A Spatial-Temporal Approach for Air Quality Forecast in Urban Areas

Eric Hsueh-Chan Lu * and Chia-Yu Liu

Abstract: The diameter of PM2.5 is less than that of 2.5 μg/m^3 particulate matter; PM2.5 is small enough to enter the body through the alveolar microvasculature and has a major impact on human health. Therefore, people are interested in the establishment of air quality monitoring and forecasting. The historical and current air quality indices (AQI) can now be easily obtained from air quality sensors. However, people are more likely to need the PM2.5 forecasting information. Based on the literature, air quality varies because of a variety of factors, such as the meteorology in urban areas. In this paper, a spatial-temporal approach is proposed to forecast PM2.5 for 48 h using temporal and spatial features. From the temporal perspective, it is considered that the AQI in a few hours may be very similar because AQI is continuous. In addition, this research reveals the relationship between weather similarities and PM2.5 similarity. It is found that the more similar the weather is, the more similar the PM2.5 value is. From a spatial perspective, it is also considered that the air quality may be similar to that of the adjacent monitoring stations. Finally, the experimental results, based on AirBox data, show that the proposed approach outperforms the two methods based on well-established measurements in terms of the PM2.5 forecast error.

Keywords: air quality forecast; spatial-temporal; urban computing; data mining; AirBox

1. Introduction

In recent years, the awareness of air pollution has been increasing. Air pollution seriously affects human health, especially atmospheric particulate matter. PM2.5 is small enough to circulate in the body through the blood and it can cause cardiovascular disease, cancer, and even death. For the sustainable protection of the environment, countries develop regulations to reduce air pollution. In recent years, research into air quality issues has been rapidly increasing, which means that people are paying more attention to air quality issues. Air quality studies have focused on three main topics: air quality monitoring, air quality causal analysis, and air quality prediction and forecasting. Air quality information is easily obtained by setting up environmental sensors or it can be downloaded from open databases. In addition to the air quality monitoring stations built by the government, many inexpensive and portable lightweight air quality monitors have been invented. Figure 1 shows a real-time PM2.5 value map made by Edimax Technology [1]. Even if current air quality is easy to obtain, it cannot yet meet our needs. People are more likely to need PM2.5 forecasting, which is similar to the way that people need weather forecasting. This research is dedicated to forecasting PM2.5 values with a construable statistics method.

In recent years, air quality forecasting issues have been associated with meteorology, traffic flow, human mobility, points of interest, road networks, and the form of a city [2–10]. Even so, air quality issues are still complex and difficult to solve. Previous studies have showed various aspects of methods using the AQ model [11], the statistics method [4,6,8–10,12–17], and machine learning [18–21], which is the most popular method in recent years. While the results of machine learning perform well, it is difficult to discern
the cause-and-effect relationship of air quality and other features. This study is dedicated to constructing a construable and efficient method.

In this paper, a spatial-temporal approach is proposed based on meteorological factors for PM2.5 forecasting. In this approach, spatial information is considered, including the location of a geographic region, and temporal information, such as meteorology data and air quality for the forecasting model. It is found that meteorology features are related to PM2.5. To easily understand the cause-and-effect relationship of meteorology and PM2.5, a weight-average-based method is proposed for forecasting. It was found using PM2.5 at different times, which was monitored over close time intervals, and under very similar weather conditions. Additionally, PM2.5 values, at the same time, are similar when the monitored locations are close to each other. In the temporal part, data that have close times and have similar weather conditions are selected as references. These references are used for PM2.5 forecasting using similar weather conditions. In the spatial part, the forecast value in the temporal part is adjacent to the smooth part. Therefore, the time series of the PM2.5 forecast values for devices that are near to each other are likely to be similar. In this way, a PM2.5 value is forecasted over the following 48 h. Temporal and spatial data are collected and several internal and external experiments are conducted. The internal experimental results show that spatial or temporal feature combinations reduce forecast errors. In external experiments, we tried to compare our approach to two methods, based on the well-established Pearson correlation coefficient (PCC) [22] and inverse distance weighting (IDW) [23] measurements. The experimental results show that the proposed approach performs well and more stably in comparison to the other two methods.

2. Related Work

Baralis et al., Lv et al., Caglierio et al., and Zhang et al. highlighted the relationship between meteorology, season, and air quality [2,3,7,9]. However, the relationship between weather features and air quality seems to be different in different areas. Baralis et al. mentioned that temperature and air quality show a positive correlation in summer, but Zheng et al. obtained the opposite results in North China. Baralis et al. and Caglierio et al. showed that human activity, such as vehicle exhausts and industrial exhaust, is the main source of PM2.5; however, some human activity is beneficial to air quality. Gromke et al.
studied streets with or without bushes, and found that the air quality of streets with bushes was better than streets without bushes [5]. In addition to the points above, there are still lots of factors that could affect air quality. Traffic structures, wildfires, and even government regulations could have an impact on air quality [24–30]. As stated above, factors that could affect air quality are diverse and complex.

Based on the collected references, air quality analysis issues are classified into real time prediction and future forecasting. The literature on these two topics emphasized the importance of feature selection and has proposed some methods based on feature combinations. The more complete the features are, the more accurate the model is. The first category focuses on real-time air quality prediction in locations without air quality stations using features and air quality data. The majority of references propose an air quality prediction algorithm with multiple features, such as meteorological, traffic, human mobility, road networks, points of interest, and time [9,13,31,32]. Cagliero et al. analyzed the characteristics of air pollution using pattern mining, and found that air pollution is serious in autumn and winter [33]. There are some approaches that predict air quality using social networks [12,14–17]. These works predict air quality by mining key words in public posts and photos from users on social networks.

Another method is future air quality forecasting by features. People are more interested in future air quality forecasting than real-time air quality inference. A great deal of research has tried to forecast air quality using multiple features [4,6,8,10]. The authors in above works collected meteorology, traffic flow, human mobility, points of interest, road networks, and city data to forecast air quality. Zhu et al. used pattern mining and the Bayesian Gaussian model to analyze where air pollution comes from [34]. Sánchez-Balseca and Pérez-Foguet also used a Bayesian approach to forecast air quality in relation to wildfire events [35]. Domanska et al. and Bouarar et al. forecasted air quality using several models [36,37]. They analyzed the distribution of different sources, including chemical pollution, atmospheric composition, and ground emissions. As deep learning and machine learning are becoming more popular, some research has tried to solve this issue using machine learning and deep learning. Liu et al. used some learning models to forecast PM2.5 [11,18–21]. The results showed that the deep learning approach outperforms most models. Almost all the references mentioned that the most important factor affecting air quality forecasting is the features. Feature information is necessary in forecasting air quality. Natural factor and human factor analyses of air pollution composition are beneficial in forecasting air quality. There are other features that remain unknown. In this research, temporal and spatial features are prepared for PM2.5 forecasting.

3. Proposed Method

This section describes the features included and the proposed approach for PM2.5 value forecasting over 48 h. We collected temporal and spatial features to forecast the PM2.5 value. Temporal features were used for preliminary forecasting and then spatial features were used to improve the forecast.

3.1. System Framework

Figure 2 shows a flowchart of the PM2.5 value forecast. The PM2.5 data, meteorology data, and weather forecast data were included for PM2.5 value forecasting. It was considered that meteorology data are highly correlated with air quality. The PM2.5 data and meteorology data were combined for each hour and were regarded as historical data. Then, the PM2.5 value was forecasted using the weather forecast data and the spatial data.
Figure 2. Framework of the spatial-temporal PM2.5 forecast approach.

There were some pre-processing tasks for PM2.5 data. Some unreasonable air monitoring devices, which are considered as being installed indoors, needed to be removed. Some weather features, such as temperature and humidity, were more stable than outdoor weather features [38]. The most important factor was that indoor air quality was less impacted by meteorology. Then, the PM2.5 data of an air quality sensor and the meteorology data of the nearest weather station were combined for each hour as historical data. Additionally, the weather forecast data were regarded as future data. There were two parts to the methodology: temporal and spatial parts. In terms of the temporal part, the preliminary PM2.5 value was forecasted using the similarity between historical meteorology data and weather forecast data. Air quality was associated with time. In addition to the data with similar meteorology features, recent data were also referred to. Therefore, recent data and data with similar weather features were detected to forecast a PM2.5 value. The preliminary forecasted PM2.5 value was evaluated using the weighted-average formula. The degree of similarity of weather features was considered to be the weight.

In terms of the spatial part, it was found that the PM2.5 values were highly correlated with distance. The PM2.5 value trends of devices near to each other are similar. Therefore, the preliminary forecasted PM2.5 value was smooth with other forecast PM2.5 from nearby devices from each forecast hour. In addition to distance, wind speed and wind direction were both taken into consideration. It was considered that the forecasted distribution per hour was a credible trend and resulted in the forecast trend being similar due to normalization. The value after smoothing was the final result of the forecast.

3.2. Utilized Features

Features that may affect PM2.5 were included for the PM2.5 forecast. A great deal of research has mentioned that meteorology data are highly related to PM2.5 [2,3,6,9,10,13,31,32,34,36,37,39]. For example, when it is not rainy, the PM2.5 value increases as the relative humidity increases. However, the PM2.5 value usually drops rapidly when it starts to rain. In addition, as wind speed increases, air quality improves; air quality is better when a sea breeze is blowing, which usually carries particulate matter into the land breeze. This
research also found meteorology data and air quality connections. Temperature, relative humidity, pressure, wind direction, wind speed, and daily rainfall were included in the meteorological data. In addition to meteorological data, the PM2.5 values had a lot to do with the spatial data. When a device monitored a good PM2.5 value, adjacent devices also monitored good values. In this study, the impacts of the atmosphere, regardless of individual factors, such as traffic emissions and factory emissions, were emphasized. These factors may have applied to time; for example, they are usually blocked during peak hours, or a factory emits exhaust gas at midnight. In this section, we will present the data used in the study. The dataset consists of temporal and spatial data. The features, types, and descriptions are introduced in Table 1.

Table 1. Summary of all features.

| Feature     | Type    | Description                                                                 |
|-------------|---------|----------------------------------------------------------------------------|
| Latitude    | Spatial | Latitude (WG84) of air quality monitoring devices                          |
| Longitude   | Spatial | Longitude (WG84) of air quality monitoring devices                         |
| Altitude    | Spatial | Altitude of air quality monitoring devices (unit: meter)                   |
| Observed Time | Temporal | Observed time of monitored data                                           |
| Temperature | Temporal | Temperature data (unit: Celsius)                                           |
| Humidity    | Temporal | Relative humidity data (unit: %)                                           |
| Wind Speed  | Temporal | Wind speed data (unit: m/s)                                                |
| Wind Direction | Temporal | Wind direction data (unit: bearing angle)                                  |
| Pressure    | Temporal | Pressure data (unit: hPa)                                                  |
| Rainfall    | Temporal | Daily rainfall (unit: mm)                                                  |
| PM2.5       | Temporal | PM2.5 value (unit: µg/m³)                                                  |

In this section, the method of PM2.5 forecasting is introduced. This study is dedicated to forecasting PM2.5 values over 48 h. Many studies have shown that meteorological data have a lot to do with air quality. We also found this connection in the data. Figure 3 shows the relationship between each feature difference and the PM2.5 value difference; it shows that, when a feature difference is smaller, the PM2.5 value difference is smaller. In terms of distance difference, the highest density was not close to zero. This was because the distance average of the data was about 6 km (Figure 3h), and data less than 6 km were not enough. Figure 3i concerns weather difference and PM2.5 difference. The weather difference \(WD(t_u, t_v)\) between times \(t_u\) and \(t_v\) was evaluated using Equation (1). There are \(m\) weather features in total. Each feature was normalized to 0–1 individually, and were compared with each observation from each time. The notation \(f_k(t)\) means the \(k\)-th weather feature at time \(t_u\). In Equation (1), it is assumed that the weight of each feature has the same value. The last figure shows that the PM2.5 value was probably similar when the weather showed high similarity. Therefore, a PM2.5 value forecast approach based on weather differences between the reference data and the target was designed.

\[
WD(t_u, t_v) = \sqrt{\sum_{k=1}^{m} [f_k(t_u) - f_k(t_v)]^2}
\]  

(1)

PM2.5 data and weather data were combined for forecasting. The method is divided into temporal and spatial parts; in the temporal part, we used the weather and forecast weather data to estimate preliminary PM2.5 with a weighted-average formula. The weight depended on weather similarity. The data that were monitored over close periods of time, and which have a high weather similarity, were selected as references for the forecast. In the spatial part, it was found that the PM2.5 time series of adjacent devices during the same period was very similar. Figure 4 shows the PM2.5 time series for adjacent devices within 5 km. We can see that the PM2.5 time series of different devices were almost the same. Therefore, we used nearby monitoring stations to adjust the estimated value of PM2.5.
Figure 3. Relationship between (a) temperature, (b) humidity, (c) pressure, (d) wind direction, (e) wind speed, (f) rainfall, (g) time, (h) distance, and (i) weather difference and PM2.5 value error.

Figure 4. PM2.5 time series of adjacent devices within 5 km.

Obtaining a similar time series of the PM2.5 forecast for adjacent devices was expected; however, an accurate forecast time series as a standard was difficult to discern. We used the median of the estimated PM2.5 value in the temporal part for each hour as a more accurate time series, rather than selecting a forecast time series that may be accurate. The time series of adjacent devices within a specific range were smoothed closer. The temporal and spatial methods will be discussed, respectively, in the following sections.
3.3. Weighted-Average Strategy to Forecast PM2.5 Value

Normally, if air quality is good at a particular moment, it will probably be good in the next moment. As mentioned in the previous section, if the weather feature similarity is high between two datasets, their PM2.5 values are very likely to be similar. It is considered that air quality values are correlated from temporal and spatial perspectives. We drew a chart to model the spatial-temporal correlation at specific locations. Recent data and similar data were used as references in this section. The reference data were selected for forecasting. The “recent data” is defined as the last observation in the historical data prior to the target that is going to be forecasted; “similar” is defined as the data in history with more similar weather features to the target period. In addition, most similar data were not recent data. The construction of Figure 5 includes how recent and similar data were selected.

![Schematic diagram of the forecasting approach.](image)

**Figure 5.** Schematic diagram of the forecasting approach.

A weighted-average-based method was proposed for forecasting. For device $d_i$, weight $w$ between times $t_u$ and $t_v$ was evaluated using Equation (2). In terms of the temporal part, the preliminary forecast ($p^T$) at time $t_u$ was evaluated with $h$ references (including recent data ($r$) and similar data ($s$)), in total, using Equation (3). In the spatial part, we proposed concentrating the forecast time series close to the median forecast values of nearby devices in each hour using Equation (4). The design idea of Equation (4) is illustrated with Figure 6. Figure 6 shows that the blue area, which is a distribution of the forecast value of nearby devices in an hour, is concentrated close to the median by $c\%$. It is expected that forecast value $p^S$ of adjacent devices are similar. Notation $D(t_u)$ is a set of temporal PM2.5 forecastings of $n$ adjacent devices. Further, notations $\text{max}D(t_u)$ and $\text{min}D(t_u)$ represent the maximum and minimum temporal PM2.5 forecast of adjacent devices ($D(t_u)$) around device $d_i$, respectively. If temporal PM2.5 forecasting ($p^T$) is larger than or equal to $\text{med}D(t_u)$, $p^T$ is normalized to $c\%$ of $\text{med}D(t_u)$ to $\text{max}D(t_u)$. Otherwise, $p^T$ is normalized to $c\%$ of $\text{min}D(t_u)$ to $\text{med}D(t_u)$, where $\text{med}D(t_u)$ means a median temporal PM2.5 forecast of adjacent devices $D(t_u)$ around device $d_i$.

\[
w(t_u, t_v) = \frac{1}{WD(t_u, t_v)}
\]

\[
p^T(t_u) = \frac{1}{\sum_{v=1}^{h} w(t_u, t_v)} \sum_{v=1}^{h} w(t_u, t_v) \times p(t_v)
\]

\[
p^S(t_u) = \begin{cases} 
\frac{p^T(t_u) - \text{med}D(t_u)}{|\text{max}D(t_u) - \text{med}D(t_u)|} & \text{if } p^T(t_u) \geq \text{med}D(t_u) \\
\frac{p^T(t_u) - \text{med}D(t_u)}{|\text{med}D(t_u) - \text{min}D(t_u)|} & \text{otherwise}
\end{cases}
\]
4. Experimental Evaluations

A series of experiments were performed using the monitoring data to evaluate the performance of our proposed PM2.5 forecasting approach. The experiment was divided into two parts: the first part was for the testing of various parameter settings and the second part was for comparison of our proposed approach to two methods, based on PCC and IDW, in terms of the air quality forecasting errors.

4.1. Experimental Datasets and Settings

This research utilized several features that were mentioned in Section 3.2 to reveal the result of a PM2.5 value forecast. AirBox data were selected for experiments [1]. AirBox was provided by the Taipei government (Taiwan, R.O.C). AirBox also cooperates with Environmental Protection Administration Executive Yuan as an air quality sensor IoT system; data from AirBox were much more abundant than from national sensors. In Figure 7, the blue dots represent 146 AirBoxes in Taipei, and the purple circles represent 21 weather stations around Taipei. To prevent inaccuracies in the weather forecast data, weather forecast data were replaced with monitored meteorological data. AirBox data were combined with the nearest weather station data every hour. The spatial data provided by Data.Taipei were used for this study. Due to AirBox users’ private rights, the coordinates of consumers were rounded to three decimal places; therefore, the location error could be about 75 m.

1. AirBox: An AirBox is a device that can monitor PM2.5 using the principles of optics, temperature, and relative humidity. It measures the number of particles and evaluates the concentration of PM2.5. Edimax Technology and Academia Sinica produce AirBox and make it available to schools and citizens. Each device monitors data every 5 min; however, weather data are monitored every hour. We averaged the monitored data from AirBox every hour to combine it with the weather data. As long as AirBox users agree, the monitoring data are uploaded to the Edimax Internet-of-Things platform and will be made available as an open data download. This was provided by Data.Taipei.

2. Weather station: The Central Weather Bureau provides meteorological data and weather forecast data. They consist of temperature, relative humidity, wind speed, wind direction, atmospheric pressure, and daily rainfall. It monitors data once an hour. We received the weather data from data.gov.
3. Location information: Location data include the latitude and longitude of each AirBox device. Due to users’ private rights, some location coordinates were only accurate to the third decimal place. These data were also received from Data.Taipei.

![Figure 7. Distribution of AirBoxes and weather stations in or around Taipei.](image)

The experiment data were collected from March 2017 to May 2018. The data were divided into a historical part and a test part. The data before January 2018 were defined as history data, and the data from January 2018 were to be forecasted. After 48 h of forecast completion, the following 48 h of the next hour were forecasted, and the last hour would be added into the historical data. For the forecasting day for every 48 h, the performance of the forecast method was analyzed using the RMSE per hour.

### 4.2. Impact on Various Parameter Settings

There were three parameters in the methodology that we focused on: amount of recent data ($r$), similar data ($s$) and adjacent devices ($n$) were detected in the forecast. In the following, we divide these parameter experiments into two sub-sections. In Section 4.2.1, the experiment is used to determine temporal parameters $r$ and $s$, and Section 4.2.2 shows an experiment for spatial parameters.

#### 4.2.1. Temporal Parameter

First, we tested the temporal parameters. This experiment tested the recent data amount ($r$) and similar data amount ($s$) only. The correlation between parameter $r$ and PM2.5 forecast at the 6th, 12th, 24th, 36th, and 48th hours are shown in Figure 8. The x-axis is parameter $r$, and the y-axis is the RMSE of the PM2.5 forecast value. It is obvious that as $r$ becomes larger, the RMSE of PM2.5 value becomes better at the 12th, 24th, 36th, and 48th hours. However, the degree of decline in RMSE is also small. For the PM2.5 forecast at the 24th hour, as $r$ increases from 1 to 30, the RMSE of PM2.5 is only reduced by about 1.5 $\mu g/m^3$. As $r$ becomes larger, the degree of decline was considered to converge. However, the trend of the 6th hour is opposite to the other curves in the figure. As $r$ became larger, the RMSE at the 6th hour became larger. Air quality is usually highly similar to the last observed value because air quality is continuous data. This is the reason why the RMSE was smaller when there were only recent data selected. In addition, as the forecasting time was further away, the RMSE became larger. Therefore, the farther the forecast hour, the harder it is to forecast.
4.2.1. Temporal Parameter

First, we tested the temporal parameters. This experiment tested the recent data and similar reference data. Figure 9 shows that when recent and similar reference data are both used for forecasting, the RMSE of the forecast results are close every hour. When the forecast is further apart, the reference value of \( r \) becomes smaller, but \( s \) is more decisive. While \( s \) may include recent data, it is possible to select long-term data from the history. This tells us that, when the forecast hour is further away, the RMSE will become larger if we only use similar reference data.

![Figure 8. The impact on PM2.5 error of parameter \( r \) on the PM2.5 value forecast.](image)

Figure 8 shows the performance of parameter \( s \). The degree of convergence of RMSE is more pronounced than it is with \( r \). However, the RMSE is greater when \( s = 1 \) than when \( r = 1 \). When they are both 30, RMSE is better than \( r \). In addition, the degree of decline of a similar reference is greater than the recent reference. This indicates that parameter \( s \) is beneficial for the PM2.5 forecast. According to the experimental results, it was determined that the RMSE results of the forecast results are close every hour. When the forecast is further apart, the reference value of \( r \) becomes smaller, but \( s \) is more decisive. While \( s \) may include recent data, it is possible to select long-term data from the history. This tells us that, when the forecast hour is further away, the RMSE will become larger if we only use similar reference data.

![Figure 9. The impact on PM2.5 error of parameter \( s \) on PM2.5 value forecast.](image)

Figure 9 shows the performance of parameter \( s \). The degree of convergence of RMSE is more pronounced than it is with \( r \). However, the RMSE is greater when \( s = 1 \) than when \( r = 1 \). When they are both 30, RMSE is better than \( r \). In addition, the degree of decline of a similar reference is greater than the recent reference. This indicates that parameter \( s \) is beneficial for the PM2.5 forecast. According to the experimental results, it was determined that the RMSE results of the forecast results are close every hour. When the forecast is further apart, the reference value of \( r \) becomes smaller, but \( s \) is more decisive. While \( s \) may include recent data, it is possible to select long-term data from the history. This tells us that, when the forecast hour is further away, the RMSE will become larger if we only use similar reference data.

Figure 10 shows the experimental results when parameters \( r \) and \( s \) are set to the same value. As the number of references increases, we can see that the RMSE improves and the degree of decline also converges. It is worth noting that the 6th hour of the RMSE does not improve as the number of references increases. When \( r \) and \( s \) are greater than 7, the RMSE becomes even worse. According to the experimental results, we found that it only takes a few \( r \) to forecast the first 6th hour, and the RMSE will perform very well.

This shows the different combinations of parameters, from 0 to 30, at the 6th, 12th, 24th, 36th, and 48th h, in Figure 11. We can see that the RMSE at the 12th, 24th, 36th, and 48th h becomes smaller as the parameters increase and the convergence is also obvious. The results from the 6th hour show that it can perform well when there is only one recent reference. When the parameter is larger, its RMSE will not improve. We can see that the impact of recent data is better than the impact of similar data at the 6th and 12th h. Similar references performed well at the 24th, 36th, and 48th h. This tells us that the recent reference is more useful in forecasting the most near future time. When we forecast further than an hour into the future, similar references are more important than recent data. Figure 11 shows that when recent and similar reference data are both used for forecasting, the RMSE of the PM2.5 forecast value is better than when only using one of them.
After these experimental analyses, the research forecast PM2.5 value was only available from one recent reference dataset for hours 1–6 and with 15 recent references and 15 similar references for hours 7–48, respectively. Due to convergence and time cost considerations, we did not use 30 references for forecasting. Figure 12 shows a comparison of different combinations of references. The blue line indicates that $r$ and $s$ are, respectively,
set at 15; the red line uses only one recent reference, and the yellow line is the combination we used. The purple line is the best RMSE of all combinations. The yellow line and purple line become more and more different as the forecast time continues; the difference is not obvious. The best combination at each forecast hour was not used to prevent overfitting. Therefore, the same reference is the setting for the 7th to 48th h.

![Figure 12. Comparison of the best r and s combinations.](image)

### 4.2.2. Spatial Parameter

Next, we added the spatial parameter and referred to the forecasted PM2.5 value from the nearest \( n \) devices. Figure 13 shows the impact of a number of spatial references surrounding devices on the RMSE of PM2.5 value forecast. We varied the number of surrounding station references \( n \) from 0 to 70, where 0 means that we did not use any spatial references for air quality forecasting. The experimental results show that, as \( n \) increases, the RMSE first drops and then increases. When \( n \) is larger than 25, the RMSE may be even worse than the results without the reference to surrounding stations. The reason for this is that the reference range is too large, resulting in a large difference in air quality. To find the best parameter setting, we scaled the results of \( n \) up from 0 to 30. We could observe that every line had a minimum RMSE when \( n \) was about 5 to 7. While the spatial parameter has a low RMSE effect on the results, we set \( n \) as 6, based on the experimental results in the following experiments.

![Figure 13. PM2.5 RMSE of different parameters of n.](image)

### 4.3. Performance

To show the performance of our proposed approach, we modified two methods based on two well-established measurements and tried some simple methods. The first was the
PCC-based method. The temporal weather similarity was calculated using PCC instead of using recent and similar references. The idea was that the PM2.5 time series was highly correlated to the weather conditions. Hence, the first method modified PCC to calculate weather similarities between the historically continuous 48 h and the target 48 h. The PM2.5 time series of the continuous 48 h with the highest weather similarity in the historical data would be extracted to forecast the target PM2.5. Due to the air quality being continuous data, the forecast value would increase or decrease based on the variety of the extracted PM2.5 time series, instead of the value of the extracted PM2.5 time series. To compare this method, we also tried to forecast using the extracting continuous 48 h PM2.5 in historical data randomly. For the second, IDW-based method, we used IDW to average the forecast time series value for the spatial inference. The temporal inference of this method and our approach was the same, but this method selected six nearby devices around the target device, selected as in Section 4.2.2 and used IDW to average the PM2.5 values as the final forecast result of PM2.5. To check our method, we did not require too many parameters, “r” and “s”, and it was forecasted using the mean weather. The results of the mean weather are shown below. We also tried to compare this with the results of the forecast from the same hour over the previous 90 days. Figure 14a shows the results of different methods in terms of RMSE and MAE. Our method was better than other methods used. Figure 14b shows the RMSE of the methods for each forecasted hour. The result shows that our proposed approach can converge and stably forecast future hours. However, after about 30 h, the 90-day results narrowly win. In Figure 14c, results from different days are compared. The results are better when they refer to more days, and the results converge within 30 to 90 days. In Figure 14d, except the result from the afternoon, our results were more accurate for more periods, and the results show that the PM2.5 in the afternoon is harder to forecast.

![Figure 14](image-url)

**Figure 14.** Evaluation of the PM2.5 forecast error for (a) different methods in RMSE and MAE, (b) different methods in various hours, (c) different days methods in various hours, and (d) different methods in various hours in days.

The acceptable error of PM2.5 forecast is introduced. Table 2 shows the grade of PM2.5 from Environmental Protection Administration Executive Yuan (R.O.C.). Given that the range of each grade is different (Table 2), the acceptable error is different in each grade. In

| Grade          | PM2.5 (μg/m³) |
|----------------|---------------|
| Good           | 0.0–15.4      |
| Moderate       | 15.5–35.4     |
| Unhealthy      | 35.5–54.5     |
| Very Unhealthy | 54.5–150.4    |
| Hazardous      | 150.5–250.4   |
|                 | 250.5–500.4   |

| Method          | RMSE | MAE  |
|-----------------|------|------|
| PCC-based       |      |      |
| Random hour     |      |      |
| Mean weather    |      |      |
| 90-days method  |      |      |
| IDW-based       |      |      |
| Our method      |      |      |

---

Table 2. Grade of PM2.5 from Environmental Protection Administration Executive Yuan (R.O.C.).
Figure 15, the bar is the standard deviation (STD) of our results in each grade. PM2.5 in data are within 150, so there are no data in grade “Very Unhealthy”. This shows that our results are within the acceptable error in each grade.

### Table 2. Grade of PM2.5 from Environmental Protection Administration Executive Yuan (R.O.C.).

| Grade                                | Good       | Moderate   | Unhealthy for Sensitive Groups | Unhealthy | Very Unhealthy | Hazardous |
|--------------------------------------|------------|------------|--------------------------------|-----------|----------------|-----------|
| PM2.5 (µg/m³)                        | 0.0–15.4   | 15.5–35.4  | 35.5–54.4                      | 54.5–150.4| 150.5–250.4    | 250.5–500.4|

![Bar graph showing acceptable error and each grade's standard deviation from our results.](image)

**Figure 15.** Acceptable error and each grade’s standard deviation from our results.

### 5. Conclusions and Future Work

Recently, people’s awareness about air pollution has increased. There is an urgent need for air quality forecasting. In this paper, we discussed the relationship between spatial-temporal data and PM2.5 values, and forecasted air quality using a spatial-temporal combination approach. In this approach, both recent and similar references were considered for temporal air quality forecasting. According to the experimental results, considering recent references can only forecast air quality in the first six hours. It is necessary to consider 15 recent and 15 similar references to forecast PM2.5 after six hours. While the experimental results show that the forecast error decreases with the increase in reference data, it shows a tendency to converge. To reduce the amount of calculations, we did not use a larger reference amount. For the spatial air quality forecast, nearby monitoring devices with similar meteorological features were considered as references to forecast the PM2.5 values. This can cause the trend of the PM2.5 forecast value to be similar for nearby devices for a period of time. The results showed that our proposed approach outperforms methods based on well-established IDW and PCC measurements and other simple methods in terms of the forecast error. The results of our proposed approach are convergent and stable. While we only forecasted the PM2.5 value for 48 h, our proposed approach could forecast further PM2.5 values. In fact, there are many events that can cause sudden increases in PM2.5 value, such as someone smoking or burning something. Sources and sinks of PM2.5 are also important. In the future, we will collect more data for seasonal analyses, and we will try to analyze the causality of PM2.5 value and traffic flow, human mobility, points of interest, natural disasters, terrain, etc. Furthermore, we will try to discover how weather conditions affect air quality, and will consider adding weight to each feature. For the spatial method, there is still a long way to go. We will test, with devices, which has the highest correlation, which are in similar terrains, and which are in similar weather conditions in terms of the spatial method. Additionally, we found that PM2.5 in the same or opposite wind directions is similar. We will also consider the relationship of upwind and leeward to improve the spatial forecast mechanism.
Author Contributions: Conceptualization, E.H.-C.L. and C.-Y.L.; methodology, E.H.-C.L. and C.-Y.L.; software, C.-Y.L.; validation, E.H.-C.L. and C.-Y.L.; formal analysis, E.H.-C.L. and C.-Y.L.; investigation, C.-Y.L.; resources, E.H.-C.L.; data curation, E.H.-C.L.; writing—original draft preparation, E.H.-C.L. and C.-Y.L.; writing—review and editing, E.H.-C.L.; visualization, C.-Y.L.; supervision, E.H.-C.L.; project administration, E.H.-C.L.; funding acquisition, E.H.-C.L. Both authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Science and Technology, Taiwan, R.O.C., grant number MOST 109-2121-M-006-013-MY2 and MOST 109-2121-M-006-005-.

Conflicts of Interest: The authors declare no conflict of interest.

References
1. Chen, L.J.; Hsu, W.; Cheng, M.; Lee, H.C. LASS: A location-aware sensing system for participatory PM2.5 monitoring. In Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services Companion, New York, NY, USA, June 2016; p. 98.
2. Baralis, E.; Cerquitelli, T.; Chiusano, S.; Garza, P.; Kavoosifar, M.R. Analyzing Air Pollution on The Urban Environment. In Proceedings of the 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 30 May–3 June 2016.
3. Cagliero, L.; Cerquitelli, T.; Chiusano, S.; Garza, P.; Ricupero, G.; Xiao, X. Modeling Correlations Among Air Pollution-Related Data Through Generalized Association Rules. In Proceedings of the IEEE International Conference on Smart Computing (SMARTCOMP), St. Louis, MI, USA, 18–20 May 2016.
4. Calculi, C.; Fassò, A.; Finazzi, F.; Pollice, A.; Turnone, A. Maximum Likelihood Estimation of the Multivariate Hidden Dynamic Geostatistical Model with Application to Air Quality in Apulia, Italy. *Environmetrics* 2015, 26, 406–417. [CrossRef]
5. Gromke, C.; Jamarkattel, N.; Ruck, B. Influence of Roadside Hedgerows on Air Quality in Urban Street Canyons. *Atmos. Environ.* 2016, 139, 75–86. [CrossRef]
6. Lu, X.; Wang, Y.; Huang, L.; Yang, W.; Shen, Y. Temporal-Spatial Aggregated Urban Air Quality Inference with Heterogeneous Big Data. In Proceedings of the International Conference on Wireless Algorithms, Systems, and Applications, Bozeman, MT, USA, 8–10 August 2016.
7. Lv, B.; Zhang, B.; Bai, Y. A Systematic Analysis of PM2.5 in Beijing and Its Sources from 2000 to 2012. *Atmos. Environ.* 2016, 124, 98–108. [CrossRef]
8. Wang, S.; Paul, M.J.; Dredze, M. Social Media as A Sensor of Air Quality and Public Response in China. *J. Med. Internet Res.* 2015, 17, 22. [CrossRef] [PubMed]
9. Zhang, Y.; Ding, A.; Mao, H.; Nie, W.; Zhou, D.; Liu, L.; Fu, C. Impact of Synoptic Weather Patterns and Inter-Decadal Climate Variability on Air Quality in The North China Plain During 1980–2013. *Atmos. Environ.* 2016, 124, 119–128. [CrossRef]
10. Zheng, Y.; Yi, X.; Li, M.; Li, R.; Shan, Z.; Chang, E.; Li, T. Forecasting Fine-Grained Air Quality based on Big Data. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015.
11. Cheng, F.Y.; Feng, C.Y.; Yang, Z.M.; Hsu, C.H.; Chan, K.W.; Lee, C.Y.; Chang, S.C. Evaluation of real-time PM2.5 forecasts with the WRF-CMAQ modeling system and weather-pattern-dependent bias-adjusted PM2.5 forecasts in Taiwan. *Atmos. Environ.* 2021, 244, 117909. [CrossRef]
12. Chen, J.; Chen, H.; Zheng, G.; Pan, J.Z.; Wu, H.; Zhang, N. Big Smog Meets Web Science: Smog Disaster Analysis Based on Social Media and Device Data on The Web. In Proceedings of the International Conference on World Wide Web, Seoul, Korea, 7–11 April 2014.
13. Hsieh, H.P.; Lin, S.D.; Zheng, Y. Inferring Air Quality for Station Location Recommendation based on Urban Big Data. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015.
14. Jiang, W.; Wang, Y.; Tsou, M.H.; Fu, X. Using Social Media to Detect Outdoor Air Pollution and Monitor Air Quality Index (AQI): A Geo-Targeted Spatiotemporal Analysis Framework with Sina Weibo (Chinese Twitter). *PLoS ONE* 2015, 10, e0141185. [CrossRef]
15. Li, Y.; Huang, J.; Luo, J. Using User Generated Online Photos to Estimate and Monitor Air Pollution in Major Cities. In Proceedings of the 7th International Conference on Internet Multimedia Computing and Service, Zhangjiajie, China, 19–21 August 2015.
16. Mei, S.; Li, H.; Fan, J.; Zhu, X.; Dyer, C.R. Inferring Air Pollution by Sniffing Social Media. In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Beijing, China, 17–20 August 2014.
17. Wang, Y.; Xu, M.; Huang, H.; Chen, S.X. A Spatio-Temporal Model for The Analysis and Prediction of Fine Particulate Matter Concentration in Beijing. *Environmetrics* 2020, 32, e2648. [CrossRef]
18. Lee, M.; Lin, L.; Chen, C.Y.; Tsao, Y.; Yao, T.H.; Fei, M.H.; Fang, S.H. Forecasting air quality in Taiwan by using machine learning. *Sci. Rep.* 2020, 10, 1–13. [CrossRef]
19. Liu, H.; Duan, Z.; Chen, C. A Hybrid Framework for Forecasting PM2.5 Concentrations using Multi-Step Deterministic and Probabilistic Strategy. *Air Qual. Atmos. Health* 2019, 12, 785–795. [CrossRef]
20. Liu, H.; Dong, S. A Novel Hybrid Ensemble Model for Hourly PM2.5 Forecasting using Multiple Neural Networks: A Case Study in China. *Air Qual. Atmos. Health* **2020**, *13*, 1411–1420. [CrossRef]

21. Liu, H.; Long, Z.; Duan, Z.; Shi, H. A New Model Using Multiple Feature Clustering and Neural Networks for Forecasting Hourly PM2.5 Concentrations, and Its Applications in China. *Engineering* **2020**, *6*, 944–956. [CrossRef]

22. Lee Rodgers, J.; Nicewander, W.A. Thirteen Ways to Look at The Correlation Coefficient. *Am. Stat.* **1988**, *42*, 59–66. [CrossRef]

23. Shepard, D. A Two-Dimensional Interpolation Function for Irregularly-Spaced Data. In Proceedings of the ACM National Conference, New York, NY, USA, 27–29 January 1968; pp. 517–524.

24. Aouizerats, B.; Van Der Werf, G.R.; Balasubramanian, R.; Betha, R. Importance of Transboundary Transport of Biomass Burning Emissions to Regional Air Quality in Southeast Asia During a High Fire Event. *Atmos. Chem. Phys.**2015**, *15*, 363–373. [CrossRef]

25. Crippa, M.; Janssens-Maenhout, G.; Dentener, F.; Guizzardi, D.; Sindelarova, K.; Muntean, M.; Van Dingenen, R.; Granier, C. Forty Years of Improvements in European Air Quality: Regional Policy-Industry Interactions with Global Impacts. *Atmos. Chem. Phys.* **2016**, *16*, 3825–3841. [CrossRef]

26. Cuchiara, G.C.; Rappenglück, B.; Rubio, M.A.; Lissi, E.; Gramsch, E.; Garreau, R.D. Modeling Study of Biomass Burning Plumes and Their Impact on Urban Air Quality: A Case Study of Santiago De Chile. *Atmos. Environ.* **2017**, *166*, 79–91. [CrossRef]

27. Lee, Y.H.; Shindell, D.T.; Faluvegi, G.; Pinder, R.W. Potential Impact of A US Climate Policy and Air Quality Regulations on Future Air Quality and Climate Change. *Atmos. Chem. Phys. Discuss.* **2016**, *16*, 5323–5342. [CrossRef]

28. Martins, V.; Moreno, T.; Mendes, L.; Eleftheriadis, K.; Diapouli, E.; Alves, C.A.; Minguillón, M.C. Factors Controlling Air Quality in Different European Subway Systems. *Environ. Res.* **2016**, *146*, 35–46. [CrossRef]

29. Millstein, D.; Wiser, R.; Bolinger, M.; Barbose, G. The Climate and Air-Quality Benefits of Wind and Solar Power in The United States. *Nat. Energy* **2017**, *2*, 17134. [CrossRef]

30. Shirmohammadi, F.; Sowlat, M.H.; Hasheminassab, S.; Saffari, A.; Ban-Weiss, G.; Sioutas, C. Emission Rates of Particle Number, Mass and Black Carbon by The Los Angeles International Airport (LAX) and Its Impact on Air Quality in Los Angeles. *Atmos. Environ.* **2017**, *151*, 82–93. [CrossRef]

31. Dong, Y.; Wang, H.; Zhang, L.; Zhang, K. An Improved Model for PM2.5 Inference based on Support Vector Machine. In Proceedings of the IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), Shanghai, China, 30 May–1 June 2016.

32. Zheng, Y.; Liu, F.; Hsieh, H.P. U-Air: When Urban Air Quality Inference Meets Big Data. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA, 11–14 August 2013.

33. Cagliero, L.; Cerquitelli, T.; Chiusano, S.; Garza, P.; Ricupero, G. Discovering Air Quality Patterns in Urban Environments. In Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing, Adjunct, Heidelberg, Germany, 12–16 September 2016.

34. Zhu, J.Y.; Zhang, C.; Zhang, H.; Zhi, S.; Li, V.O.; Han, J.; Zheng, Y. p-Causality: Identifying Spatiotemporal Causal Pathways for Air Pollutants with Urban Big Data. In *IEEE Trans. Big Data* **2018**, *4*, 571–585. [CrossRef]

35. Sánchez-Balseca, J.; Pérez-Foguet, A. Modelling Hourly Spatio-Temporal PM2.5 Concentration in Wildfire Scenarios using Dynamic Linear Models. *Atmos. Res.* **2020**, *242*, 104999. [CrossRef]

36. Bourar, I.; Brasseur, G.; Granier, C.; Petersen, K.; Doumbia, E.H.T.; Wang, X.; Fan, Q.; Gauss, M.; Peuch, V.H.; Pommier, M.; et al. Monitoring and Forecasting Air Quality over China: Results from the PANDA Modeling System. In Proceedings of the International Global Atmospheric Chemistry (IGAC) Science Conference, Breckenridge, CO, USA, 26–30 September 2016.

37. Domanska, D.; Lukasik, S. Handling High-Dimensional Data in Air Pollution Forecasting Tasks. *Ecol. Inform.* **2016**, *34*, 70–91. [CrossRef]

38. Zhuang, Y.; Lin, F.; Yoo, E.H.; Xu, W. Airsense: A Portable Context-Sensing Device for Personal Air Quality Monitoring. In Proceedings of the Workshop on Pervasive Wireless Healthcare, Hangzhou, China, 22 June 2015.

39. Zheng, Y.; Chen, X.; Jin, Q.; Chen, Y.; Qu, X.; Liu, X.; Sun, W. A Cloud-Based Knowledge Discovery System for Monitoring Fine-Grained Air Quality. 2014. Available online: http://research.microsoft.com/apps/pubs/default (accessed on 27 May 2021).