM2.5 concentrations (126.8 μg/m^3) as an example, it contains datasets associated only with winter and spring. The significant seasonal variations of the measured PM2.5 concentrations result from different pollution sources as well as climatic and meteorological conditions, which suggests further controls of PM concentrations are needed, especially in winter and spring. This results are consistent with previous studies that declared the frequency of PM2.5 pollution was the highest in spring and winter in Taiwan (Fu et al., 2014; Yang et al., 2016). The background aerosol concentration was the highest in cold seasons (late winter to early spring) in northern Taiwan, which was significantly affected by the air-flow speed because high speed surface winds corresponding to high air flows could transport the aerosol from China to Taiwan. This could be that as cold high-pressure systems originating from Siberia move southward, the peripheral circulation usually transports the Asian haze to downstream areas, such as Korea, Japan, and Taiwan (Zhang et al., 2015). A recent study also indicated that high aerosol loadings observed over northern Taiwan could be associated with long-range transported dust particles and anthropogenic pollutants from the Asian Continent as well as local anthropogenic emissions (Hung et al., 2019). In contrast, air quality is better in summer in Taiwan owing to better air diffusion conditions (Wu et al., 2019). These seasonal patterns can be explained by the differences in the meteorological conditions and in the strength of the aerosol sources. Our analytical results provide more evidences to support these findings.

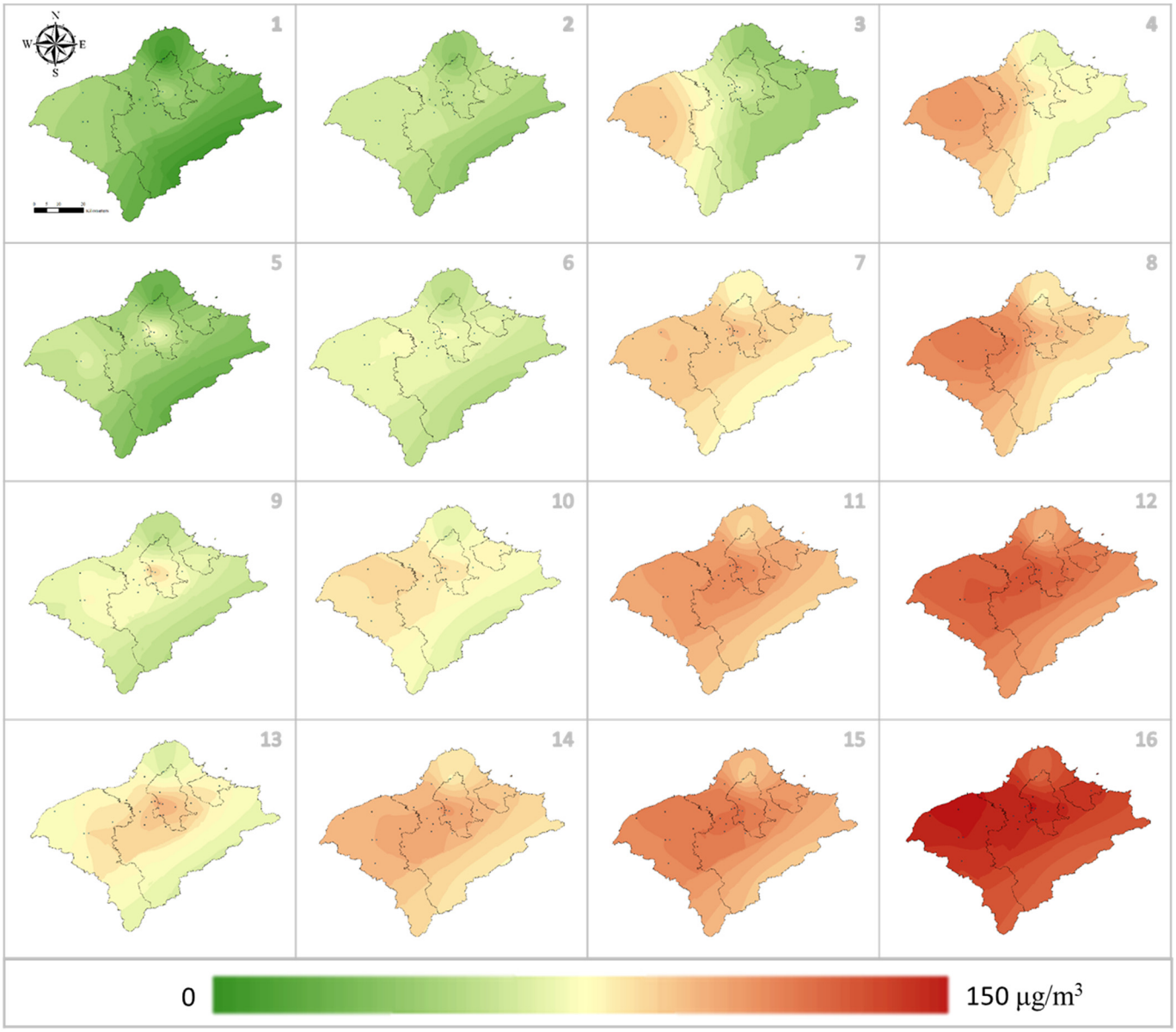


Fig. 5. SOM topological map presenting the two-dimensional spatial distribution of PM2.5 concentrations in the study area.

4.1.3. Daily perspective

There are studies exploring the differences in PM2.5 concentrations between day and night. For instance, Ye et al. (2017) presented the day-to-night mass ratios of some selected chemical species in PM2.5 and highlighted the dominant contribution of secondary processes to the major aerosol components in Changzhou, China. Also, Ge et al. (2017) indicated the enhancement of secondary aerosol formation was likely a dominant cause for the increase of PM2.5 concentrations during daytime. Pérez-Ramírez et al. (2012) revealed the increases of the fine mode radius and of the fine mode contribution to aerosol optical depth (AOD) during nighttime and explained the variations by the changes of the local aerosol sources and by the meteorological conditions between daytime and nighttime, as well as aerosol aging processes.

To gain more insights into the changes in PM 2.5 concentration between day and night and to better understand regional PM2.5 pollutant dynamics from ground-based observations, three time periods (i.e., 8-hour period; 0–7, 8–15, 16–23) in a day are clearly illustrated and analyzed based on the long-term hourly monitoring datasets clustered by the SOM. The results of the three-tier bar in each neuron could be summarized as follows.

(1) Inconsistent patterns of the ratios corresponding to three periods were observed between the neurons on the top (i.e., 1st, 2nd, 3rd, and 4th) and the neurons on the bottom (i.e., 13th, 14th, 15th, and 16th);
(2) For those neurons with the highest averaged PM2.5 concentrations (i.e., 12th and 16th neurons), high concentrations occurred more frequently during the period of 16–23 (from evening to midnight) than during the period of 0–7 (midnight to early morning); and
(3) For those neurons with the lowest averaged PM2.5 concentrations (i.e., 1st and 2nd neurons), low concentrations occurred most likely during the period of 0–7.

These results suggest that there may exhibit a relationship between PM2.5 concentration and human activities, while the

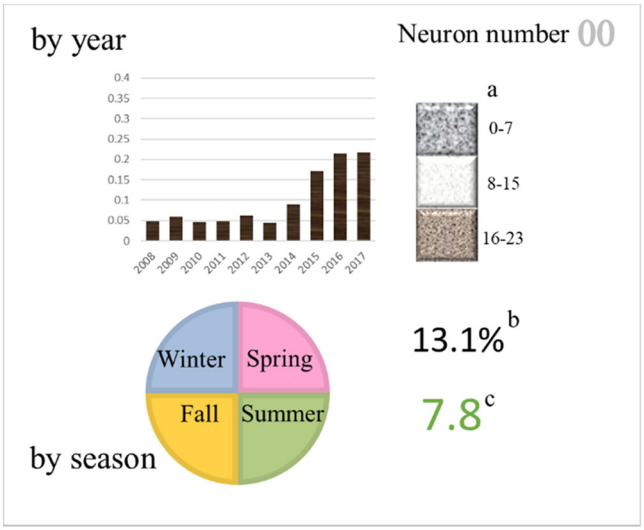


Fig. 6. Main features (ratio) of PM2.5 concentrations at yearly, seasonal and daily scales in each cluster (2008–2017).

observed day-to-night inconsistency may also be caused by different wind directions that transport air masses from local emission sources during daytime and nighttime as well as by long-range transported dust particles and anthropogenic pollutants from the China. Thus, it is necessary to take a more comprehensively spatio-temporal analysis coupled with human activities and emission

sources to figure out the complex nonlinear air pollution phenomena.

The analysis of the clustering results (i.e., the topological map) indicate that the SOM can very effectively extract and visibly represent the spatio-temporal features from the regional long-term daily monitoring datasets. Besides, the SOM can assist in the skillful summarization and visualization of the yearly trend, seasonal effect, and daily variation of long-term regional PM2.5 distributions.

4.2. Prediction of high PM2.5 concentrations

According to the air quality standard defined by the TW_EPA, air quality above 35 μg/m^3 is unhealthy for sensitive populations. Thus, we are more concerned about high-pollution events. According to the SOM clustering results of PM2.5 concentrations, high pollution datasets are clustered into the 8th, 11th, 12th, 14th, 15th, and 16th neurons, as shown in Fig. 7. There are a total of 12,114 datasets in those neurons (2008–2017). We next explore the causes of PM2.5 pollution and establish a prediction model, especially for those high PM2.5 events in recent years (2015–2017). After further inspection, there were only 2040 datasets of 193 high PM2.5 events recorded during 2015–2017, accounting for about 16.8% of the total (12,114) datasets of high pollution events over 10 years (2008–2017). We notice that all the datasets clustered in the 16th neuron (the highest averaged PM2.5 concentrations) do not fall within the period of 2015–2017.

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

Fig. 7. Clustered neurons of SOM for high PM2.5 concentrations (brown color). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 8 shows the number of occurrences of pollution events along the prolonged period at an hourly scale for the 2040 h (datasets) of 193 high pollution events. We observed that most of the prolonged periods were <12 h and only 8 events lasted for >24 h. We noticed from those high pollution events that most of the northern Taiwan exposed to poor air quality and the long-duration large-scale air pollution was caused mainly by transboundary (overseas) air pollution. During those smog periods, the northern Taiwan commonly suffered serious air pollution.

Fig. 9 shows the hourly PM2.5 concentrations at four different types of air quality monitoring stations (Yangming, Dayuan, Yonghe, and Zhongshan) for a long-duration high-pollution event that occurred during 2016/2/5 23:00 and 2016/2/7 3:00, where the horizontal axis on the top represents the neuron number of the constructed SOM. As shown, the 12th neuron (purple color) had the highest averaged PM2.5 concentrations while the 15th neuron (red color) had the second-highest ones. We noticed that the smog from China was notified to the public by the TW_EPA on 2016/2/5, coinciding with the occurrence of this high pollution event. In this high pollution incident, the peak concentration was recorded from 14:00 to 17:00 on 2016/2/5 and the highest concentration that reached 121 μg/m^3 appeared at Station Zhongshan.

Based on the occurrences of high-pollution events, we next assessed the factors affecting PM2.5 concentrations for further investigation and modeling. The Gamma Test was used to select the crucial variables for modeling PM2.5 concentrations. The noise estimation values generated by all possible input combinations were evaluated, and the input combination that produced the smallest noise was selected as the optimal input combination for the prediction model. The factors affecting PM2.5 concentrations at different time horizons were also identified to establish multi-step-ahead prediction models (i.e., T + 1, T + 4, and T + 8) for PM2.5 concentrations. The results of factor selection using the Gamma Test indicated that the important factors affecting PM2.5 concentration were temperature, relative humidity, wind speed, ozone and PM10 in general. Based on the Gamma Test results, various BPNN models were constructed (denoted as Model 1), where input variables consisted of the five selected factors and PM2.5 at the current time. Fig. 10 shows the network architecture diagram of Model 1. The data of high pollution events clustered in the 8th, 11th, 12th, 14th, and 15th neurons were used for model training and testing. A total of 2720 datasets collected during 2015–2018 were used, where 2040 datasets (2015–2017) were for training and the remaining 680 datasets (2018) were for testing.

In practice, it is common to establish a prediction model with all monitored variables as inputs based on full observational datasets. Therefore, we also established a BPNN model with inputs of all the 18 variables (Table 2) monitored at 25 stations based on full observational datasets, denoted as Model 2 (Fig. 11) serving as a benchmark. A total of 35,065 datasets collected during 2015 and 2018 were used, where

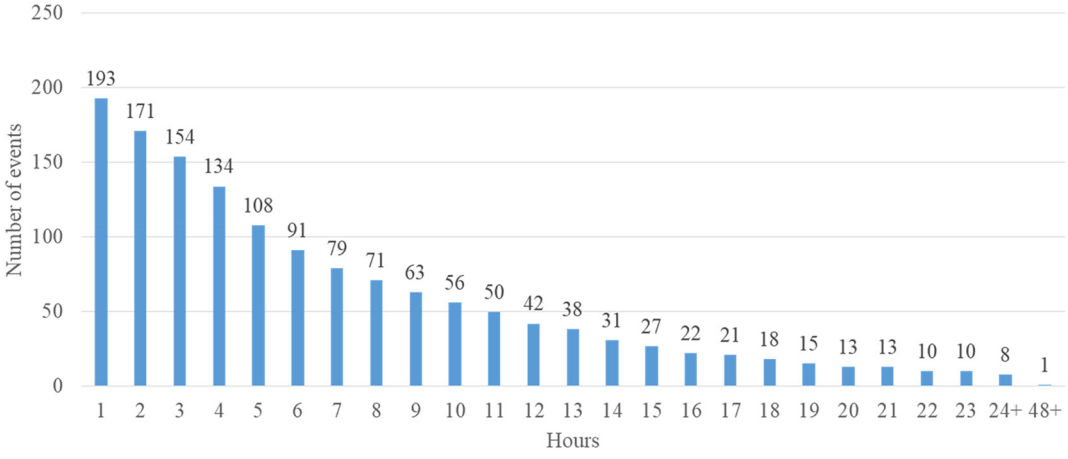


Fig. 8. Statistics of the prolonged period of 193 high pollution events (2040 hourly datasets).

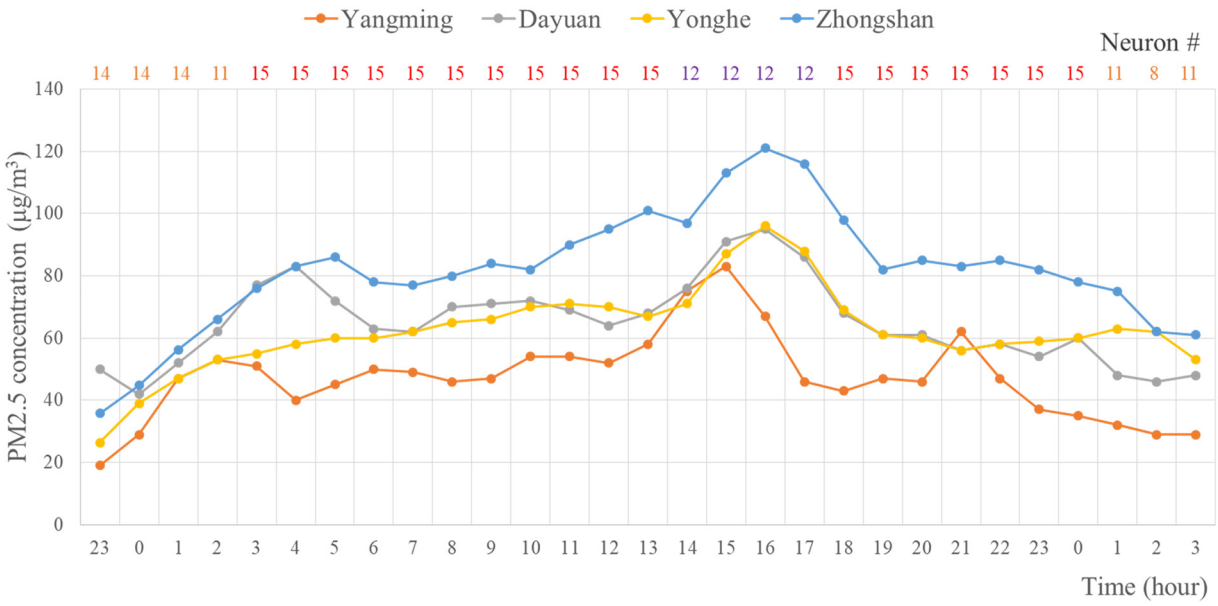


Fig. 9. Time series of a high PM2.5 concentration event occurred during 2016/2/5 23:00 and 2016/2/7 03:00 and the corresponding neurons of the constructed SOM at four stations in northern Taiwan.

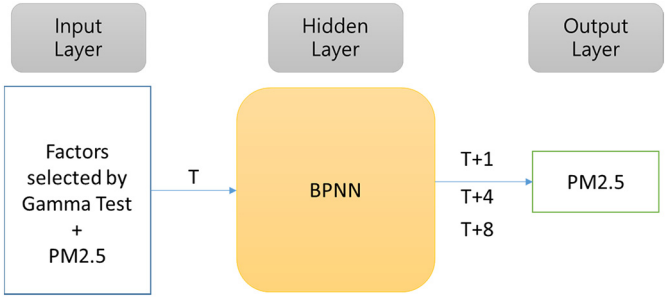


Fig. 10. Model 1 with BPNN architecture based on the Gamma Test results for multi-step-ahead PM2.5 prediction.

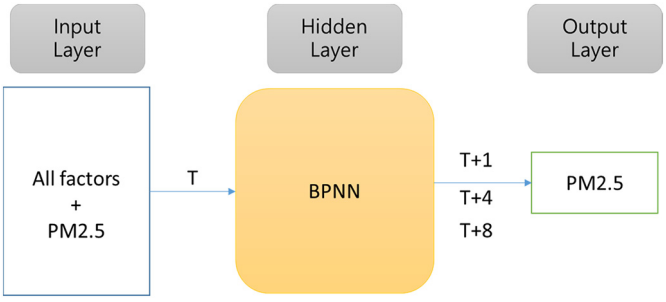


Fig. 11. Model 2 with BPNN architecture based on all 18 input variables for multi-step-ahead PM2.5 prediction.

26,305 datasets (2015–2017) were for training and 8760 datasets (2018) were for testing. We expected the ability of Model 2 in predicting high pollution events would be inferior because the number of data of high pollution events accounted for 7.8% of full observational datasets only. Therefore, we particularly extracted from Model 2 results the predicted values of high pollution events, denoted as Model 2[a], for comparison purpose. The performance of Models 1 and 2 is shown in Table 3. It appears that the performances of the two models in both training and testing cases dramatically decreases (much smaller R^2 and NSE as well as much larger RMSE values) as the prediction horizon increases. The results show that Model 2 has the best performances (higher R^2 and NSE as well as smaller RMSE values) in all the cases. However, this is mainly because it uses a large number of datasets for training and testing, where the great portion of small PM2.5 concentrations significantly reduces RMSE values as well as increases R^2 values.

Table 2
Monitored variables at 25 air quality monitoring stations.

#	Item	Units
1	SO_2	ppm
2	CO	ppb
3	O_3	ppm
4	PM_{10}	μg/m³
5	$PM_{2.5}$	μg/m³
6	NO_x	ppb
7	NO	ppb
8	NO_2	ppb
9	THC	ppm
10	NMHC	ppm
11	CH_4	ppm
12	Temperature (TEMP)	°C
13	Rainfall	mm
14	Relative humidity (RH)	%
15	Wind_Speed (WS, instantaneous value)	m/s
16	Wind_Direct (WD, instantaneous value)	Degrees
17	WS_HR (hourly average)	m/s
18	WD_HR (hourly average)	Degrees

Table 3
Comparison of PM2.5 prediction by Model 1 and Model 2.

Model	Training (2015–2017)			Testing (2018)		
	R^2	RMSE	NSE[b]	R^2	RMSE	NSE
T + 1 (Model 1)	0.78	6.09	0.63	0.76	6.40	0.61
T + 1 (Model 2)	0.85	4.81	0.70	0.85	4.66	0.71
T + 1 (Model 2[a])	0.72	6.52	0.61	0.68	7.23	0.56
T + 4 (Model 1)	0.36	10.57	0.36	0.32	10.42	0.37
T + 4 (Model 2)	0.58	7.75	0.51	0.58	7.55	0.52
T + 4 (Model 2[a])	0.28	11.65	0.30	0.30	11.46	0.31
T + 8 (Model 1)	0.24	11.29	0.32	0.17	12.25	0.26
T + 8 (Model 2)	0.42	9.27	0.42	0.41	9.25	0.42
T + 8 (Model 2[a])	0.16	13.40	0.19	0.11	14.32	0.14

[a] Results correspond to the predicted values of high pollution events only.
[b] Nash-Sutcliffe model efficiency coefficient.

Not surprisingly, Model 2[a] is inferior to Model 2. Besides, Model 1, a tailored-made model for high pollution events, indeed outperforms (higher R2 and lower RMSE values) Model 2[a] in all the cases.

To further demonstrate the reliability and applicability of the constructed models, the multi-step-ahead prediction results of Model 1 and Model 2 together with the monitored PM2.5 concentrations (real values) for a high pollution event recorded at Station Yonghe are shown in Fig. 12. Fig. 12(a) shows that predictions obtained from both models at T + 1 are quite close (small errors) to the real values. It is worth noting that Model 1 (the yellow line) can predict the peak value more accurately and its pattern is similar to the observed (real) time series while Model 2 (the gray line), however, has a delay phenomenon in prediction. Fig. 12(b) and (c) show that the prediction results of the two models at T + 4 and T + 8 do match the observational trend but fail to accurately predict high (peak) PM2.5 concentrations. Besides,

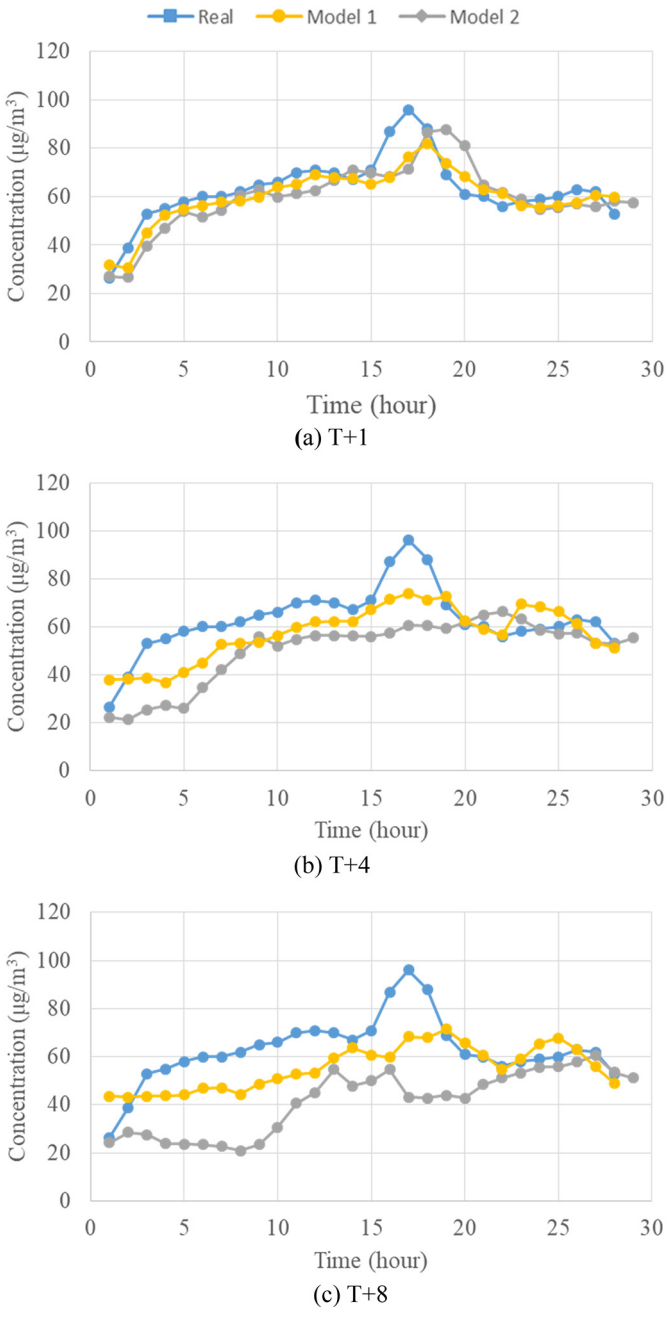


Fig. 12. Performance comparison of Model 1 and Model 2 at horizons T + 1, T + 4, and T + 8 in the testing stages at Station Yonghe.

Model 2 in general underestimates PM2.5 concentrations and has larger prediction errors than Model 1. As for the time to peak error of the employed models shown in Fig. 12, the time shift of peak concentration generally occurs for both models. For instance, there is a 1-hour delay for Model 1 while 2-hour delay for Model 2 at T + 1. As for T + 8, Model 1 has a 2-hour delay whereas Model 2 has a 10-hour delay. Both models under estimated peak concentrations in all the cases. However, Model 1 performs better than Model 2 at peak concentrations by producing smaller prediction errors at all three horizons. Therefore, it is worth establishing a specific prediction model during the occurrence of each high pollution event, a crucial condition gaining much more concerns.

5. Conclusions

This study used machine learning techniques to explore the complex spatio-temporal PM2.5 features based on a large number of high-dimensional (25 stations) hourly monitored PM2.5 concentrations in northern Taiwan. We demonstrated the SOM could skillfully encode the high-dimensional structure into a two-dimensional feature map to form a "topology" that could cluster similar features of regional PM2.5 concentrations in the constructed map to extract the complex spatio-temporal PM2.5 features. The major findings based on the constructed topology of the SOM are two-fold.

1) The spatio-temporal interrelationships of PM2.5 concentrations could be visually displayed in the SOM neurons. The designed graphs can effectively summarize all the averaged PM2.5 concentrations associated with 25 monitoring stations at various time scales in each neuron of the constructed SOM. Therefore, we could explore the yearly trend, seasonal effect, and daily variation of the regional PM2.5 concentrations to visually examine the spatio-temporal interrelationships of PM2.5 concentrations among different monitoring stations at various time scales.
2) In the SOM topological map, each cluster has its own spatial and temporal relationships. The temporal behavior of PM2.5 concentrations showed that the annual pollution trend did improve from 2007 to 2017, where most of the high pollution data were recorded before 2012; seasonal variations indicated PM2.5 concentrations were significantly worse in spring and winter than in summer and autumn; and pollution variations for three dayparts (hours 0–7, 8–15, and 16–23) were related to human activities.

The high pollutant events clustered in the neurons of the SOM were further investigated to demonstrate their usefulness and benefit in modeling multi-step-ahead PM2.5 prediction. Two BPNN models (Model 1 with 6 key input variables, and Model 2 with 18 input variables) were established to predict PM2.5 concentrations at time horizons T + 1, T + 4, and T + 8. The datasets of the high pollutant events extracted from the SOM clustering results were used to train and test Model 1. Model 2 served as a benchmark, where all the datasets collected during 2015 and 2018 were used to train and test the model. The results indicated Model 1 outperformed Model 2 in all the cases based on high pollution datasets, which provided the extra usefulness (benefit) of the SOM clustering results in modeling multi-step-ahead PM2.5 prediction for high pollution events. We conclude that the SOM results of the spatio-temporal analysis can offer the characteristics of pollution variation from point to regional scales and pollution control strategies can be formulated more effectively for specific regions, providing a useful reference for air pollution management.

CRediT authorship contribution statement

Fi-John Chang: Funding acquisition, Methodology, Project administration, Supervision, Writing - review & editing. **Li-Chiu Chang:** Methodology, Project administration, Resources, Supervision. **Che-Chia**

Kang: Data curation, Formal analysis, Software, Validation, Writing - original draft. **Yi-Shin Wang:** Data curation, Formal analysis, Writing - original draft. **Angela Huang:** Data curation, Visualization, Writing - original draft.

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.

Acknowledgement

This study is supported by the Ministry of Science and Technology, Taiwan (MOST: 108-2119-M-002-017-A). The datasets provided by the Environmental Protection Administration of Taiwan are acknowledged.

References

Alimissis, A., Philippopoulos, K., Tzanis, C.G., Deligiorgi, D., 2018. Spatial estimation of urban air pollution with the use of artificial neural network models. Atmos. Environ. 191, 205–213.
Aristodemou, E., Boganegra, L.M., Mottet, L., Pavlidis, D., Constantinou, A., Pain, C., ... ApSimon, H., 2018. How tall buildings affect turbulent air flows and dispersion of pollution within a neighbourhood. Environmental pollution 233, 782–796.
Basagaña, X., Triguero-Mas, M., Agis, D., Pérez, N., Reche, C., Alastuey, A., Querol, X., 2018. Effect of public transport strikes on air pollution levels in Barcelona (Spain). Sci. Total Environ. 610, 1076–1082.
Brokamp, C., Rao, M.B., Fan, Z.T., Ryan, P.H., 2015. Does the elemental composition of indoor and outdoor PM2.5 accurately represent the elemental composition of personal PM2.5? Atmos. Environ. 101, 226–234.
Callén, M.S., Iturmendi, A., López, J.M., 2014. Source apportionment of atmospheric PM2. 5-bound polycyclic aromatic hydrocarbons by a PMF receptor model. Assessment of potential risk for human health. Environ. Pollut. 195, 167–177.
Chang, F.J., Chang, L.C., Kao, H.S., Wu, G.R., 2010. Assessing the effort of meteorological variables for evaporation estimation by self-organizing map neural network. J. Hydrol. 384 (1–2), 118–129.
Chang, L.C., Shen, H.Y., Chang, F.J., 2014. Regional flood inundation nowcast using hybrid SOM and dynamic neural networks. J. Hydrol. 519, 476–489.
Chang, F.J., Tsai, Y.H., Chen, P.A., Coynel, A., Vachaud, G., 2015. Modeling water quality in an urban river using hydrological factors–data driven approaches. J. Environ. Manag. 151, 87–96.
Chang, F.J., Chang, L.C., Huang, C.W., Kao, I.F., 2016. Prediction of monthly regional groundwater levels through hybrid soft-computing techniques. J. Hydrol. 541, 965–976.
Chang, L.C., Chang, F.J., Yang, S.N., Tsai, F.H., Chang, T.H., Herricks, E.E., 2020. Self-organizing maps of typhoon tracks allow for flood forecasts up to two days in advance. Nat. Commun. 11 (1), 1–13.
Derwent, R.G., Parrish, D.D., Galbally, I.E., Stevenson, D.S., Doherty, R.M., Naik, V., Young, P.J., 2018. Uncertainties in models of tropospheric ozone based on Monte Carlo analysis: tropospheric ozone burdens, atmospheric lifetimes and surface distributions. Atmos. Environ. 180, 93–102.
Elangasinghe, M.A., Singhal, N., Dirks, K.N., Salmond, J.A., Samarasinghe, S., 2014. Complex time series analysis of PM10 and PM2. 5 for a coastal site using artificial neural network modelling and k-means clustering. Atmos. Environ. 94, 106–116.
Faigl, J., Kulich, M., Vonásek, V., Přeučil, L., 2011. An application of the self-organizing map in the non-Euclidean Traveling Salesman Problem. Neurocomputing 74 (5), 671–679.
Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., Wang, J., 2015. Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. Atmos. Environ. 107, 118–128.
Foehn, A., Hernández, J.G., Schaefli, B., De Cesare, G., 2018. Spatial interpolation of precipitation from multiple rain gauge networks and weather radar data for operational applications in Alpine catchments. J. Hydrol. 563, 1092–1110.
Fu, G.Q., Xu, W.Y., Yang, R.F., Li, J.B., Zhao, C.S., 2014. The distribution and trends of fog and haze in the North China Plain over the past 30 years. Atmos. Chem. Phys. 14 (21), 11949–11958.
Ge, X., He, Y., Sun, Y., Xu, J., Wang, J., Shen, Y., Chen, M., 2017. Characteristics and formation mechanisms of fine particulate nitrate in typical urban areas in China. Atmosphere 8, 62.
Han, J.C., Huang, Y., Li, Z., Zhao, C., Cheng, G., Huang, P., 2016. Groundwater level prediction using a SOM-aided stepwise cluster inference model. J. Environ. Manag. 182, 308–321.
He, Q., Zhou, G., Geng, F., Gao, W., Yu, W., 2016. Spatial distribution of aerosol hygroscopicity and its effect on PM2.5 retrieval in East China. Atmos. Res. 170, 161–167.
Heikkinen, M., Poutiainen, H., Liukkonen, M., Heikkinen, T., Hiltunen, Y., 2011. Subtraction analysis based on self-organizing maps for an industrial wastewater treatment process. Math. Comput. Simul. 82 (3), 450–459.
Hung, W.T., Lu, C.H.S., Wang, S.H., Chen, S.P., Tsai, F., Chou, C.C.K., 2019. Investigation of long-range transported PM2.5 events over Northern Taiwan during 2005–2015 winter seasons. Atmos. Environ. 217, 116920.
Ji, X., Yao, Y., Long, X., 2018. What causes PM2.5 pollution? Cross-economy empirical analysis from socioeconomic perspective. Energy Policy 119, 458–472.
Jones, A.J., Evans, D., Kemp, S.E., 2007. A note on the Gamma test analysis of noisy input/output data and noisy time series. Phys. D 229, 1–8.
Karaca, F., Camci, F., 2010. Distant source contributions to PM10 profile evaluated by SOM based cluster analysis of air mass trajectory sets. Atmos. Environ. 44 (7), 892–899.
Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biol. Cybern. 43 (1), 59–69.
Koncar, N., 1997. Optimisation Methodologies for Direct Inverse Neurocontrol. (PhD Thesis). Department of Computing, Imperial College of Science, Technology and Medicine, University of London.
Kow, P.Y., Wang, Y.S., Zhou, Y., Kao, I.F., Issermann, M., Chang, L.C., Chang, F.J., 2020. Seamless integration of convolutional and back-propagation neural networks for regional multi-step-ahead PM2.5 forecasting. J. Clean. Prod. 261, 121285.
Lanzaco, B.L., Olcese, L.E., Querol, X., Toselli, B.M., 2017. Analysis of PM2.5 in Córdoba, Argentina under the effects of the El Niño Southern Oscillation. Atmos. Environ. 171, 49–58.
Li, Y., Jiang, P., She, Q., Lin, G., 2018. Research on air pollutant concentration prediction method based on self-adaptive neuro-fuzzy weighted extreme learning machine. Environ. Pollut. 241, 1115–1127.
Marchetti, S., Hassan, S.K., Shetaya, W.H., El-Mekawy, A., Mohamed, E.F., Mohammed, A.M., ... Mantecca, P., 2019. Seasonal Variation in the Biological Effects of PM2. 5 from Greater Cairo. International journal of molecular sciences 20 (20), 4970.
Mishra, D., Goyal, P., Upadhyay, A., 2015. Artificial intelligence based approach to forecast PM2.5 during haze episodes: a case study of Delhi, India. Atmos. Environ. 102, 239–248.
Nash, J.E., 1970. River flow forecasting through conceptual models, I: a discussion of principles. J. Hydrol. 10, 398–409.
Newman, A.M., Cooper, J.B., 2010. AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number. BMC Bioinforma. 11 (1), 117.
Noori, R., Sabahi, M.S., Karbassi, A.R., 2010. Evaluation of PCA and gamma test techniques on ANN operation for weekly solid waste predicting. J. Environ. Manag. 91, 767–771.
Nowak, D.J., Hirabayashi, S., Bodine, A., Hoehn, R., 2013. Modeled PM2. 5 removal by trees in ten US cities and associated health effects. Environ. Pollut. 178, 395–402.
Orun, A., Elizondo, D., Goodyer, E., Paluszczyszyn, D., 2018. Use of Bayesian inference method to model vehicular air pollution in local urban areas. Transp. Res. Part D: Transp. Environ. 63, 236–243.
Park, Y., Kwon, B., Heo, J., Hu, X., Liu, Y., Moon, T., 2019. Estimating PM2.5 concentration of the conterminous United States via interpretable convolutional neural networks. Environ. Pollut. 256, 113395. https://doi.org/10.1016/j.envpol.2019.113395 113395.
Pérez-Ramírez, D., Lyamani, H., Olmo, F.J., Whiteman, D.N., Alados-Arboledas, L., 2012. Columnar aerosol properties from sun-and-star photometry: statistical comparisons and day-to-night dynamic. Atmos. Chem. Phys. 12 (20), 9719–9738.
Perrone, M.G., Gualtieri, M., Consonni, V., Ferrero, L., Sangiorgi, G., Longhin, E., ... Camatini, M., 2013. Particle size, chemical composition, seasons of the year and urban, rural or remote site origins as determinants of biological effects of particulate matter on pulmonary cells. Environmental pollution 176, 215–227.
Pisoni, E., Farina, M., Carnevale, C., Piroddi, L., 2009. Forecasting peak air pollution levels using NARX models. Eng. Appl. Artif. Intell. 22 (4–5), 593–602.
Raza, W., Kim, K.Y., 2008. Shape optimization of wire-wrapped fuel assembly using Kriging metamodeling technique. Nucl. Eng. Des. 238 (6), 1332–1341.
Salavati, N., Strak, M., Burgerhof, J.G.M., de Walle, H.E.K., Erwich, J.J.H.M., Bakker, M.K., 2018. The association of air pollution with congenital anomalies: an exploratory study in the northern Netherlands. Int. J. Hyg. Environ. Health 221 (7), 1061–1067.
Serrien, B., Verhaeghe, N., Verhaeghe, S., Tassignon, B., Baeyens, J.P., 2018. Evaluation of coordination hysteresis in a multidimensional movement task with continuous relative phase and self-organizing maps. Hum. Mov. Sci. 60, 162–174.
Sosa, B.S., Porta, A., Lerner, J.E.C., Noriega, R.B., Massolo, L., 2017. Human health risk due to variations in PM10-PM2.5 and associated PAHs levels. Atmos. Environ. 160, 27–35.
Stingone, J.A., Pandey, O.P., Claudio, L., Pandey, G., 2017. Using machine learning to identify air pollution exposure profiles associated with early cognitive skills among us children. Environ. Pollut. 230, 730–740.
Timmermans, R., Kranenburg, R., Manders, A., Hendriks, C., Segers, A., Dammers, E., ... van der Gon, H.D., 2017. Source apportionment of PM2. 5 across China using LOTOS-EUROS. Atmospheric Environment 164, 370–386.
Wu, T., Li, Y., 2013. Spatial interpolation of temperature in the United States using residual kriging. Appl. Geogr. 44, 112–120.
Wu, M.C., Hong, J.S., Hsiao, L.F., Hsu, L.H., Wang, C.J., 2017. Effective use of ensemble numerical weather predictions in Taiwan by means of a SOM-based cluster analysis technique. Water 9 (11), 836.
Wu, C.H., Tsai, I.C., Tsai, P.C., Tung, Y.S., 2019. Large–scale seasonal control of air quality in Taiwan. Atmos. Environ. 214, 116868.
Xu, Y., Ho, H.C., Wong, M.S., Deng, C., Shi, Y., Chan, T.C., Knudby, A., 2018. Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level PM2.5. Environ. Pollut. 242, 1417–1426.
Yang, Y., Liao, H., Lou, S., 2016. Increase in winter haze over eastern China in recent decades: roles of variations in meteorological parameters and anthropogenic emissions. J. Geophys. Res. Atmos. 121 (21), 13–050.
Ye, Z., Li, Q., Ma, S., Zhou, Q., Gu, Y., Su, Y., ... Ge, X., 2017. Summertime day-night differences of pm2. 5 components (inorganic ions, oc, ec, wsoc, wson, hulis, and pahs) in changzhou, china. Atmosphere 8 (10), 189.

Yu, H., Russell, A., Mulholland, J., Odman, T., Hu, Y., Chang, H.H., Kumar, N., 2018. Cross-comparison and evaluation of air pollution field estimation methods. Atmos. Environ. 179, 49–60.

Zaman, N.A.F.K., Kanniah, K.D., Kaskaoutis, D.G., 2017. Estimating particulate matter using satellite based aerosol optical depth and meteorological variables in Malaysia. Atmos. Res. 193, 142–162.

Zhang, Q., Quan, J., Tie, X., Li, X., Liu, Q., Gao, Y., Zhao, D., 2015. Effects of meteorology and secondary particle formation on visibility during heavy haze events in Beijing, China. Sci. Total Environ. 502, 578–584.

Zhang, Y., Qu, S., Zhao, J., Zhu, G., Zhang, Y., Lu, X., ... Wang, H., 2018. Quantifying regional consumption-based health impacts attributable to ambient air pollution in China. Environment international 112, 100–106.

Zhao, D., Chen, H., Li, X., Ma, X., 2018. Air pollution and its influential factors in China's hot spots. J. Clean. Prod. 185, 619–627.

Zheng, X., Xu, X., Yekeen, T.A., Zhang, Y., Chen, A., Kim, S.S., ... Huo, X., 2016. Ambient air heavy metals in PM2. 5 and potential human health risk assessment in an informal electronic-waste recycling site of China. Aerosol Air Qual Res 16 (2), 388–397.

Zhou, Y., Chang, F.J., Chang, L.C., Kao, I.F., Wang, Y.S., 2019a. Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts. J. Clean. Prod. 209, 134–145.

Zhou, Y., Chang, F.J., Chang, L.C., Kao, I.F., Wang, Y.S., Kang, C.C., 2019b. Multi-output support vector machine for regional multi-step-ahead PM2.5 forecasting. Sci. Total Environ. 651, 230–240.