IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 6, 2018, accepted March 24, 2018, date of publication March 28, 2018, date of current version April 23, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2820164

Improving the Accuracy and Efficiency of PM2.5 Forecast Service Using Cluster-Based Hybrid Neural Network Model

SACHIT MAHAJAN[1,2], (Student Member, IEEE), HAO-MIN LIU[3], TZU-CHIEH TSAI[4], AND LING-JYH CHEN[3], (Senior Member, IEEE)

[1]Social Networks and Human Centered Computing Program, Academia Sinica, Taipei 11529, Taiwan
[2]Taiwan International Graduate Program, National Chengchi University, Taipei 116, Taiwan
[3]Institute of Information Science, Academia Sinica, Taipei 11529, Taiwan
[4]Department of Computer Science, National Chengchi University, Taipei 116, Taiwan

Corresponding author: Sachit Mahajan (sachitmahajan90@gmail.com)

This work was supported by the Ministry of Science and Technology of Taiwan and Academia Sinica under Grant MOST 105-2221-E-001-016-MY3, Grant MOST 106-3114-E-001-004, and Grant AS-104-SS-A02.

ABSTRACT Information and communication technologies have been widely used to achieve the objective of smart city development. A smart air quality sensing and forecasting system is an important part of a smart city. One of the major challenges in designing such a forecast system is ensuring high accuracy and an acceptable computation time. In this paper, we show that it is possible to accurately forecast fine particulate matter (PM2.5) concentrations with low computation time by using different clustering techniques. An Internet of Things framework comprising of Airbox devices for PM2.5 monitoring has been used to acquire the data. Our main focus is to achieve high forecasting accuracy with reduced computation time. We use a hybrid model to do the forecast and a grid based system to cluster the monitoring stations based on the geographical distance. The experiments and evaluation is done using Airbox devices data from 557 stations deployed all over Taiwan. We are able to demonstrate that a proper clustering based on geographical distance can reduce the forecasting error rate and also the computation time. Also, in order to further evaluate our system, we have applied wavelet-based clustering to group the monitoring stations. A final comparative analysis is done for different clustering schemes with respect to accuracy and computational time.

INDEX TERMS Internet of Things, forecasting, smart cities, neural networks.

I. INTRODUCTION

In the recent years, a lot of effort has been made in order to develop a framework for smart cities. A smart city project can be considered as a project that integrates Information and Communication Technology (ICT) and Internet of Things (IoT) to provide better facilities, lifestyle to the people and also promote development. Lately, IoT has revolutionized the smart city initiative and IoT devices have become the technological backbone of smart cities [1]. IoT can be considered as an evolution of Internet into an omnipresent network of smart interconnected objects that sense and interact with the outside physical world [2]. The world has witnessed the problems related to degrading air quality. It has become a topic of concern all over the world. Industrial growth and urbanization have been some of the biggest reasons behind the deteriorating air quality. Environment Protection Agencies (EPA) in different countries around the world have taken initiatives to continuously monitor the air quality and keep a check on the variations in the pollutants level. Talking about smart cities, not only we need continuous air quality monitoring but we need to develop a system that accurately forecasts future air quality. There are various kind of pollutants based on human and environmental factors that get diffused in the air. One of the most important among all the pollutants is fine particulate matter whose size is 2.5 micrometers or less also known as PM2.5. These particles are responsible for causing serious health damages and can lead to respiratory problems [3]. Also, increasing air pollution has raised many concerns which are not only related to public health but also the social economies [4], [5].

In some of the previous works, the researchers have tried to solve real-world problems like rainfall-runoff modelling [6], [7] and river forecasting with low computation time [8]. Similarly, efforts have been made and many

initiatives have been taken in the area of PM2.5 monitoring in the past few years. Most of the conventional approaches which rely on having air quality monitoring stations deployed strategically and operated by environmental agencies or other environment monitoring organizations. But again, these methods are not perfect and have some drawbacks. The monitoring stations are expensive and huge. This creates problems during large-scale deployment. They can hardly be deployed at a high density. As a result, air pollution dispersion models have to be utilized in order to estimate PM2.5 concentrations in-between different stations [9]. As a consequence of it, the accuracy of these estimations is sensitive to wind direction and wind speed, terrain and distance between the nearby stations. Also, the measurement results from the conventional systems is effective in representing the well-mixed atmospheric pollution only but it can hardly represent the air quality in our living surroundings [10], [11].

Currently, there are some organizations with platforms for monitoring and inferring air quality parameters. Most of them follow the strategy which involves data collection and analysis followed by some action. Then the whole process is repeated again. The issues with data collection is the cost associated with it and the difficulties to replicate it. So what is needed is a smart system which is cost-effective and incorporate technologies which can help create a reliable system. Without any doubt, the problem can be addressed using a large scale IoT based system which includes data capturing, analysis, management and data analytics [12]. With the real-time data gathering and evaluation using the latest sensing and computational techniques, we aim to develop a system that will serve the citizens and improve the decision making of the authorities to turn a city into a smart city.

In this work, we deal with a large amount of PM2.5 data obtained from the IoT devices deployed around Taiwan. The challenging task is to have a model that is scalable and can be implemented in real-time. So, it becomes important to have a prediction model that can achieve high prediction accuracy with low computation time. We use a data-centric and grid based approach which uses real-time data from Airbox Project [13] to perform the experiments and evaluation.

The novelty and contribution of this paper is summarized as under:

1) We design a PM2.5 prediction model that combines a neural network based hybrid model and clustering techniques like grid based clustering and wavelet based clustering.
2) The proposed method is used for hourly PM2.5 prediction using real-time Airbox data obtained from monitoring devices deployed all over Taiwan.
3) A comparative analysis is performed for PM2.5 prediction with both grid based clustering and wavelet based clustering in terms of prediction accuracy and computation time.
4) Based on the comparative analysis results, the best combination of prediction model and clustering technique is selected. The model is used to provide real-time PM2.5 forecast service and the results are made publicly available online[1].

The remainder of this paper is organized as follows. Section II provides a review of the related works that have been used for air quality forecast and also we talk about the disadvantages of the earlier research works and how our work improves the forecasting. In Section III, we describe the proposed Hybrid neural network model (HNNM) with detailed explanations. We explain the different components of the model and the the task performed by each component. Also, the parameters for evaluation are explained. In Section IV, cluster based method is shown with detailed interpretation of the analyzed data results in four main areas in Taiwan. The comparison between clustering using wavelet transform and clustering using purely geometric features is also shown in this section. Furthermore, we describe our data-set with time series data analysis method in Section V. In Section VI, we explain the results obtained for different setups and analyze them in detail. We evaluate our model by doing a comparative analysis of the results obtained by implementing different clustering approaches. We try to explain the trade-off that happens between accuracy and the computation time. Finally, we summarize our conclusions and future work in Section VII.

II. RELATED WORKS

Researchers have been trying to solve real-world problems using machine learning techniques. Neural networks have been widely used to solve classification problems [14] and prediction problems [15], [16]. In one of the related works [17], Zheng *et al.* used a data based approach to perform PM2.5 prediction for the next 48 hours. They implemented a linear regression and neural network based prediction model. They considered meteorological data, weather forecast data and air quality data from the monitoring station. In [18], Grover *et al.* proposed a Deep Hybrid Model for weather forecasting. They considered the weather forecasting as a spatio-temporal data-intensive challenge. Though it didn't forecast PM2.5 but it predicts temperature, dew point and wind. They trained predictive models and combined them with neural networks to show the improved forecasting accuracy. In one of the other works [19], Lary *et al.* used a combination of remote sensing and meteorological data with the ground-based PM2.5 observations. Some of the researchers have implemented machine learning techniques on big data to perform the computation [20]. The authors have tried to address the problems due to rapid urbanization,industrial growth. They have proposed a framework that combines sensing and computation to perform data monitoring and analytics. In another recent work [21], a deep learning air prediction model was proposed which considers both spatial and temporal correlations. Zheng *et al.* [22] have used real-time air quality data, traffic flow data, human mobility data

[1] https://pm25next.lass-net.org/

from different points of interests. They have presented a semi-supervised learning approach to predict the air quality and have evaluated the model by comparing the results with other base-line models. Our work produces the accuracy at the similar level but there is a difference between the two datasets. Our approach is different and easy to implement as we focus on only historical PM2.5 value.

Some of the previous research works dealt with time series forecasting using Autoregressive Integrated Moving Average (ARIMA) models and classifiers based on neural networks [23], [24]. There are some models based on Support Vector Machines to perform the forecasting as shown by Sapankevych and Sankar [25]. It has been seen that a lot of research has revolved around using machine learning algorithms. But there are still many ways in which this technique can be exploited when it comes to PM2.5 prediction for small intervals of time. Most of the related work mentioned above forecasts PM2.5 on a daily basis or on an hourly basis. In [26], Syafei *et al.* performed prediction of pollutants like Nitrogen Dioxide (NO_2), PM10 (particulate matter 10 micrometers or less) and Ozone (O_3) concentrations for the next 30 minutes by considering spatial and temporal factors. In [27], Pires *et al.* applied a comparative analysis approach of five linear models to forecast PM10 concentration mean on a daily basis.

It has also been observed that most of the works revolve around performing air quality forecast for a particular monitoring station. As there are hundreds of monitoring stations installed, some researchers implemented different grouping algorithms to cluster the stations. In [28], Chen *et al.* used a cluster analysis approach to understand spatial PM variation in the environment. But this method required continuously looking for the optimal number of clusters which is a tedious job. Huang *et al.* [29] used cluster analysis and wavelet transform based approach to understand the characteristics of PM2.5 in a particular region. They tried to understand the regional distribution based on the wavelet features. They used the data for 13 sites and divided them into three clusters. Some of the researchers also used k-means clustering [30] to cluster cities which exhibited similar pollution characteristics. Their approach combined PM2.5 levels and medical records for patients for particular cities. They studied yearly variation in the PM2.5 levels and survival rate among the patients.

Though there have been quite a few works in the related field, still there are some areas that can be worked upon. Most of the above mentioned techniques rely on feeding some features into the model like traffic data, industrial emission etc.The features correspond to one particular location and the model is then implemented on all the remaining stations. It can be easily understood that different areas have different PM2.5 levels because on different sources of emission. So accurately performing forecasting using a generic models for all the stations is not really a feasible option.To address this issue, we introduce the concept of clustering the monitoring stations into grids based on the geographical distance between the stations. We assumed that air pollution has locality. Thus, monitoring nodes deployed in a certain location show similar behaviour because of close proximity to each other. This assumption was verified by studying the data we used. We could observe that the PM2.5 level of monitoring nodes close to each other had similar PM2.5 concentrations without large variations. And in the later stages, we try to implement wavelet based clustering to further evaluate the clustering based approach and improve the proposed system's accuracy and reduce the computation time.

III. HYBRID NEURAL NETWORK MODEL

Here we discuss in detail the Hybrid Neural Network Model (HNNM) we have used in the research to perform the prediction task. Before we describe the actual framework of the model, we will discuss about some important time-series forecasting models which are an important element of the Hybrid model.

A. AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) MODEL

An ARIMA model is a widely used model for time-series forecasting and is considered to be a robust model. Another important factor which makes this model so popular is the ease to understand and use it. For forecasting, the first step involves identifying the model which is followed by parameters estimation and then finally a diagnostic check is done. Generally, an ARIMA (p,d,q) model consists of integers p, d and q respectively. p represents the autogregressive (AR) part, d represents the integrated (I) part and q represents the moving average (MA) part. It has to be noticed that the values of all these components has to be greater than or equal to zero [31]. If there is time-series Z_t, where t denotes an integer and the time-series Z_t consists of real numbers which are dependent on the values at given time t. Th equation for an ARIMA (p,d,q) model is shown below.

$$(1 - B_s)^d (1 - \sum_{i=1}^{p} \phi_l B_s^l) Z_t = (1 + \sum_{i=1}^{p} \theta_l B_s^l) E_t \qquad (1)$$

In the above equation, B_s corresponds to the backward shift operator, ϕ_l and θ_l correspond to autoregressive and moving part parameters respectively and E_t is the error term.

ARIMA models have been implemented in a wide variety of applications. The problems related to prediction of wind speed, energy consumption etc., things that can be represented in a time-series format with sufficient data can be modelled using ARIMA technique [32].

B. NEURAL NETWORK AUTOREGRESSION (NNAR) MODEL

In the past few years, Artificial Neural Networks (ANN) have been widely used for the problems related to time-series forecasting. The important property of ANNs is that they can effectively model the complex relationships between the input variables and the output variables. In an NNAR model, the input is on the form of a lagged time-series and the output

denotes the predicted time-series value. In [33], Hyndman and Athanasopoulos described an NNAR $(p, P,k)m$ model where p and P denote the lagged seasonal and non seasonal input values respectively. k corresponds to the number of the hidden layers and m denotes the seasonality. The NNAR model consists of two basic functions. The first one is the Linear combination function. It is represented as

$$Z_t = m_t + \sum_{i=1}^{n} w_{i,t} y_t \tag{2}$$

In the above equation, m_t and $w_{i,t}$ are obtained directly from the data. y_t corresponds to the lagged time-series values. The second function is called the activation function and is shown as

$$f(Z_t) = 1/(1 + e^{-z_t}) \tag{3}$$

The activation function helps in reducing the effect of extreme inputs which makes the network robust to outliers. We have used a feed-forward neural network in this work. It is based on non-linear autoregressive mode for time-series forecasting.

C. HYBRID MODEL FOR FORECASTING

For the forecast purpose, we have implemented a hybrid model. In general, a time-series can be divided into linear as well as non-linear components. For linear components, ARIMA model works well [34] but it does not capture the non-linear components quite well. So another technique is needed that can capture non-linear components. To tackle this issue we use ANN [34]. We tested two methods for weighting the forecasts of the two contributing models. Initially, we assigned more weight to the model which had better in-sample performance. Later on we assigned equal weights to both contributing models. It was found out that the overall performance was better when both the models were assigned equal weights. In Fig.1, we show a flowchart of information flow in the hybrid model. The first step of the PM2.5 model includes setting the model requirements i.e. checking the past time-series data and making sure that the data is enough to perform the prediction. In the next step, the data is cleaned and input parameter is selected. Then training and testing data sets are selected for the model. The training parameters are adjusted as the the algorithm requirements. For e.g. depending on the configuration of the NNAR model, we define how many historical values of the input parameter would be selected. The performance of the designed model is analyzed and then the best model is selected to do PM2.5 forecasting. The evaluation of the model is done using the root mean square error (RMSE) and mean absolute error (MAE) measurements. Both the parameters are scale dependent measurements. With RMSE, it is better to observe bigger deviations and MAE is easier to interpret. In [35], Zhang provided a description about how a hybrid model can be implemented. A hybrid model

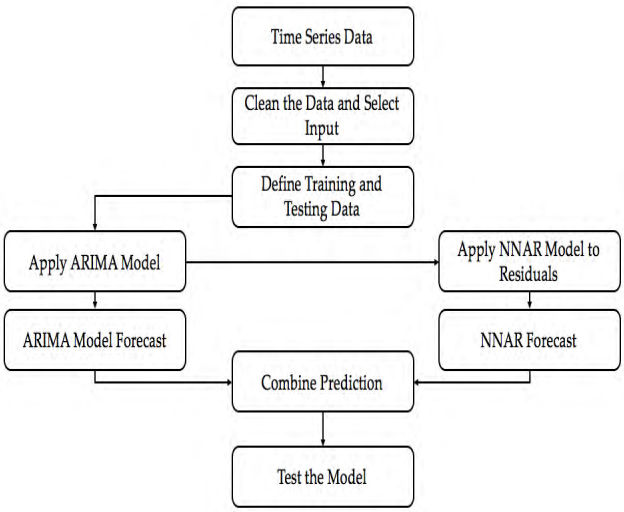

FIGURE 1. Flowchart for the hybrid model.

can be represented as

$$Z_t = X_t + Y_t \tag{4}$$

X_t corresponds to the linear components and Y_t corresponds to the non-linear components. Initially, we estimate the linear and non-linear components from the data. ARIMA model handles the linear components of the time-series data and as a result the residuals are generated which are in the form the non-linear components. If we assume r_t are the residuals at time t, then

$$r_t = Z_t - F_t \tag{5}$$

In the above equation, F_t corresponds to the forecast value for time t. Residuals can be later on modelled used the neural network. Let us assume that there are n input nodes, so the neural network can be represented as

$$r_t = f(r_{t-1}, r_{t-2}, \ldots.., r_{t-n}) + E \tag{6}$$

The function f is a non-linear function and E is the random error value. Our model uses an ARIMA (3, 1, 1) model and an NNAR (9, 5, 1) model which uses 9 lagged inputs and the hidden layer consists of 5 nodes. The parameter combination for different models was obtained after testing different setups and then selecting the parameters with the best performance.

An important part of the analysis of the model involves analyzing the scalability. The proposed monitoring infrastructure should scale with the increasing number of monitoring nodes. We started with lesser number of monitoring nodes in one particular region and then expanded our system over other major regions in Taiwan covering almost 557 monitoring stations.

IV. CLUSTER-BASED HYBRID NEURAL NETWORK MODEL

Our proposed cluster-based hybrid neural network model is shown in Figure 2. Initially, the real-time PM2.5 data from the

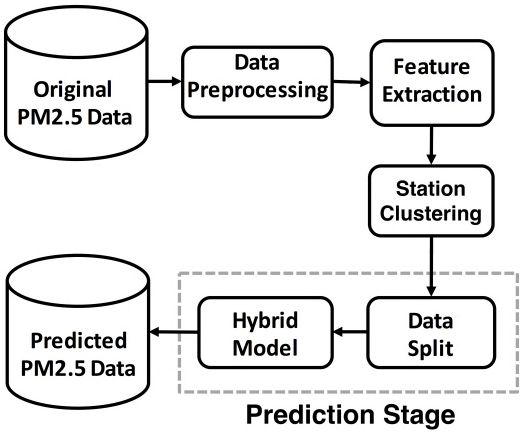

FIGURE 2. Overall architecture of the proposed cluster-based HNNM.

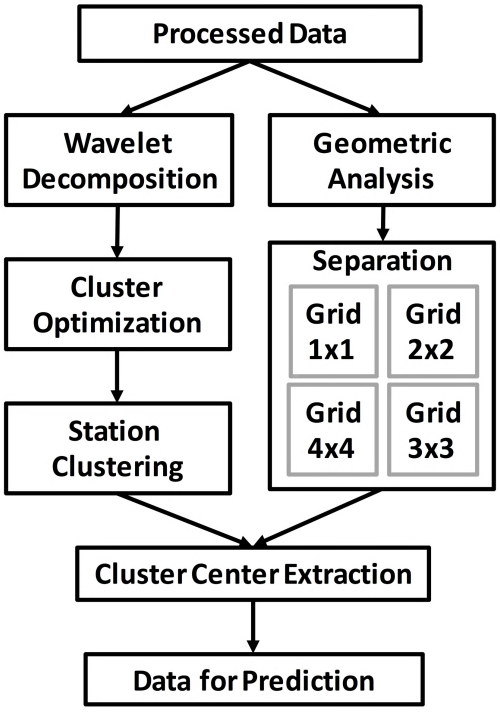

FIGURE 3. Feature extraction and station clustering flow chart.

monitoring stations is taken for analysis. The data that we get from the monitoring stations is sampled at every 5 minutes. As the sampling time is low, it is not easy to do a detailed analysis of the changes in PM2.5 level. And also due to some physical and environmental factors, some devices show a ambiguous behaviour. To tackle this problem we perform the pre-processing of the data. The data is converted into hourly data. The next two parts of the system are feature extraction and station clustering. The details of these two parts are shown in Figure 3. The processed data is then fed into two types of clustering approach. The first one is geometric-based clustering approach, which separates the stations into four kinds of grids according to their geometric distributions. The number of grids ranges from 1x1 to 4x4. The other is wavelet-based clustering approach, which firstly decomposes PM2.5 data into approximate coefficients and detail coefficients. After the decomposition, stations are clustered according to specific coefficients. Moreover, in order to obtain the best clustering result, we evaluate the result by Silhouette Evaluation approach to determine the number of clusters. At last, in cluster center extraction stage, the center of the cluster is calculated to be the input for prediction model.

The next stage involves the prediction task. First, the data is split into training and testing data set as per the requirements. Second, the hybrid model is trained using the historical PM2.5 data and later the testing is performed to check the prediction accuracy of the model. Finally, the predicted PM2.5 value is obtained and can be used to provide a service to the citizens.

A. GRID-BASED CLUSTERING APPROACH

In order to reduce the computation time of the prediction for all the monitoring stations, we applied clustering approach before implementing the prediction.

First of all, we divided all the stations into different clusters according to their geographic locations. Then, we applied the prediction model on the average value of time series data in each cluster. According to the distribution of stations in Taichung, we performed experiments to divide all the stations into one-by-one to four-by-four clusters. One-by-one case denotes that we predict the whole region with only the average value of time series data in that region, shown in Figure 4(a). In two-by-two case, we divided stations into four clusters according to the median value of their latitude and longitude, as it is shown in Figure 4(b). In three-by-three case, we divided stations into nine clusters according to the 33^{th} quantile value and 67^{th} quantile value of their latitude

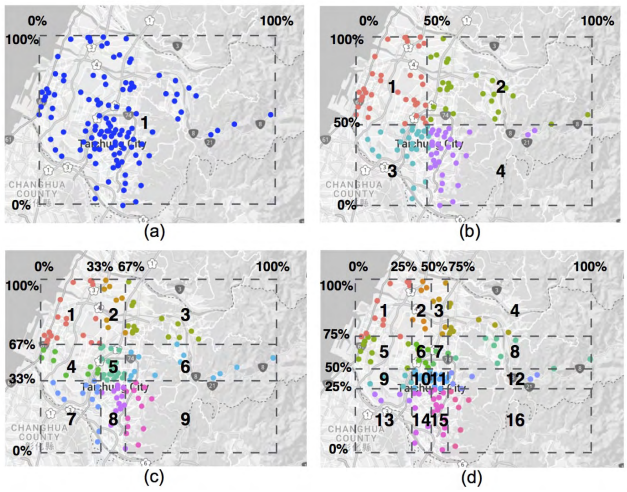

FIGURE 4. (a) Without clustering method (b) 2x2 Clustering method (c) 3x3 Clustering method (d) 4x4 Clustering method.

and longitude, as it is shown in Figure 4(c). And similarly it is done for four-by-four case as shown in Figure 4(d).

B. WAVELET-BASED CLUSTERING APPROACH

With the aim of predicting the next two hours of PM2.5, first level detail feature (D_1), which equals to one sample per two hours, is extracted to proceed with the clustering approach. Among several clustering method, Ascending Hierarchical clustering method with Euclidean as distance method and Ward's Linkage show comparably good result. The aim in Ward's method is to join cases into clusters such that the variance within a cluster is minimized.

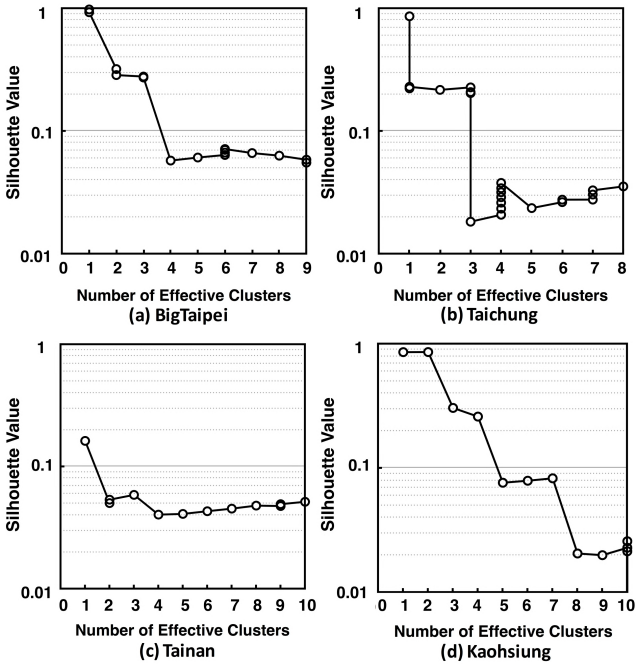

FIGURE 5. Clustering evaluation results of four major cities in Taiwan.

Besides, in order to optimize the number of clusters in each location, we applied Silhouette Evaluation, an approach for validation of consistency within clusters of data and to determine the optimized number of clusters (k). The result is shown in Figure 5. Different colors refer to different clusters in the figure. The nodes with the same color are in one cluster. In this figure, clustering evaluations from four major cities are presented. The x-axis denotes the number of effective clusters while y-axis denotes the Silhouette value. High silhouette value denotes that the cohesion of the cluster is high and the separation between the clusters is clear. In our clustering evaluation, rather than counting the number of clusters, we calculated the number of effective clusters, which are only formed from at least three stations. This approach eliminates the influence of the extreme cases. For example, if we formed five clusters with two of them containing only one station, we will count the number of effective clusters as three. Moreover, from experimental results, we discovered that those cases with silhouette value higher than 0.25 are being largely influenced by extreme cases. Therefore, our aim is to find out the largest silhouette value under 0.25. According to multiple factors, including geometry, various potential polluted sources and traffic condition, different effective number of clusters is shown different cities.

The clustering result is shown in Figure 6. The result for each area reflects some features related to the geography of that city. In Figure 6(a), the clusters of BigTaipei City (Taipei City & New Taipei City) is not clear to recognize. However, we can discern that there are two clusters being separated by Tamsui River. Figure 6(b) presents obviously three clusters in Taichung City. Moreover, these clusters are related to the natural barriers such as highway number three, central city and mountains. Figure 6(c) denotes the cluster distribution of Tainan City with clear separation according to highway number one and central city. At last, the cluster distribution of Kaohsiung City is shown in Figure 6(d). The distribution of apparent four clusters are influenced by central city and mountains. Therefore, through wavelet based clustering method, the trend and feature of PM2.5 time series data could be kept in each cluster. Besides the geometric clustering method, we applied wavelet based clustering approach to obtain higher prediction accuracy with smaller number of clusters. First, we calculated the average value of the time series data in each cluster. Second, by taking the average value as the center of each cluster, we fed this time series average value to the input of our Hybrid prediction model.

V. PM2.5 DATA ANALYSIS

In this section, we will explain in detail the Airbox Project and also elaborate the data set used for this work. Furthermore, we show PM2.5 time series data analysis in this section. First, we wanted to obtain an understanding of the underlying forces and structure that produced the observed data. Second, we tried to build a model to fit the data and proceed to forecasting.

A. AIRBOX SYSTEM

The Airbox Project which began in Taiwan started with pilot deployment of IoT systems for PM2.5 monitoring. The objective of this project is to motivate and encourage people to voluntarily participate in PM2.5 sensing. LASS (Location Aware Sensing System) community is the main source of inspiration behind this project. The community tries to engage people to participate in PM2.5 sensing and also encourage them to develop devices for sensing PM2.5 by themselves. There are many benefits of having such a sensing system. One of the important benefits is the particulate matter concentration can be monitored at a fine spatio-temporal granular level. The data can be easily accessed by people which makes analysis of data easy [36]. Sensing devices for the Airbox Project are designed and developed by professional manufacturers. During the device installation, it is made sure that the places have a regular power supply and internet connection.

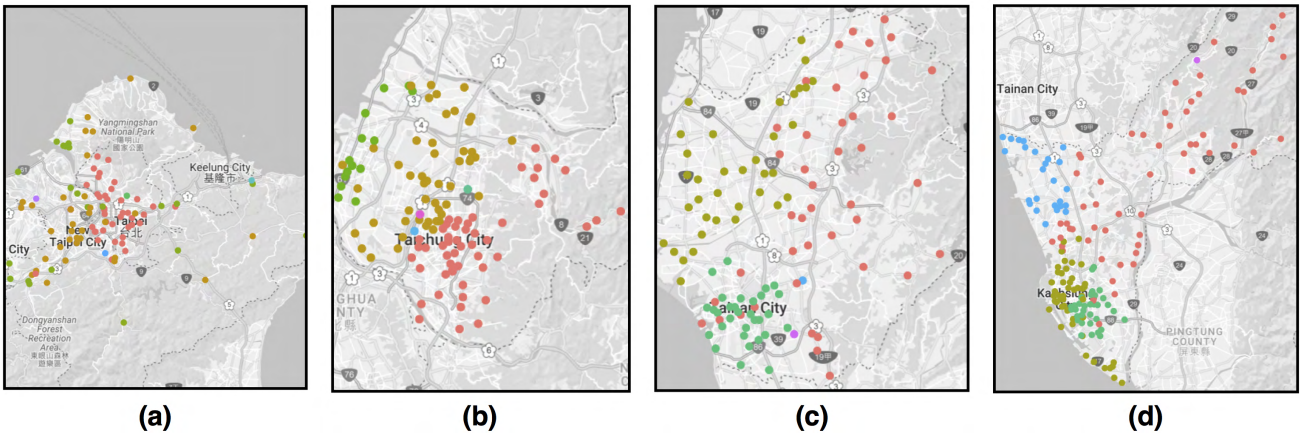

FIGURE 6. Wavelet-based clustering results of four major cities in Taiwan: (a) BigTaipei City (b) Taichung City (c) Tainan City (d) Kaohsiung City.

B. AIRBOX DATASET

For this research work, the data was collected from the Airbox devices installed all over Taiwan. The measurement data was taken for the time period between January 18, 2017 and February 17, 2017. Most of the Airbox devices are installed in elementary schools around the region with regular power connection and internet supply. This makes sure that the data which is recorded is reliable. So for accurate forecasting, we considered the data from reliable sources. Manufacturers claim that the sampling frequency for Airbox devices is every five minutes. It was found that for 80% of the devices the inter-sampling rate was around six minutes.For the rest, it was near around twelve minutes. Another factor to be studied is that there is a standby time between two samples collection. It is somewhere around five minutes and it takes one minute to perform the sampling. So the total time sums up to be around six minutes. And if an error occurs and the device is unable to record the first measurement, the inter-sampling time becomes twelve minutes. When the data is sampled every six minutes, it is not possible to notice a lot of variations in the PM2.5 level and also the chances to have outliers are more likely. So for this study, we converted the data into hourly data as done in most of the research works and then performed the experiments. The distribution of PM2.5 measuring stations is shown in Table 1. The average variation in each major area is shown in Figure 7.

TABLE 1. Distribution of PM2.5 measuring stations.

Location	Number of Stations
BigTaipei	110
Taichung	134
Tainan	118
Kaohsiung	195

C. TIME SERIES PM2.5 DATA

In this section, we would focus on the first aspect. As it is shown in recent studies, PM2.5 is generated from multiple factors, including factories combustion, incense burning from

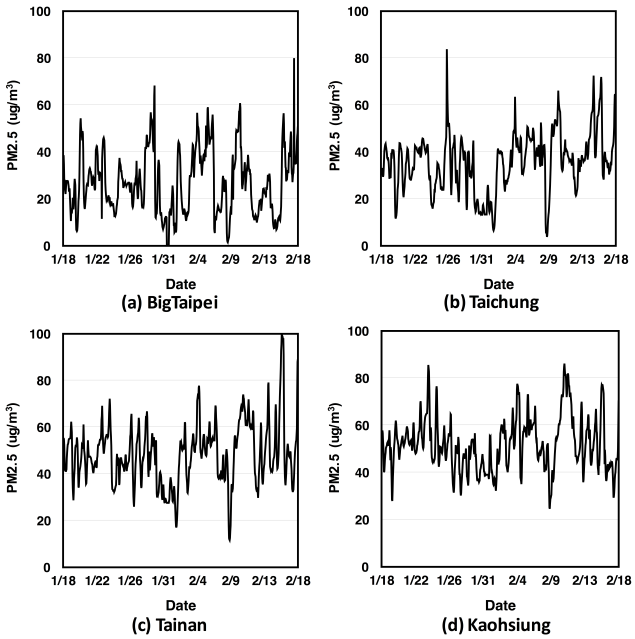

FIGURE 7. PM2.5 data variation in four major cities.

the temples, polluted air emission from vehicles etc. Therefore, it is easy to understand that there would be a fluctuation in the pattern in certain areas according to a relative strong factor. Besides, the pattern might be related to a certain period of time in a day. In order to find out underlying forces that lead to certain patterns of PM2.5 time-series data, we try to investigate PM2.5 data by time-frequency analysis.

Most common approaches that transfer data from time domain to frequency domain are Fourier transform and wavelet transform. Fourier transform can convert signal in time domain to frequency domain by integrating over the whole time axis. However, if the signal is not stationary, then the frequency composition is a function of time, we cannot tell when a certain frequency rises. Therefore, Short-time Fourier transform (STFT) is proposed to solve the problem. The window is designed to extract a small portion of the

signal and then take Fourier transform. However, there is a limitation of the frequency component detection on the spectrum because of the fixed-size window. In contrast with STFT, wavelet transform overcomes the previous problem. It is designed to strike a balance between time domain and frequency domain. Therefore, we can see very low frequency components as well as very high frequency components by this approach. The continuous wavelet transform ($CW(a, b)$) is defined by the following equation6

$$CW(a,b) = \int_{-\infty}^{\infty} f(t) \cdot \psi^*_{a,b}(t)dt \tag{7}$$

where

$$\psi_{a,b} = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{8}$$

and a denotes a scale parameter, b denotes translation parameter and ψ denotes mother wavelet.

In order to reduce the redundancy and to make the wavelet transform more practical, by defining the sampling grid $a = {a_0}^m$ and $b = b_0{a_0}^m$, we can obtain the discrete wavelet transform coefficients ($DW(m, n)$) as the following equation

$$DW(m,n) = \int_{-\infty}^{\infty} f(t) \cdot \psi^*_{m,n}(t)dt \tag{9}$$

where

$$\psi_{m,n}(t) = \frac{1}{\sqrt{a_0^m}}\psi\left(\frac{t - nb_0a_0^m}{a}\right) \quad m, n \in Z. \tag{10}$$

with a_0 as a specified fixed dilation stop parameter set at a value greater than one and b_0 as a location parameter which must be greater than zero.

D. WAVELET DECOMPOSITION

Mallat's wavelet is one of the most attractive wavelets for digital implementation, because it describes the wavelet transform in terms of digital filtering and sampling. The filtering is done iteratively using low-pass (h) and high-pass (g) components, whose sampling ratios are powers of 2. The equations denoting high-pass and low-pass filtering operation are shown in equation (11) and equation (12), respectively.

$$Y_{high}[2^j t] = \sum_k h_{j+1}[k] f[2^{j+1}t - k] \tag{11}$$

$$Y_{low}[2^j t] = \sum_k g_{j+1}[k] f[2^{j+1}t - k] \tag{12}$$

Through cascading approach, shown in Figure 8, wavelet decomposition can be accomplished. In this figure, $f[n]$, the original PM2.5 time series data is passed through a high pass as well as a low pass filter. Since half the frequencies of the signal have now been removed, half the samples can be discarded according to Nyquist's rule. The filter output of the low-pass filter $g[n]$ in the diagram above is then subsampled by 2 and further processed by passing it again through

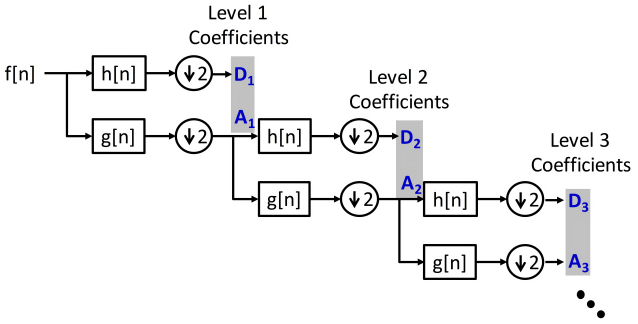

FIGURE 8. Cascading and filter banks scheme.

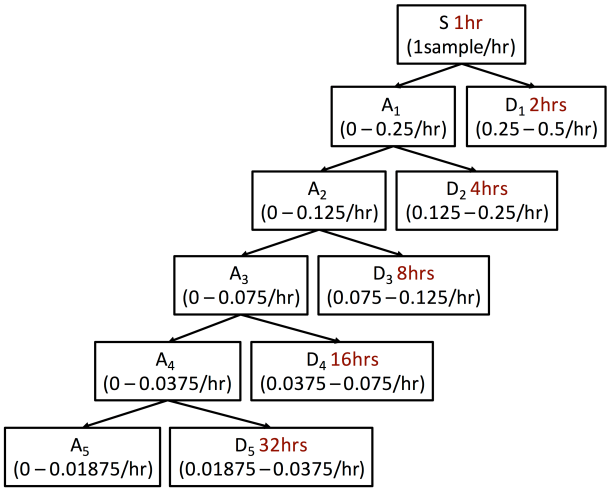

FIGURE 9. PM2.5 data wavelet decomposition scheme.

a new low-pass filter g and a high-pass filter h with half the cut-off frequency of the previous one. In each decomposition level, D denotes detail coefficient and A denotes approximate coefficient. For the simplification of wavelet decomposition calculation, we applied Haar wavelet, which is one of the simplest and most popular wavelets. In our experiment, we decomposed the PM2.5 data into 5 levels with different frequency bands. The decomposition scheme is shown in Figure 9. S denotes original PM2.5 signal with the sampling frequency (fs) equals to 1 sample per hour. After the decomposition, approximate feature ($A_1 - A_5$) and detail feature ($D_1 - D_5$) are obtained. We applied D_1 to D_5 as detail features of two hours, four hours, eight hours, sixteen hours and thirty two hours sampling rate to do following clustering processes. The result of wavelet decomposition on one of the stations is shown in Figure 10, where its approximate feature and detail features are exhibited. D_1 contains the variation with the highest frequency components, whose sampling rate equals to one sample per two hours sampling rate. D_5 contains the variation with the lowest frequency components, whose sampling rate equals to one sample per thirty two hours.

VI. RESULTS AND EVALUATION

In this section we will elaborate the results achieved with grid based clustering and wavelet based clustering. Also we will

TABLE 2. Results of computation time and relative error with grid/wavelet-based prediction.

Location	BigTaipei						Taichung					
Cluster Type*	1x1	2x2	3x3	4x4	w/o	C_6	1x1	2x2	3x3	4x4	w/o	C_3
Relative Error (%)	0.63	0.78	0.74	0.38	0.074	**0.3**	0.25	0.22	0.21	0.2	0.04	**0.2**
Computation Time (sec)	3.8	15.2	34.2	60.8	431.8	**17.34**	3.08	12.32	27.72	49.28	596.2	**9.6**
Location	Tainan						Kaohsiung					
Cluster Type*	1x1	2x2	3x3	4x4	w/o	C_3	1x1	2x2	3x3	4x4	w/o	C_7
Relative Error (%)	0.3	0.26	0.2	0.16	0.015	**0.21**	0.4	0.27	0.22	0.14	0.026	**0.23**
Computation Time (sec)	3.1	12.4	27.9	49.6	313	**8.1**	3.9	15.6	35.1	62.4	604	**19.6**
* 1x1 to 4x4: Grid-Based Clustering / C_n: n clusters under Wavelet-Based Clustering / w/o: without clustering method												

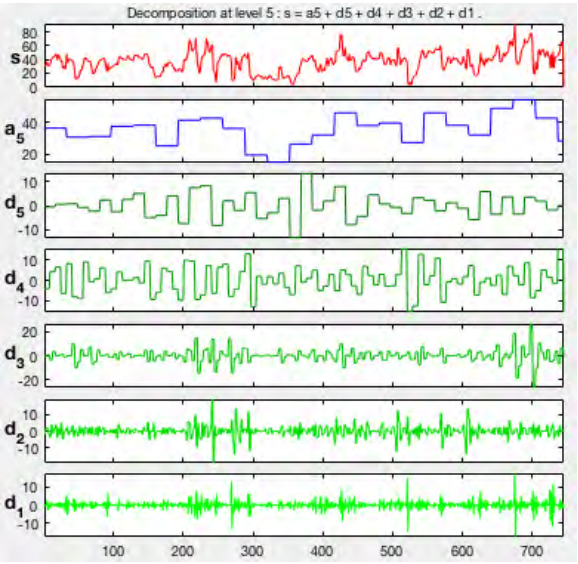

FIGURE 10. PM2.5 data wavelet decomposition result.

evaluate our system by showing how accurately the system performs the prediction when it comes to all major regions of Taiwan including all 557 monitoring stations.

A. PREDICTION RESULTS

The experiments and evaluation were done using the real-time Airbox data obtained from devices from all over Taiwan. To perform the forecast, the Hybrid Model was trained using the three weeks hourly historical data. And the next one week hourly data was used for testing the model. We first forecast the hourly PM2.5 value for all the 557 monitoring stations and compared the forecast with the actual observed values.

The parameters chosen to compare the results were relative error and computation time. For the computation time, we considered the time elapsed for completing the entire computation process for performing the prediction. Initially, we carried out our first approach which was based on clustering the monitoring stations according to the geographical distance. The results can be analyzed from Table 2. It can be observed that for all the locations, the relative error is highest when we don't cluster the stations into different grids and consider them as one cluster. Although the computation time is low but it doesn't serve our purpose as our aim is to lower the prediction error as well. Similarly, we perform the experiments for Gird 2x2, 3x3 and 4x4. The results improved significantly. We can observe that for Big Taipei the relative error reduced to 0.38 for Grid 4x4 and the computation time was 17.34 seconds. For Taichung the relative error reduced to 0.20 for Grid 4x4 and computation time was 49.28 seconds. Similarly for Tainan and Kaohsiung, the relative error dropped to 0.16 and 0.14 respectively for Grid 4x4. And the computation time was 49.6 seconds and 62.4 seconds respectively. From all these results it can be inferred that when we clustered the monitoring stations using 4x4 Grid, the error reduced significantly for all the regions and there was a slight increase in the computation time. This can be explained as a trade-off which occurs when accuracy increases and the monitoring stations are divided into many clusters.

With Grid based clustering the results we achieved were acceptable as our main objective of reducing the error rate was achieved. But in order to further reduce the computation time, we focused on application of wavelet based clustering to further improve our system and reduce the computation time. With the application of wavelet clustering on Big Taipei region, all the stations were divided into 6 clusters. The result showed significant improvement with relative error dropping to 0.30 and computation time dropped to 17.34 seconds. For Taichung we divided the stations into three clusters, the relative error dropped to 0.20 with a significant decrease in the computation time which was 9.6 seconds. For Tainan, we divided the stations into three clusters, the relative error recorded was 0.21 with computation time of 8.1 seconds. And finally for Kaohsiung, seven clusters were made. The relative error recorded was 0.23 and computation time was 19.6 seconds.

B. EVALUATION OF THE PROPOSED MODEL

To further evaluate our system, we compared the results for all four major regions of Taiwan. We focused on improving the prediction accuracy and reducing the computation time. As our objective is to use this system as a real-time application, we expected to reduce the computation time for all the regions in Taiwan to twenty seconds or lower while making sure that the prediction accuracy also stays in the acceptable region. The results shown in Figure 11 depicts that grid-based clustering approach significantly reduces the computation

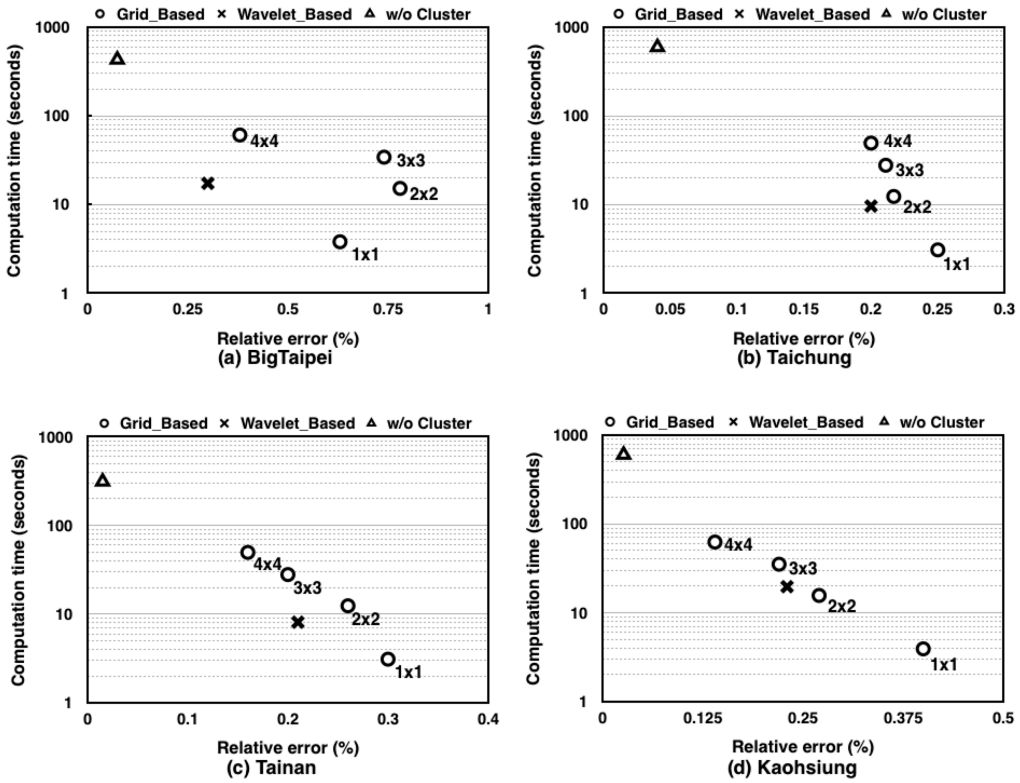

FIGURE 11. Comparison of computation time and relative error with grid/wavelet-based prediction in four areas.

time as compared to the case when we perform prediction without clustering. From Figure 11, it can be observed that for all the regions, the results obtained after wavelet-based clustering are more effective. If we do a comparative analysis of computation time between grid based clustering and wavelet based clustering approach, we can see a huge reduction in the computation time. For Big Taipei, the computation time was reduced to 17.34 seconds which is around 71% reduction from the design with Grid 4x4 based clustering. In the case of Taichung, the computation time was reduced to 9.6 seconds which is around 80% reduction from the time spent when Grid 4x4 clustering was used. For Tainan, the computation time 8.1 seconds which is around 84% less than what was achieved when Grid 4x4 clustering was used. Finally for Kaohsiung, the computation time was reduced to 19.6 seconds which is near about 69% less than what was obtained when Grid 4x4 clustering was applied. It can be inferred from all the results that the wavelet based clustering not only reduces the relative error but also the computation time is reduced significantly. We were able to achieve an acceptable trade-off between the accuracy and the computation time.

VII. CONCLUSION AND FUTURE WORK

In this paper, we proposed a framework that uses IoT technology and machine learning techniques to accurately forecast PM2.5 concentration with low computation time in Taiwan. A Hybrid model which utilizes an ARIMA model and an NNAR model was used to do the forecast. Initially, we followed a grid based method based on geographical distance to efficiently group 557 monitoring stations deployed over four major regions of Taiwan. Grid size was varied in order to understand the trade-off between the accuracy and the computation time. In the later part, wavelet based clustering was used to cluster the monitoring stations. The prediction results with wavelet based clustering showed significant improvement in the accuracy and reduction in the computation time. For all four regions, the computation time was found to be below 20 sec which was significantly low. On an average, there was 76% reduction in the computation time with wavelet-based clustering as compared to grid based clustering method. Also, we were able to significantly reduce the relative error to as low as 0.2. Based on the results, we implemented the hybrid model using wavelet based clustering for providing PM2.5 prediction as a real-time service.

There are certain limitations of this work that we would like to address in future. We would like to improve the user interface for providing PM2.5 prediction as a real-time service. Also, we would like to extended this study by testing different models and by considering other environmental factors like wind speed, wind direction and temperature. These studies can produce significant results which can be used by environmental pollution monitoring agencies for policy making.

ACKNOWLEDGMENT

The authors wish to thank Edimax Inc. and the LASS community for their support, technical advice and administrative assistance.

REFERENCES

[1] J. Jin, J. Gubbi, S. Marusic, and M. Palaniswami, "An information framework for creating a smart city through Internet of Things," *IEEE Internet Things J.*, vol. 1, no. 2, pp. 112–121, Apr. 2014.

[2] K. A. Delic, "On resilience of IoT systems: The Internet of Things (ubiquity symposium)," *Ubiquity*, vol. 2016, no. 2, p. 1, Feb. 2016.

[3] Y.-F. Xing, Y.-H. Xu, M.-H. Shi, and Y.-X. Lian, "The impact of $PM_{2.5}$ on the human respiratory system," *J. Thoracic Disease*, vol. 8, no. 1, pp. 69–74, Jan. 2016.

[4] X. Tang, "An overview of air pollution problem in megacities and city clusters in China," in *Proc. AGU Spring Meeting Abstracts*, May 2007.

[5] VUFO—NGO Resource Centre Vietnam. (Sep. 19, 2013). *Vietnam Named Among Top Ten Nations With Worst Air Pollution*. [Online]. Available: http://www.ngocentre.org.vn/news/vietnam-named-among-top-ten-nations-worst-air-pollution

[6] K.-W. Chau, "Use of meta-heuristic techniques in rainfall-runoff modelling," *Water*, vol. 9, no. 3, p. 186, 2017.

[7] W.-C. Wang, D.-M. Xu, K.-W. Chau, and S. Chen, "Improved annual rainfall-runoff forecasting using PSO-SVM model based on EEMD," *J. Hydroinform.*, vol. 15, no. 4, pp. 1377–1390, 2013.

[8] R. Taormina, K.-W. Chau, and B. Sivakumar, "Neural network river forecasting through baseflow separation and binary-coded swarm optimization," *J. Hydrol.*, vol. 529, pp. 1788–1797, Oct. 2015.

[9] M. Markiewicz, "A review of mathematical models for the atmospheric dispersion of heavy gases. Part I. A classification of models," *Ecol. Chem. Eng. S*, vol. 19, no. 3, pp. 297–314, Jul. 2012.

[10] S.-C. C. Lung, I.-F. Maod, and L.-J. S. Liu, "Residents' particle exposures in six different communities in Taiwan," *Sci. Total Environ.*, vol. 377, no. 1, pp. 81–92, May 2007.

[11] S.-C. C. Lung, P.-K. Hsiao, T.-Y. Wen, C.-H. Liu, C. B. Fu, and Y.-T. Cheng, "Variability of intra-urban exposure to particulate matter and CO from asian-type community pollution sources," *Atmos. Environ.*, vol. 83, pp. 6–13, Feb. 2014.

[12] J. Lanza *et al.*, "Large-scale mobile sensing enabled internet-of-things testbed for smart city services," *Int. J. Distrib. Sensor Netw.*, vol. 11, no. 8, p. 785061, 2015.

[13] *$PM_{2.5}$ Open Data Portal*. Accessed: Jan. 21, 2018. [Online]. Available: http://pm25.lass-net.org/en/

[14] S. Zhang and K.-W. Chau, "Dimension reduction using semi-supervised locally linear embedding for plant leaf classification," in *Proc. Int. Conf. Intell. Comput.*, 2009, pp. 948–955.

[15] V. Gholami, K. W. Chau, F. Fadaee, J. Torkaman, and A. Ghaffari, "Modeling of groundwater level fluctuations using dendrochronology in alluvial aquifers," *J. Hydrol.*, vol. 529, pp. 1060–1069, Oct. 2015.

[16] P. Sefeedpari, S. Rafiee, A. Akram, K.-W. Chau, and S. H. Pishgar-Komleh, "Prophesying egg production based on energy consumption using multi-layered adaptive neural fuzzy inference system approach," *Comput. Electron. Agricult.*, vol. 131, pp. 10–19, Dec. 2016.

[17] Y. Zheng *et al.*, "Forecasting fine-grained air quality based on big data," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 2267–2276.

[18] A. Grover, A. Kapoor, and E. Horvitz, "A deep hybrid model for weather forecasting," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 379–386.

[19] D. J. Lary, T. Lary, and B. Sattler, "Using machine learning to estimate global $PM_{2.5}$ for environmental health studies," *Environ. Health Insights*, vol. 9, no. 1, pp. 41–52, 2015.

[20] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, p. 38, 2014.

[21] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions," *Environ. Sci. Pollution Res.*, vol. 23, no. 22, pp. 22408–22417, 2016.

[22] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-Air: When urban air quality inference meets big data," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 1436–1444.

[23] C. Voyant, M. Muselli, C. Paoli, and M.-L. Nivet, "Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation," *Energy*, vol. 39, no. 1, pp. 341–355, 2012.

[24] L. Chen and X. Lai, "Comparison between arima and ann models used in short-term wind speed forecasting," in *Proc. Asia–Pacific Power Energy Eng. Conf. (APPEEC)*, 2011, pp. 1–4.

[25] N. I. Sapankevych and R. Sankar, "Time series prediction using support vector machines: A survey," *IEEE Comput. Intell. Mag.*, vol. 4, no. 2, pp. 24–38, May 2009.

[26] A. D. Syafei, A. Fujiwara, and J. Zhang, "Prediction model of air pollutant levels using linear model with component analysis," *Int. J. Environ. Sci. Develop.*, vol. 6, no. 7, p. 519, 2015.

[27] J. C. M. Pires, S. I. V. Sousa, M. C. Pereira, M. C. M. Alvim-Ferraz, and F. G. Martins, "Management of air quality monitoring using principal component and cluster analysis—Part I: SO_2 and PM_{10}," *Atmos. Environ.*, vol. 42, no. 6, pp. 1249–1260, 2008.

[28] T. Chen, J. He, X. Lu, J. She, and Z. Guan, "Spatial and temporal variations of $PM_{2.5}$ and its relation to meteorological factors in the urban area of Nanjing, China," *Int. J. Environ. Res. Public Health*, vol. 13, no. 9, p. E921, 2016.

[29] P. Huang, J. Zhang, Y. Tang, and L. Liu, "Spatial and temporal distribution of $PM_{2.5}$ pollution in Xi'an city, China," *Int. J. Environ. Res. Public Health*, vol. 12, no. 6, pp. 6608–6625, 2015.

[30] M.-A. Kioumourtzoglou, E. Austin, P. Koutrakis, F. Dominici, J. Schwartz, and A. Zanobetti, "$PM_{2.5}$ and survival among older adults: Effect modification by particulate composition," *Epidemiology*, vol. 26, no. 3, pp. 321–327, 2015.

[31] C. Christodoulos, C. Michalakelis, and D. Varoutas, "Forecasting with limited data: Combining ARIMA and diffusion models," *Technol. Forecast. Social Change*, vol. 77, no. 4, pp. 558–565, 2010.

[32] E. Cadenas, W. Rivera, R. Campos-Amezcua, and C. Heard, "Wind speed prediction using a univariate ARIMA model and a multivariate NARX model," *Energies*, vol. 9, no. 2, p. 109, 2016.

[33] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. Melbourne, VIC, Australia: OTexts, 2014.

[34] S. Mahajan, L.-J. Chen, and T.-C. Tsai, "An empirical study of $PM_{2.5}$ forecasting using neural network," in *Proc. Int. IEEE Conf. Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 2017, pp. 327–333.

[35] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, Jan. 2003.

[36] L.-J. Chen, W. Hsu, M. Cheng, and H.-C. Lee, "Demo: LASS: A location-aware sensing system for participatory $PM_{2.5}$ monitoring," in *Proc. 14th Annu. Int. Conf. Mobile Syst., Appl., Services Companion (MobiSys Companion)*, New York, NY, USA, 2016, p. 98.

SACHIT MAHAJAN received the B.Tech. degree in ECE from Punjab Technical University, India, in 2012, and the M.S. degree in communication engineering from the University of Manchester, U.K., in 2013. He is currently pursuing the Ph.D. degree in social networks and human centered computing with the Network Research Laboratory, Academia Sinica. His research interests include machine learning, data science, and Internet of Things.

HAO-MIN LIU received the B.S. and M.S. degrees in electronics engineering from National Chiao Tung University, in 2012 and 2015, respectively. He is currently a full-time Research Assistant with the Network Research Laboratory, Academia Sinica. His research interests include Internet of Things, signal processing, and data analysis.

TZU-CHIEH TSAI received the B.S. degree in electrical engineering from National Taiwan University in 1988, the M.S. degree in electrical engineering from the University of Southern California in 1991, and the Ph.D. degree in computer science from the University of California at Los Angeles, Los Angeles, in 1996. He was the Chair of the Department of Computer Science from 2005 to 2008, and the Director of the Master's program in digital content and technologies with National Chengchi University (NCCU), Taipei, Taiwan, from 2013 to 2015. He is currently an Associate Professor with the Computer Science Department, and the Vice Dean of the Office of Research and Development, NCCU. His recent research work includes ad hoc networks, delay-tolerant networks, mobile commerce, mobile cloud computing, and wearable computing.

LING-JYH CHEN (SM'12) received the B.Ed. degree in information and computer education from National Taiwan Normal University in 1998, and the M.S. and Ph.D. degrees in computer science from the University of California at Los Angeles, Los Angeles, in 2002 and 2005, respectively. His research interests are wireless networks, mobile computing, network measurements, and social computing.