Environment International 134 (2020) 105305

Contents lists available at ScienceDirect

Environment International

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

The improvement of spatial-temporal resolution of $PM_{2.5}$ estimation based on micro-air quality sensors by using data fusion technique

Yuan-Chien Lin*, Wan-Ju Chi, Yong-Qing Lin

Dept. of Civil Engineering, National Central University, Taoyuan City 32001, Taiwan

ARTICLE INFO

Handling Editor: Dr. Xavier Querol

Keywords:
$PM_{2.5}$
Micro-air quality sensors
Data fusion
Spatial-temporal estimation

ABSTRACT

With the rapid development of the Internet of things (IoTs) and modern industrial society, forecasting air pollution concentration, e.g., the concentration of $PM_{2.5}$, is of great significance to protect human health and the environment. Accurate prediction of $PM_{2.5}$ concentrations is limited by the number and the data quality of air quality monitoring stations. In Taiwan, the spatial and temporal data of $PM_{2.5}$ concentrations are measured by 77 national air quality monitoring stations (built by Taiwan EPA). However, the national stations are costly and scarce because of the highly precise instrument and their size. Therefore, many places are still out of coverage of the monitoring network. Recently, under the framework of IoTs, there are hundreds of portable air quality sensors called "AirBox" developed jointly by the Taiwan local government and a private company. By virtue of its low price and portability, the AirBox can provide a higher resolution of space-time $PM_{2.5}$ measurement. However, the spatiotemporal distribution is different between AirBox and EPA stations, and data quality and accuracy of AirBox is poorer than national air quality monitoring stations. Thus, to integrate the heterogeneous $PM_{2.5}$ data, the data fusion technique should be used before further analysis.

In this study, we propose a new data fusion method called multi-sensor space-time data fusion framework. It is based on the Optimum Linear Data Fusion theory and integrating with a multi-time step Kriging method for spatial-temporal estimation. The method is used to do heterogeneous data fusion from different sources and data qualities. It is able to improve the estimation of $PM_{2.5}$ concentration in space and time. Results have shown that by combining $PM_{2.5}$ concentration data from 1176 low-cost AirBoxes as additional information in our model, the estimation of spatial-temporal $PM_{2.5}$ concentration becomes better and more reasonable. The r^2 of the validation regression model is 0.89. Under the approach proposed in this study, we made the information of the micro-sensors more reliable and improved the higher spatial-temporal resolution of air quality monitoring. It could provide very useful information for better spatial-temporal data analysis and further environmental management, such as air pollution source localization, health risk assessment, and micro-scale air pollution analysis.

1. Introduction

Due to the rapid development of modern industrial society, coupled with the rise of global climate change and people's environmental awareness, the problem of air pollution has become increasingly serious and has gradually become the focus of the attention of people. In recent years, people and the governments worldwide have noticed the importance and the several negative impacts of the fine particulate matters ($PM_{2.5}$) (Jacob and Winner, 2009; Kan et al., 2012). In addition, there are many high-level air pollution incidents, such as the intensification of sand desertification and the malicious emissions of dishonest factories, which have led to many serious environmental and public health problems (Yu et al., 2012, 2013). These not only become an important issue that needs an urgent solution in Taiwan but have also become the focus of common concerns of all countries in the world, whether developing or developed. United Nations (UN) and World Health Organization (WHO) warned that air pollution in major global cities will pose a grave threat to the lives of millions of people and the deteriorating air conditions make the world's health care services face challenges that are more serious. Governments will, therefore, suffer huge economic losses. An article recently published in the world's top journal "Nature" further pointed out that air pollution caused more deaths annually than malaria and the AIDS virus. In addition, in many countries, the lethality of air pollution is up to 10 times higher than the death toll caused by traffic accidents (Lelieveld et al., 2015).

Airborne Particulate Matters (PM) is usually divided into PM_{10}

* Corresponding author.
E-mail address: yclin@ncu.edu.tw (Y.-C. Lin).

https://doi.org/10.1016/j.envint.2019.105305
Received 7 November 2018; Received in revised form 22 September 2019; Accepted 2 November 2019
Available online 15 November 2019

(particle size $\leq 10\,\mu m$) and $PM_{2.5}$ (particle size $\leq 2.5\,\mu m$) according to their particle size. Among them, $PM_{2.5}$ is easier to suspend in the air due to its finer particle size and its easier generation of chemical reactions with other air pollutants. Once the finer particulate matters enter the respiratory system and can more easily reach deeply in the lungs, even penetrate the alveoli, entering the cardiovascular system and being distributed through the body with the systemic circulation of blood. Therefore, the harm of $PM_{2.5}$ to human health and ecology is more direct than other particulate matters (Yu et al., 2013, 2015). In fact, in recent years, many studies have further pointed out that airborne aerosols, especially $PM_{2.5}$, are the main source that directly or indirectly lead to several fatal diseases such as pneumonia, asthma, allergies, cardiovascular diseases, cancer and even Alzheimer's disease (Atkinson et al., 2014; Leiva et al., 2013; Lelieveld et al., 2015; Tecer et al., 2008; Wu et al., 2015).

Long-term exposure to the human body under high concentrations of $PM_{2.5}$ leads to various diseases easily. Therefore, well monitoring of $PM_{2.5}$ is particularly important for air pollution-related diseases risk assessment and public health. In recent years, due to the rise of environmental awareness coupled with the development of information technology, the world has gradually taken it seriously. Early years, the Taiwan Environmental Protection Administration (Taiwan EPA) did not have any $PM_{2.5}$ observations, but several monitoring stations have gradually started to monitor $PM_{2.5}$ concentration since 2003 in Taiwan. Currently, almost all Taiwan EPA stations have $PM_{2.5}$ observations throughout the air quality observations network, as one of the basic observation key pollutants.

Recently, due to the development of 4G communication, Wi-Fi and other wireless communication technologies, and the popularization of portable devices and various environmental monitoring devices (Atzori et al., 2010; Madakam et al., 2015), the Internet of Things (IoTs) combine with environmental monitoring have become a new scope of environmental science, called Environmental Internet of Things. At present, almost all major developed countries have listed the Internet of Things as an important technology development direction in the country. Including the smart environment, smart healthcare, smart transportation, smart home and smart city, all of which are the key points of the future development of the Internet of Things (Zanella et al., 2014). Therefore, the concept of integrating and applying the Internet of Things and big data analysis is the future trend in the field of environmental science and the key point that we urgently need to develop. Under the framework of IoT, we can systematically analyze the device information originally distributed in different time-space perspectives and continuously collect environmental big data. Then, utilize data mining and spatial-temporal data analysis techniques to extract valuable information from environmental big data and provide them to the government for reference to more accurate policymaking or further academic analysis.

Furthermore, the application of big data analysis has begun to receive the attention of the world under the condition of substantial growth of data volume. In addition to the applications in information science, natural sciences, industrial technology, commercial finance, business marketing, and healthcare, almost all disciplines need to be applied to big data (Chen et al., 2014; Fosso et al., 2015; Labrinidis and Jagadish, 2012; Mayer-Schönberger and Cukier, 2013; McAfee et al., 2012; Sagiroglu and Sinanc, 2013). By using machine learning approach, Zheng, et al. developed an U-Air system to combine several kinds of heterogeneous big data to estimate air quality, such as meteorology, traffic flow, human mobility, structure of road networks, and point of interests (POIs) (Zheng et al., 2013). However, the application of the concept of the Internet of Things and big data in environmental science is still in its infancy, and the relevant literature is still in short supply. Therefore, it will be the future research trend and point that we urgently need to develop (Niu et al., 2013). Especially for environmental-related big data often contains information with time and space. Therefore, in the analysis, we need to consider the spatial-temporal distribution and characteristics in order to further extract the value of environmental data.

Since 2017, the Taiwan EPA plans to invest billions of dollars in funding for the development of tens of thousands of environmental air quality sensing IoTs throughout Taiwan. It is even more understandable that the government attaches great importance to this issue by looking forward to improving the ability of future environmental quality forecasting and management and complementing traditional cost-effective environmental monitoring stations. Recently, an open-source community for low-cost $PM_{2.5}$ micro-air quality sensors built by volunteer public participants called “LASS (Location Aware Sensing System)” was developed by Dr. Chen’s group in the Institution of Information Science, Academia Sinica (Chen et al., 2017). Taipei municipal government and other local governments have also gradually built hundreds of micro-air quality sensors called “AirBox” in various regions. Combining with the Internet of Things and cloud technologies, the spatial and temporal changes of $PM_{2.5}$ concentration are continuously monitored every five minutes. AirBox has the advantages of small, low cost, and easy to carry and install. Therefore, it is possible to monitor areas not covered by conventional stations. At present, the number of such miniature sensors is about hundreds and is still increasing. Thus, if the monitoring data of AirBox can be used for analysis, it will be of great help to the increase of space-time resolution. The drawback of micro-air quality sensors, however, is that they are relatively poor in terms of accuracy and reliability as compared to the instruments used in traditional EPA stations because of their lower cost. In addition, since the public is easy to purchase and install such a micro-air quality sensor, the sensing error can be easily amplified if the micro-sensors are not placed correctly, so the uncertainty is higher.

On the other hand, Taiwan Environmental Protection Administration has set up 77 fixed air quality monitoring stations. There are many monitoring instruments deployed in these stations, which need to be built in the more open space around them. Special personnel carries out regular maintenance, and the construction and maintenance costs are relatively high, which makes it impossible to build in many places. Therefore, in terms of spatial distribution, its density is relatively sparse, and the distribution in various counties and cities is also uneven. Due to the above conditions, the space-time resolution of these traditional stations is relatively low, compared with the micro-air quality sensor, i.e., AirBox. However, the advantages of these instruments are that the sensing instruments of Taiwan EPA used are relatively sophisticated and the measurement errors are small. The data accuracy, precision, and credibility of Taiwan EPA monitoring stations are much higher. These heterogeneous environmental monitoring data from the different scale and sources require further data fusion or data assimilation to integrate huge amounts of heterogeneous data before further big data analysis. Through the integration of environmental big data, we can greatly improve the time and space resolution of environmental monitoring data so that future precision in both analysis and forecasting can be improved.

Therefore, in this study, in order to maintain the advantages of $PM_{2.5}$ observation data from both Taiwan EPA and AirBox, we propose a novel application of data fusion method based on the Optimum Linear Data Fusion theory and integrating with multi-time step Kriging method for spatial-temporal estimation. The main purpose of this study is to use the proposed method to do heterogeneous data fusion from different sources and qualities and improve the spatial-temporal estimation of $PM_{2.5}$ concentration in Taiwan. By combining $PM_{2.5}$ concentration data from AirBox as additional information in our data fusion analyzing procedure, the better and more reasonable estimation results of spatial-temporal $PM_{2.5}$ concentration are expected. Furthermore, under the approach proposed in this study, the improved higher spatial-temporal resolution could provide very useful information for a better spatial-temporal data analysis and further environmental management, such as air pollution source localization, health risk assessment, and micro-scale air pollution analysis. Moreover, the

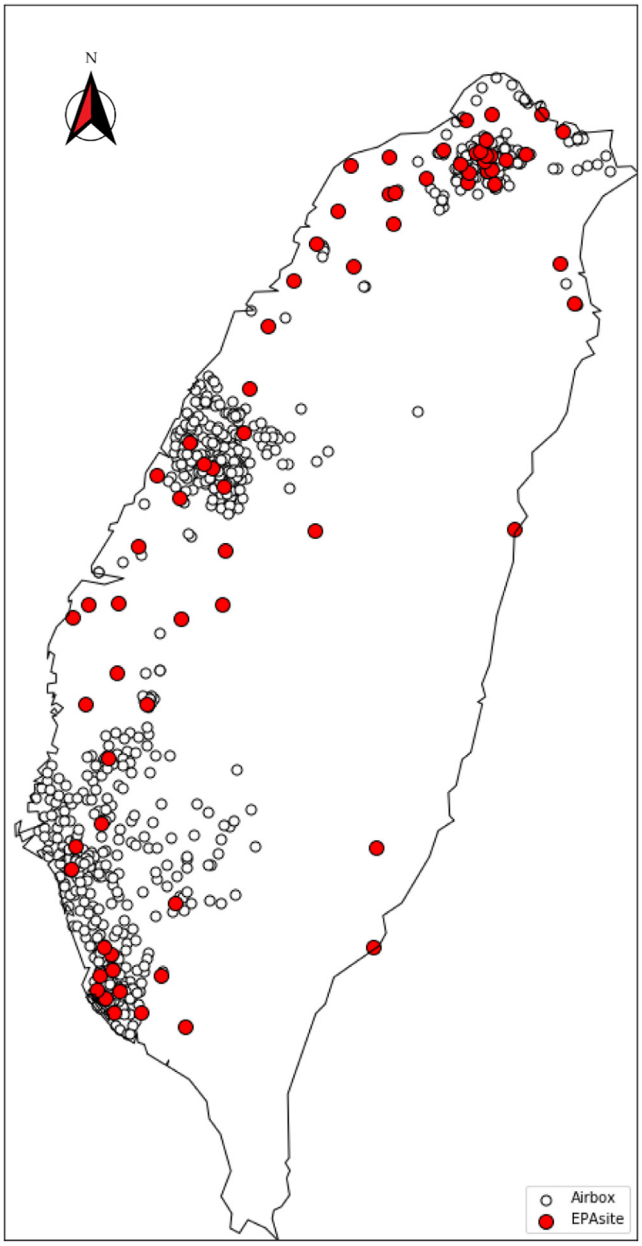

Fig. 1. The red solid dots in Fig. 1 are the Taiwan EPA air quality monitoring station, and the white dots are the AirBox micro-sensors. It can be seen that the distribution and density of these two sensors are significantly different.

most important objective of this study is to make large AirBox data more credible and scientific value.

2. Study area and materials

The study area of this study is the main island area of Taiwan, which is located in East Asia and surrounded by the Pacific Ocean and Taiwan Strait (Fig. 1). In the figure, the areas where the sensors are dense are all the major metropolitan areas in Taiwan, i.e., Taipei, Taichung, and Kaohsiung metropolitan area, from Northern to Southern of Taiwan. Among them, the red solid dots in Fig. 1 are the Taiwan EPA air quality monitoring station, and the white dots are the AirBox micro-sensors. In this study, we excluded three EPA stations built on the offshore islands. Therefore, the total number of EPA stations used is 74. It can be seen that the distribution and density of these two sensors are significantly different. The device appearances of two commonly used Airboxes including Edimax Airbox and LASS Airbox are shown in Fig. 2, which are also used in this study.

The experimental datasets were retrieved from the Taiwan Environmental Protection Agency and AirBox micro sensors data from EDIMAX Technology Co., Ltd. In this study, the hourly $PM_{2.5}$ automatic monitoring data from Taiwan EPA monitoring station and 5 min sampling frequency of 1176 AirBox micro sensors data from October 14 to 27, 2016 were selected. From basic statistics (Table1), the variance of micro-sensor AirBox data is significantly larger than the Taiwan EPA monitoring data. Moreover, several values are outliers that do not conform to common values for AirBox data. The rates of missing value are 1.24% and 22.86% for Taiwan EPA monitoring data and AirBox data, respectively. Therefore, datasets are pre-processed to filter outliers missing values for AirBox data. In this study, the interpolation method is used for processing. Table 1 is the statistical information of the pre-processed data.

In terms of space-time statistical properties, Fig. 3 shows the empirical spatial-temporal covariance and their corresponding fitted exponential covariance model of Taiwan EPA monitoring data and AirBox data. From the empirical spatial-temporal covariance, the range of AirBox data is much smaller than Taiwan EPA monitoring data, either in time or in space. On the other hand, the sill of AirBox data is much higher than the Taiwan EPA monitoring data. It shows the high variability and instability of the AirBox compared to the Taiwan EPA data in both time and space.

3. Methods

When observing the same state of air pollutants, the observation data generated by different sensors often have uncertainties for various reasons, especially for low-cost sensors (AirBox). To obtain accurate results, that is, the best estimation of the state, it is necessary to fuse the

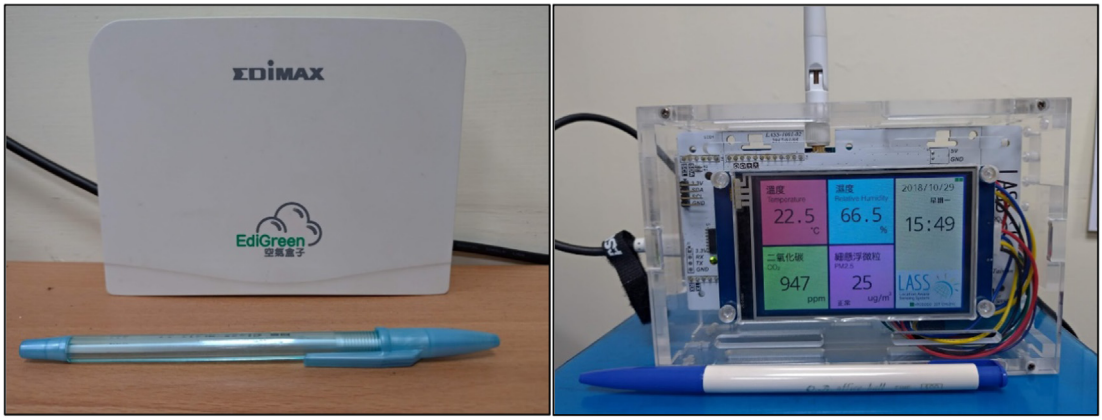

Fig. 2. The device appearances of two commonly used Airboxes. Left: EDIMAX Airbox, right: LASS Airbox.

Table 1
The comparison of basic statistics information of Taiwan EPA Monitoring Station and AirBox $PM_{2.5}$ concentration data from October 14 to 27, 2016.

Variable	Unit	Range	Mean	St. Dev	Sampling Frequency	Number of Stations
$PM_{2.5}$(EPA)	$\mu g/m^3$	[2.0, 142.0]	20.985	17.064	1 h	74
$PM_{2.5}$(AirBox)	$\mu g/m^3$	[1.0, 211.0]	36.430	23.585	5 min	1176

data of different sensors. Multi-sensor data fusion technology has been applied in many fields. Optimum Linear Data Fusion is one of the algorithms based on the least squares method for multi-sensor data fusion. This algorithm can obtain good estimation results when the multi-sensors are independently observed.

Based on this method, The overall procedure of this study combines the heterogeneous data generated by the above two different sources of $PM_{2.5}$ observation data from both Taiwan EPA and AirBox (low-cost sensor) by combining Ordinary Kriging to perform the spatial-temporal estimation of $PM_{2.5}$ concentration. Because the monitoring locations of Taiwan EPA and AirBox are different, Ordinary Kriging is used to estimate the spatial-temporal data into the same sites. Finally, we obtain an optimal estimation of the higher resolution of $PM_{2.5}$ space-time variation maps by the Optimal Linear Data Fusion (OLDF) technique.

3.1. Optimal Linear data fusion (OLDF)

This method combines the independent observations of different sensors to obtain the best state estimation and suitable weights for all the component retrieval systems by observing the error (Li, 2012; Wu et al., 2011). Based on Least Square (LS) approach, the Optimal Linear Data Fusion (OLDF) technique is to fuse and combine signals from different sensors by optimal estimation algorithm. Assume two independent sensors, i and j, observing the same object, e.g., $PM_{2.5}$ concentration in the same location. The observation result are x_i, x_j, respectively. It should satisfy the unbiased condition but have different variances σ_i^2 and σ_j^2, as following equation:

$$\begin{cases} E[x_i] = \theta, \quad Var[x_i] = \sigma_i^2 \\ E[x_j] = \theta, \quad Var[x_j] = \sigma_j^2 \end{cases} \tag{1}$$

To obtain a new estimated value x_{ij}, the two observations from different sensors x_i, x_j are linearly fused using a weighted average method:

$$x_{ij} = ax_i + bx_j \tag{2}$$

where a and b are weights, and $b = 1 - a$. Solving a and b such that the variance of the new estimated value x_{ij} should be less than the variance of the original two observations, namely:

$$Min \quad Var[x_{ij}] = E[x_{ij} - \theta]^2 \tag{3}$$

since x_i and x_j are independent observations, (Eq. (3)) is equivalent to:

$$Min \quad Var[x_{ij}] = a^2\sigma_i^2 + b^2\sigma_j^2 \tag{4}$$

using Lagrangian multiplier to find the maximum value of Eq. (4):

$$a = \frac{\sigma_j^2}{\sigma_i^2 + \sigma_j^2}, \quad b = \frac{\sigma_i^2}{\sigma_i^2 + \sigma_j^2} \tag{5}$$

then, the new estimated value x_{ij} is:

$$x_{ij} = \frac{\sigma_j^2 x_i + \sigma_i^2 x_j}{\sigma_i^2 + \sigma_j^2} \tag{6}$$

It can be seen from Eq. (6) that the the weight is inversely proportional to the number of variances. The larger the variance of sensors, the smaller the weight. The variance of the new estimated value is:

$$\mathrm{Var}[x_{ij}] = \left(\frac{1}{\sigma_i^2} + \frac{1}{\sigma_j^2}\right)^{-1} \tag{7}$$

Eq. (7) shows that after the fusion process, the variance of new estimated value is smaller than the variance of any of the original observations. Thus, the new fused optimal estimated value is derived based on two originally independent observations.

3.2. Modified spatial-temporal estimation technique by Kriging

The Kriging method is one of the most important Spatial Interpolation Methods in Geostatistics and has been widely used in soil, groundwater, atmosphere, image processing, and many other fields. The theory was originally proposed by Krige (1952), and later researched and developed by the French mathematician Georges

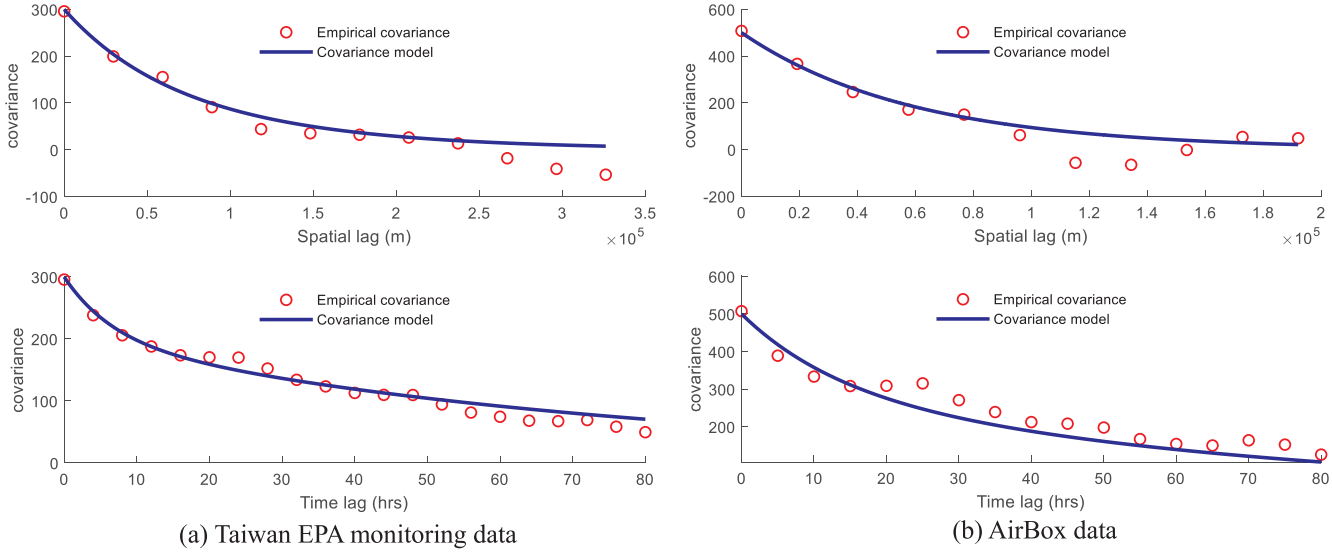

Fig. 3. The empirical spatial/temporal covariance and their corresponding exponential covariance model of (a) Taiwan EPA monitoring data and (b) AirBox data.

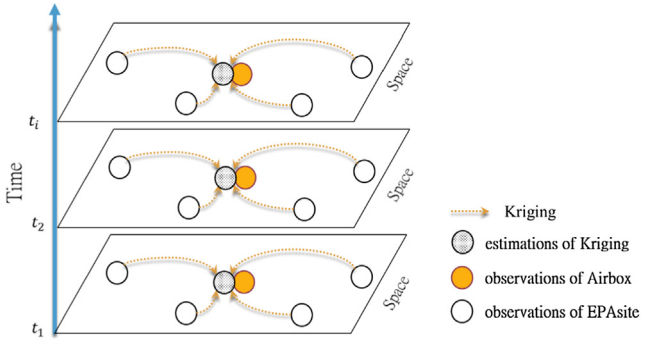

Fig. 4. The framework of multi-sensor space-time data fusion based on Ordinary Kriging and Optimal Linear Data Fusion and Modified Spatial-temporal Estimation Technique by Kriging.

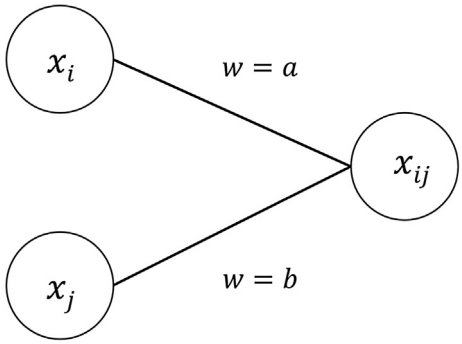

Fig. 5. The concept figure of Optimal Linear Data Fusion (OLDF).

Matheron, and named Kriging based on Krige's name (Matheron, 1963).

Supposed there is a temporal dependency in time slices for each station, so we only need to make spatial estimates based on spatial dependence, i.e., the Kriging method. Here in this study for spatial-temporal estimation, we modified the Kriging method for each time step one by one to perform not only space but also time estimation. In other words, we applied Kriging spatial estimation for each time slice, i.e., $t = 0$, $t = 1$, $t = 3$,..., $t = n$, by using loop-Kriging estimation based on Python programming language. Refers to Zheng et al. (2013), the basic concept is shown in Fig. 4.

Generally, for the spatial distribution of air pollutant concentration, the atmospheric dispersion model is commonly used for computer simulation to obtain the approximate spatial distribution of pollutants. However, with the development of modern industrial society, the sources of pollutants and environmental changes in the atmosphere become more and more complex, and the parameters required by the model are increasing, which not only leads to an increase in computational cost and uncertainty but also reduces the accuracy of the model. With the development of increasing monitoring stations and low-cost sensors of continuously spatial-temporal air quality monitoring, therefore, some researchers later investigated the spatial distribution of air pollutants from the perspective of geostatistics based on real observations (Anselin and Le Gallo, 2006; Bayraktar and Turalioglu, 2005; Li and Heap, 2014; Wong et al., 2004). The results based on data-driven also show that the Kriging method can describe the reasonable spatial distribution of real air quality.

The Kriging method uses the observations of known points around it to weight the estimated values of unknown points. The estimated results satisfy the Best Linear Unbiased Estimation (BLUE). Different basic assumptions also determine the different types of Kriging, such as Simple Kriging, Ordinary Kriging, and Universal Kriging. The Ordinary Kriging method is used in this study. The basic theory will be introduced from the assumptions, the Best Linear Unbiased Estimation, and the semi-variogram function.

The Ordinary Kriging method assumes that the spatial random variable $Z(x)$ has a second-order stationarity assumption, namely:

$$E[Z(x)] = E[Z] = u \tag{8}$$

$$Var[Z(x)] = \sigma^2 \tag{9}$$

The Best Linear Unbiased Estimation (BLUE) has the following three conditions:

a. Linearity: The Ordinary Kriging linearity estimation is:

$$\hat{Z}_o = \sum_{i=0}^{N} \lambda_i Z_i \tag{10}$$

$$\sum_{i=0}^{N} \lambda_i = 1 \tag{11}$$

where Z_i is the observed value at the known point x_i, $\hat{Z}_o$ is the estimated value at the unknown point x_o, and λ_i is the weight of the point x_i to the point x_o, and the sum of the weights should be 1.

b. Optimization: Minimize the variance of the difference value between the estimated and observed values:

$$\text{Min} \quad Var(\hat{Z}_o - Z_o) \tag{12}$$

c. Unbiased: The expected value of the difference between the estimated and observed values is 0:

$$E(\hat{Z}_o - Z_o) = 0 \tag{13}$$

For the semi-variogram function, assume that any two points x_i and x_j in the space, the Euclidean distance between these two points are d_{ij}, and the expected value and the variance are only related to the distance, and regardless of the positions of the points, Eq. (14) expresses the semi-variogram and covariance between any two points in the space.:

$$\gamma(d_{ij}) = \frac{1}{2} E[(Z_i - Z_j)^2] \tag{14}$$

where Z_i and Z_j are the values at the points x_i and x_j, respectively, and $\gamma(d_{ij})$ is a semi-variogram. Ordinary Kriging assumes that the spatial random variable is second-order stationary, and the semi-variant $\gamma(d_{ij})$ and the covariance $cov(d_{ij})$ are complementary functions. In order to obtain λ_i so that it satisfies Eqs. (12) and (13), and according to the relationship with complementarity, Lagrangian Multiplier, μ can be introduced to obtain the matrix of weight coefficients:

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} & 1 \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma_{1o} \\ \gamma_{2o} \\ \vdots \\ \gamma_{no} \\ 1 \end{bmatrix} \tag{15}$$

After obtaining λ_i, we can substitute into Eq. (10) to obtain the estimated value $\hat{Z}_o$.

3.3. Multi-sensor space-time data fusion framework

In this study, we proposed the multi-sensor space-time data fusion framework based on Ordinary Kriging and Optimal Linear Data Fusion (OLDF) to perform the weighted integration of data from the EPA monitoring station and the micro-sensor AirBox. The basic concept is shown in Fig. 4. Where t_i is the hourly time series value, that is, the two observations x_i and x_j of the two sensors are weighted and combined to obtain a new best estimate x_{ij}. See Fig. 5. According to the optimal linear data fusion method, the weights a and b depend on the variance of the observations (AirBox) or estimated observations by the modified spatial-temporal estimation technique by Kriging (EPA), that is, the

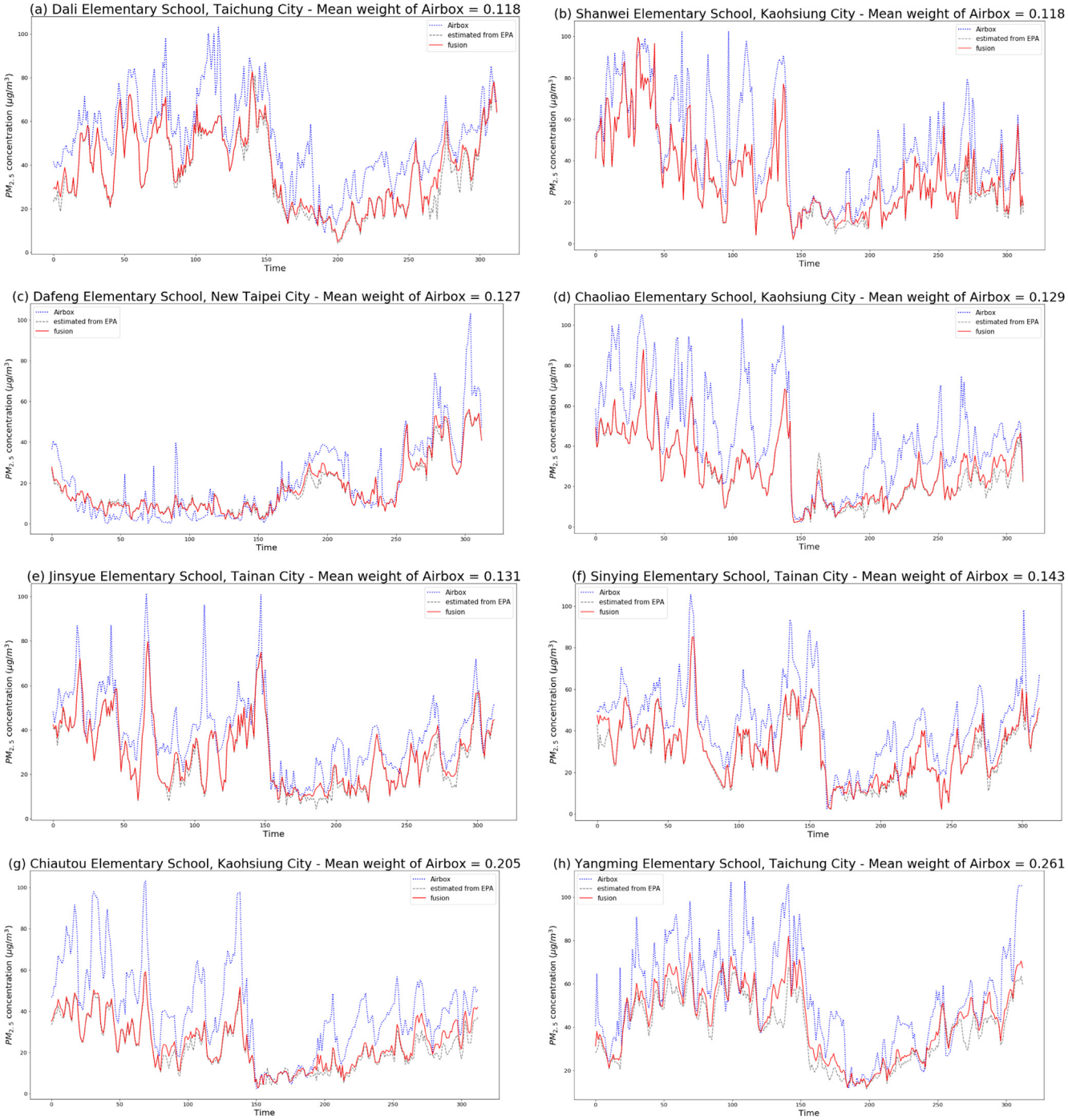

Fig. 6. The time series for AirBox (blue dashed line), Kriging estimated from Taiwan EPA (gray dashed line), and Fusion data (red solid line) at representative $PM_{2.5}$ AirBox micro-sensor stations with "low" average AirBox weights. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

variance of x_i and x_j. In this study, the variance of $PM_{2.5}$ of the two sensors is measured in days. Through this framework, the data of AirBox becomes soft information, that is, where there are many EPA stations in the vicinity, their variance of spatial estimation is naturally small, so the fusion will be close to the EPA value; in the area lacks of EPA stations, it can not only rely on the information provided by the Airbox, but also retain the large-scale EPA trend changes information. On the other hand, the weight is also affected by the variance of the AirBox itself, so it is also possible to exclude the AirBox information that has problems, high variability, or many extreme values, and to extract more credible data added to the fusion process.

4. Results

The results of the fused data combining Taiwan EPA and AirBox data by using multi-sensor space-time data fusion framework are demonstrated as the following figures. The fusion estimation is applied for the entire country. In order to demonstrate the fusion results of different regions and weights, we selected 24 (8×3) AirBox sites with high data integrity as representatives from a total of 1176 stations according to 3 different average AirBox weights: low, medium, and high. Figs. 6–8 show the demo of fused time series data at following $PM_{2.5}$ AirBox micro-sensor sites with low, medium, and high average AirBox

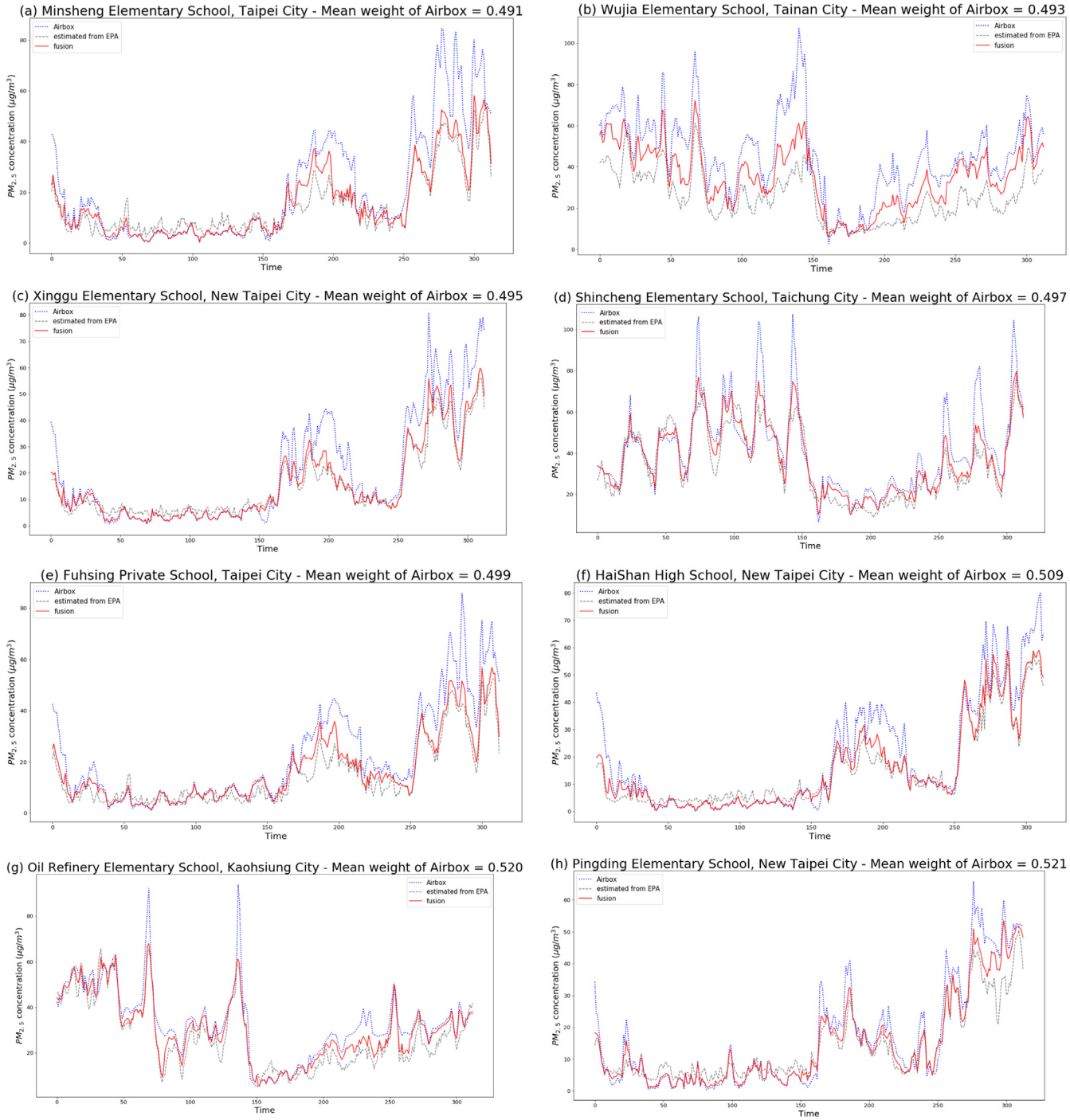

Fig. 7. The time series for AirBox (blue dashed line), Kriging estimated from Taiwan EPA (gray dashed line), and Fusion data (red solid line) at representative $PM_{2.5}$ AirBox micro-sensor stations with "medium" average AirBox weights. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

weights, respectively. Since the weights are updated with the daily calculations, we use the average weight of each site for analysis when selecting. The short blue dashed line is the observed value of the AirBox micro-sensor at this site. The gray long dashed line is the observed value from Taiwan EPA estimated by the spatial-temporal Kriging estimation method to this site. The red solid line is the result of fusion value combining Taiwan EPA and AirBox data based on the proposed framework in this study.

It is obvious that the variations of time series and overall trends for both Taiwan EPA and AirBox are similar in every location, however, the values are different. In general, the value of AirBox (blue dashed line) is higher than Taiwan EPA (gray dashed line). However, sometimes the value of the AirBox will be lower than the value of EPA, especially in the case of low concentrations (e.g., Figs. 7a, e, f, h, 8d, g). The values of the AirBoxes are more extreme at extreme values, that is, at high concentrations of $PM_{2.5}$, AirBoxes tend to measure higher than EPA, while on the other hand, at low concentrations, most of them are lower than EPA. Furthermore, the variation also shows that the measure of the dispersion of Airbox (Std. Dev. = 23.585) is higher than Taiwan EPA monitoring stations (Std. Dev. = 17.064), which is also quantified by Standard Deviation shown in Table 1. Therefore, we can find that both time series measurements continue to remain inconsistent.

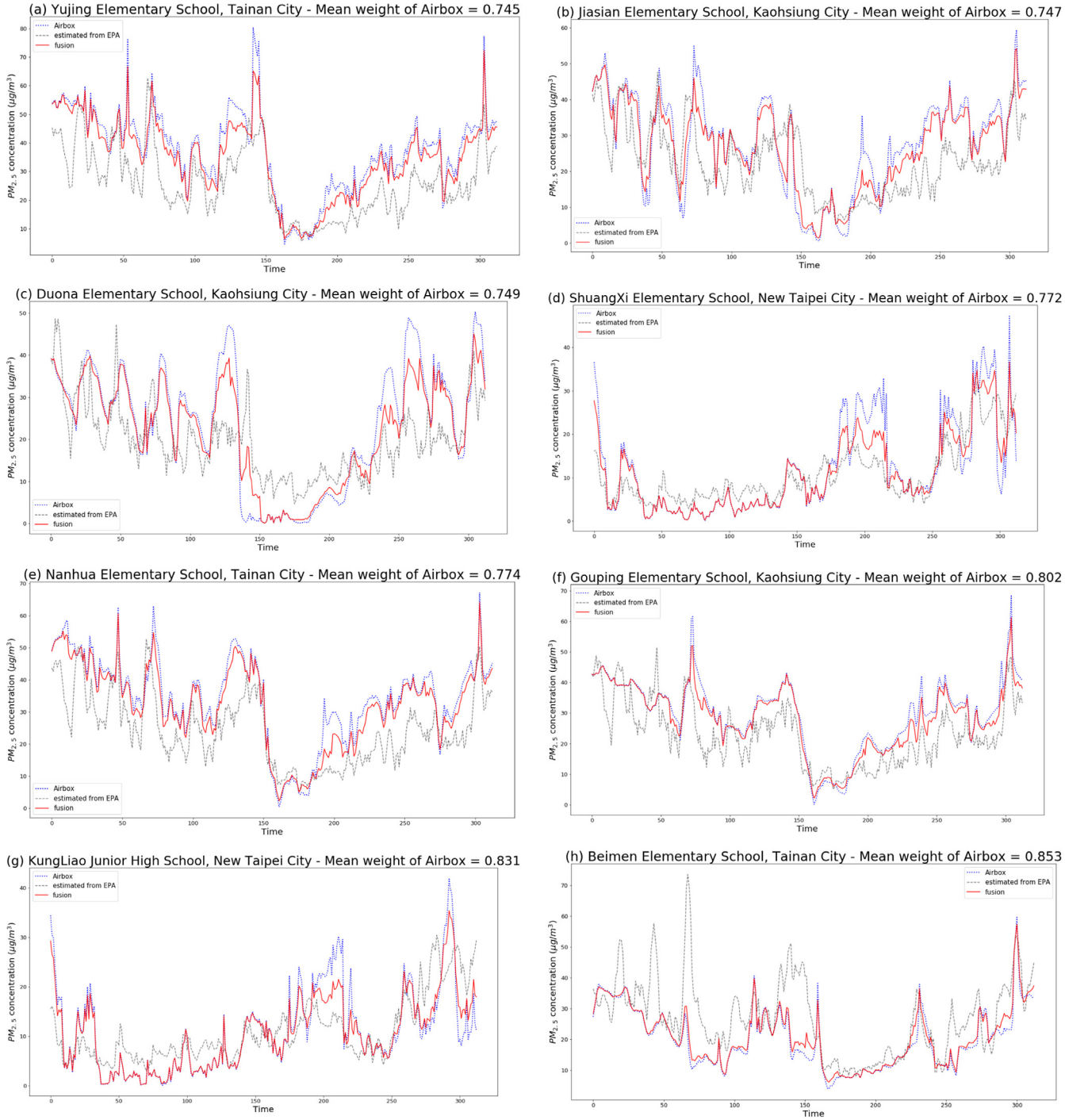

Fig. 8. The time series for AirBox (blue dashed line), Kriging estimated from Taiwan EPA (gray dashed line), and Fusion data (red solid line) at representative $PM_{2.5}$ AirBox micro-sensor stations with "high" average AirBox weights. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 6 shows the time series for AirBox (blue dashed line), Kriging estimated from Taiwan EPA (gray dashed line), and Fusion data (red solid line) at representative $PM_{2.5}$ AirBox micro-sensor stations with "low" average AirBox weights, which means that the fusion result may generally follow with the variation of EPA. Most of these AirBox stations are located very close to the EPA station. Under the condition of the low variance of EPA, the fusion results are naturally biased towards EPA, which is what we expected. In other words, where the EPA has a measurement, the framework chooses to believe that using complex and accurate measuring instruments and methods by EPA more to produce our fusion results, and their degree of variation is relatively small, so these areas have higher EPA weights. The AirBox information with less data quality can be used here as auxiliary soft information. For example, the Dali Elementary School in Fig. 6a, the AirBox is located just 165 m away from the EPA station (Dali District Office). Similarly, Fig. 6b Shanwei Elementary School and EPA Linyuan Station; Fig. 6c Dafeng Elementary School and EPA Xindian Station; Fig. 6d Chaoliao Elementary School and EPA Daliao Station; Fig. 6f Sinying Elementary School and EPA Sinying Station, they are all located in the same corresponding campus. Fig. 6e Jinsyue Elementary School and EPA Tainan Station (Zhongshan Junior High School) are also close to each other, and the distance between the two campuses is only about 150 m.

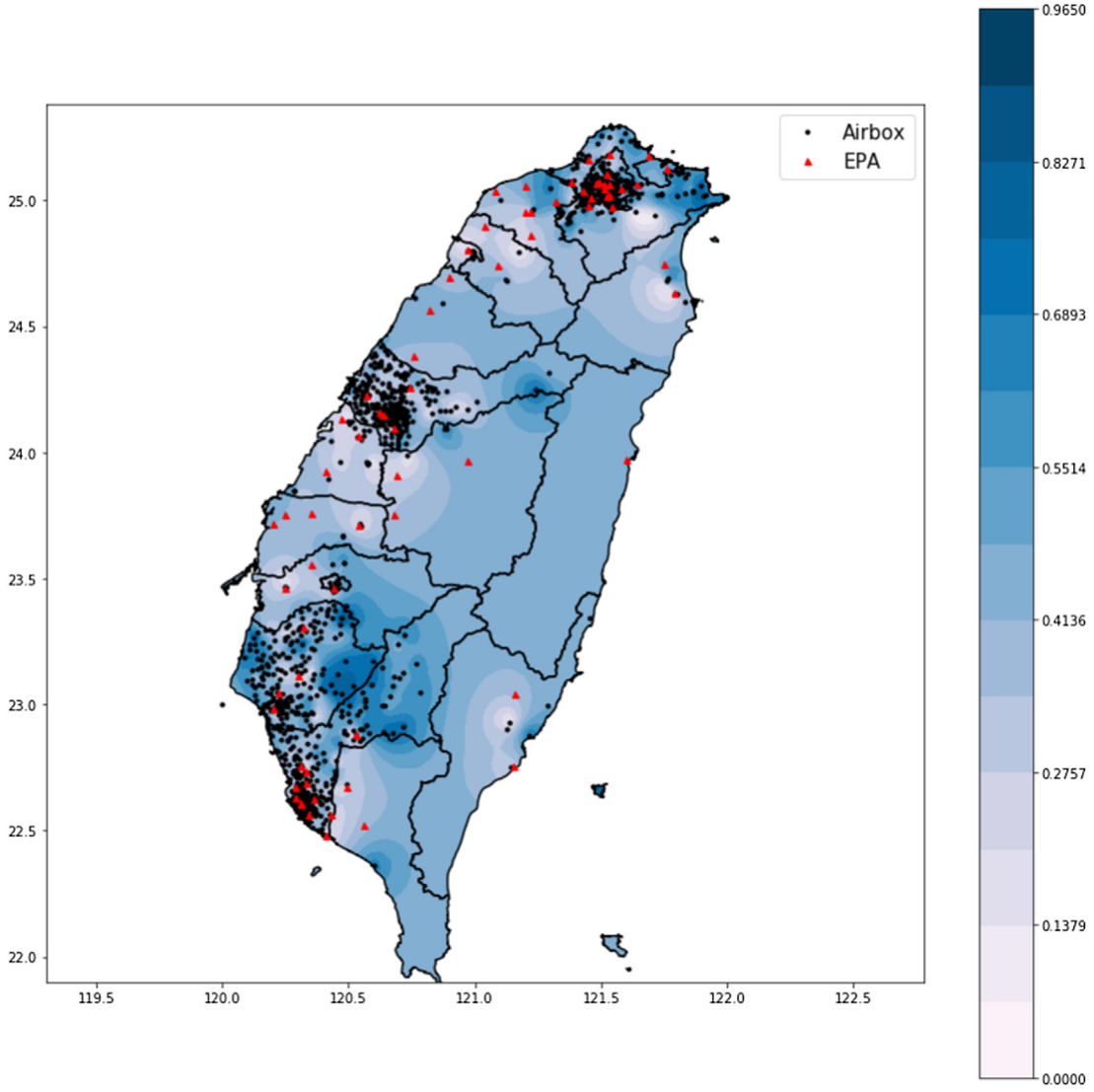

Fig. 9. The map of the average weight distribution of the AirBox.

Furthermore, Fig. 6g the distance between Chiautou Elementary School and EPA Chiautou Station (Chiautou District Office) is also less than 400 m. Thus, most of them are exactly located in the same or neighbor campus. Although the average weight of the AirBoxes in these stations is lower in Fig. 6 and the variation after fusion is generally close to the changes in EPA. However, if EPA is unreasonably changing at certain times and even inconsistent with the changes in AirBoxes, the results of the fusion will automatically produce reasonable adjustments. These phenomena are also shown in Fig. 6 for all representative stations.

As the distance from the EPA station increases, the weight of the Airbox gradually increases. For example, the medium mean weight of Airbox about 0.5 (Fig. 7), they are all about 3–5 km from the nearest EPA national station. Furthermore, many of these stations are still located in urban areas, but no national-level stations provide local monitoring, so AirBoxes provide good reference information in these areas and provide nearly half of the weight, which is also a future recommendation as a very important reference area for Taiwan EPA to build new stations. In Fig. 7, generally, the overall trend of the two time series is similar, but the Airbox has higher values especially during peak events. The result of the fusion in Fig. 7 also retains the common trend changes in the two time series and adjusts many extreme values. Moreover, through the characteristics of the dynamic weight used in this framework, all the fusion results will not exactly lie in the middle, even though their average weight is about 0.5. It will automatically adjust the results of the fusion with the daily variation to determine which one should be biased more.

As for the High AirBox weight stations shown in Fig. 8, they are farther away from the EPA national stations, mostly in the suburbs, even close to the mountains, but there are still many people living. For example, Fig. 8a Yujing Elementary School is 16.5 km from the nearest EPA Tainan Shanhua Station, Fig. 8b 22.71 km from the nearest EPA station to Jiasian Elementary School, and Fig. 8c 19.19 km from the nearest EPA station to Duona Elementary School. Therefore, through the framework proposed in this study, these areas with lacking national monitoring data are more depending on the information provided by the AirBox. After fusion, we can also incorporate some of the information provided by the EPA station.

By using the multi-sensor space-time data fusion framework shown in Figs. 6–8, we take into account the advantages of the two measurement data and find optimized fusion time series based on their different reliability in different trends, variation, and value changes. The red line can be seen as a more reasonable time series for this station, adjusting the uncertainty of the AirBox. If the two time series have different trends, for example, for the sudden troughs that appeared for Taiwan EPA data or the peaks of AirBox in Fig. 7, the fusion data avoid the impact of high variability data and adjust them into data that are more

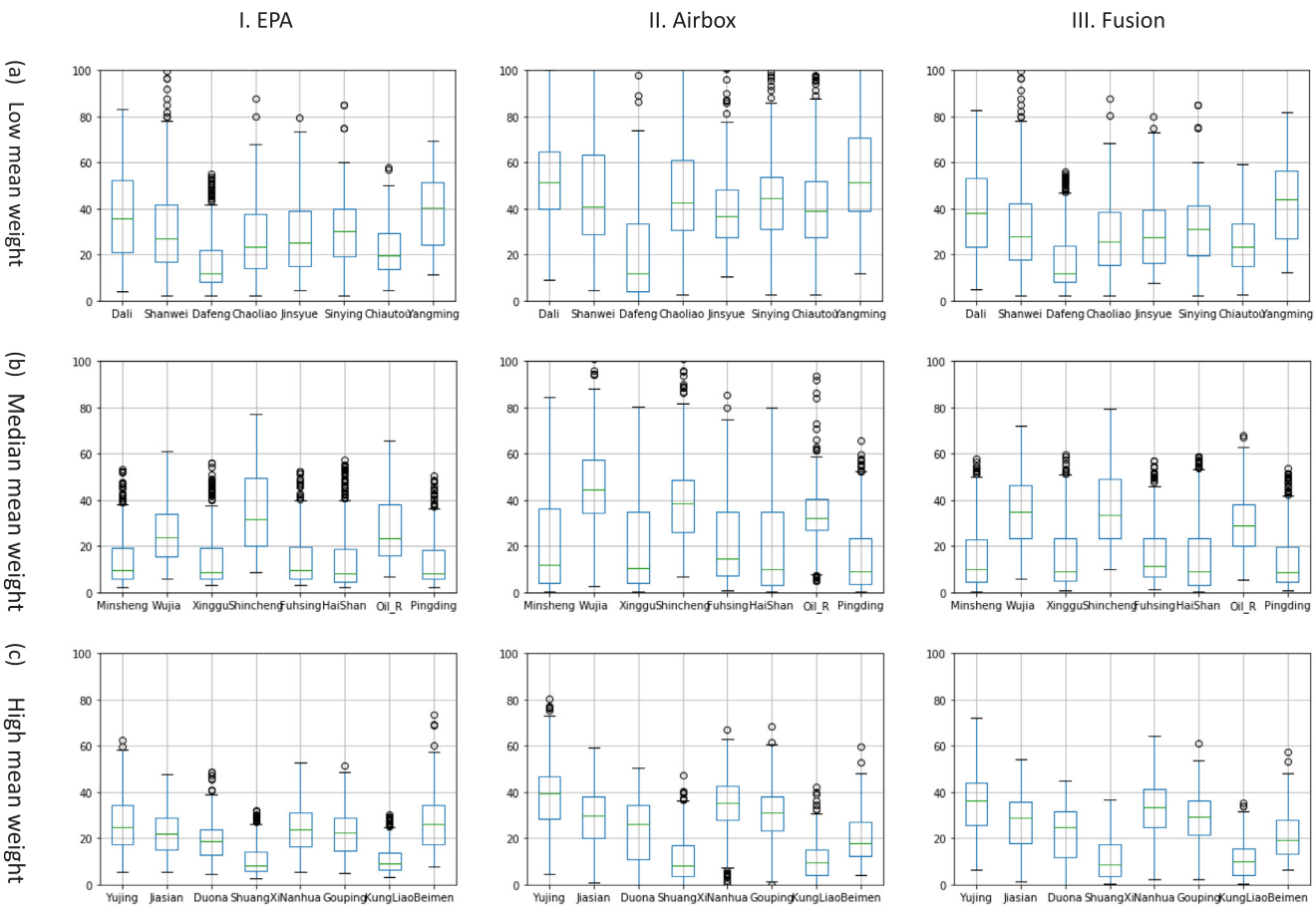

Fig. 10. The boxplots for AirBox, Kriging estimated Taiwan EPA, and Fusion data at demo $PM_{2.5}$ AirBox micro-sensor stations corresponding with the representative stations shown in Figs. 6–8. y-axis is $PM_{2.5}$ concentration ($\mu g/m^3$).

Table 2
The *p*-values of testing statistical significant differences between EPA data and fusion result by two-tailed *t*-test.

(a) Low mean weight	Dali	Shanwei	Dafeng	Chaoliao	Jinsyue	Sinying	Chiautou	Yangming
	0.2568	0.5615	0.6580	0.3283	0.3090	0.2267	0.0172	0.0014
(b) Median mean weight	Minsheng	Wujia	Xinggu	Shincheng	Fuhsing	HaiShan	Oil_R	Pingding
	0.3483	0.0000	0.5065	0.0615	0.0290	0.5561	0.0254	0.4399
(c) High mean weight	Yujing	Jiasian	Duona	ShuangXi	Nanhua	Gouping	KungLiao	Beimen
	0.0000	0.0000	0.0031	0.7858	0.0000	0.0000	0.3630	0.0000

reliable. If the two time series have similar trends, then this method can find an eclectic weight adjustment to optimized spatial-temporal data. Furthermore, the optimized fusion data also considered spatial variability from Kriging estimation. It can automatically and consistently comply with EPA's stable and accurate monitoring information in the vicinity of the EPA station area; in the areas that are lack of the EPA station information, it automatically follows the information provided by AirBox, but can also maintain the overall trend provided by the EPA. In addition, not only the representatives mentioned in the article, we apply the fusion procedure for the entire study area. Therefore, there are also many stations located between these representative stations, and we have provided the best-fused information to improve the spatial resolution of $PM_{2.5}$ concentration and understand their large-scale overall distribution and local changes.

In order to demonstrate that the average fusion weight of the data calculated based on this framework is related to the distribution of the EPA stations and understand the spatial distribution of the weights, the average weight distribution map of the AirBox is shown in Fig. 9. The AirBox sites with lower weights are closer to the EPA station, so their fusion results tend to similar to the EPA stations due to smaller variance, so the weight is smaller, as shown in Fig. 6. On the other hand, the higher weights of Airbox stations represent the $PM_{2.5}$ concentration estimated from EPA stations that are less credible here due to the areas are far away from the EPA stations, so it relies on the soft information provided by the Airbox (Fig. 8). In Fig. 9, the darker the color, the higher the average weight of the Airbox at the Airbox site, mostly in the suburbs or areas where there is no EPA station nearby.

The EPA stations are currently mostly located in the urban area. The number of suburbs is quite insufficient. Although the setup of Airbox has crossed the boundaries of the urban area, there are already many Airboxes installed in the suburbs, which are farther away from the EPA station. At present, due to the convenience of people's installation, densely populated areas are still the main place for deployment. However, from the process of data fusion in this study, it can be seen from the weight distribution map (Fig. 9) that the area with the highest average weight of the Airbox represents the lack of reliable EPA stations in the local neighborhood to provide more accurate $PM_{2.5}$ concentration information, so rely on the information provided by the Airbox. For

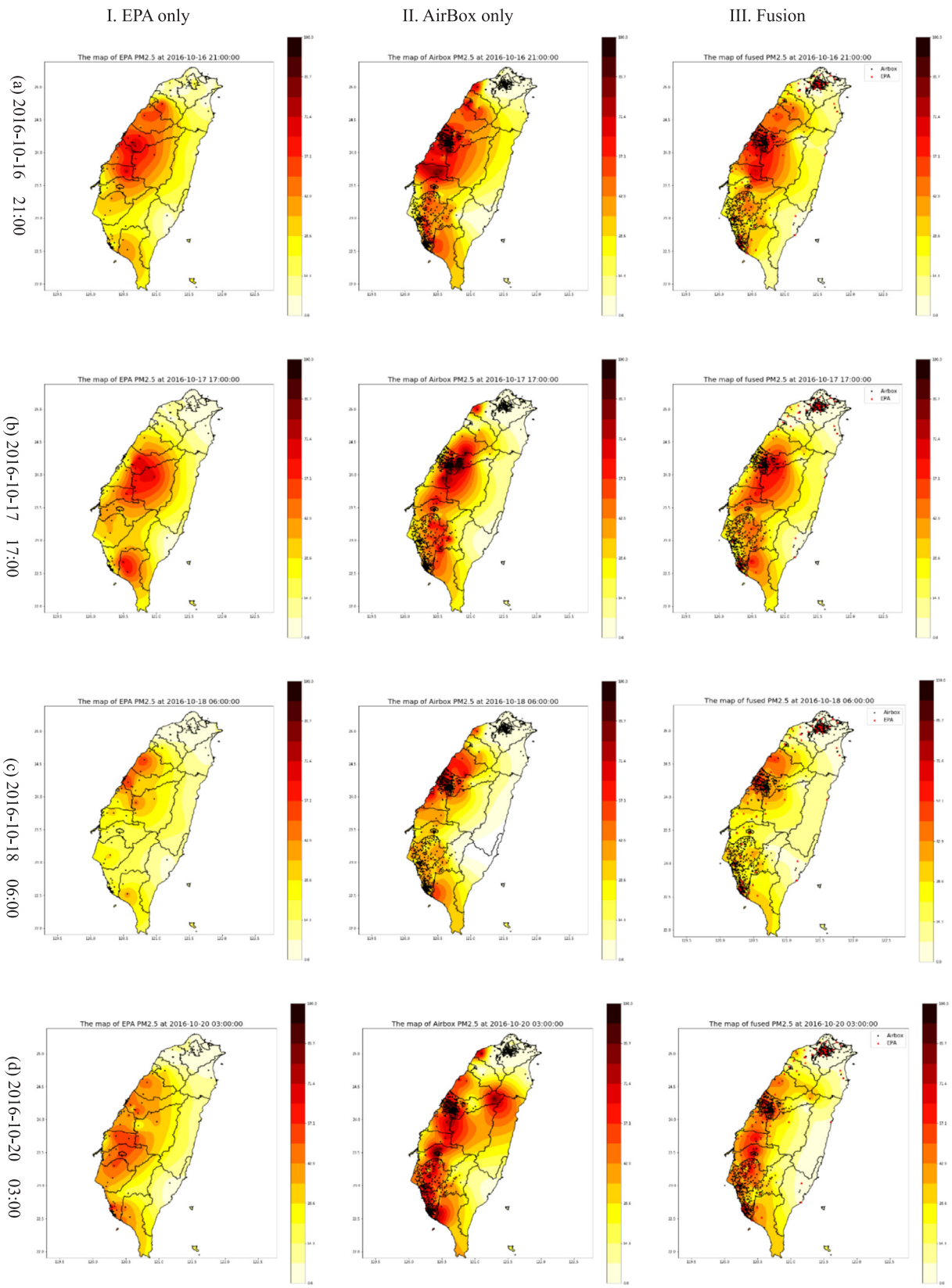

Fig. 11. The spatial distribution maps of $PM_{2.5}$ concentration ($\mu g/m^3$) before and after using multi-sensor space-time data fusion framework at different time-slices: (a) 2016-10-16 21:00 (b) 2016-10-17 17:00 (c) 2016-10-18 06:00 (d) 2016-10-20 03:00. I. only EPA data used, II. only AirBox data used, III. fused data for generating the spatial estimation map.

Table 3
The statistical comparison analysis of the spatial pattern differences for Fig. 11, including the mean difference in pixel scores, standard deviation, *t*-value, *p*-value for the *t*-test, percent difference, and Kappa value.

Time	Comparison maps	*Mean*	*Std. Dev*	*t value*	*Pr* > |*t*|	%	*Kappa*
2016-10-16 21:00	EPA-Fusion	12.7645	63.7304	11.9536	< 0.0001	15.5533	0.7778
	EPA-Airbox	22.7048	58.8530	20.9410	< 0.0001	12.7999	0.7427
	Airbox-Fusion	−9.9403	48.5069	−8.7776	< 0.0001	12.7833	0.7451
2016-10-17 17:00	EPA-Fusion	11.8980	64.2280	11.2049	< 0.0001	14.2485	0.7681
	EPA-Airbox	13.9184	63.2466	12.9846	< 0.0001	15.7957	0.7078
	Airbox-Fusion	−2.0205	45.0432	−1.8058	0.0709	9.7915	0.7716
2016-10-18 06:00	EPA-Fusion	12.3172	102.5425	20.6707	< 0.0001	27.9124	0.5612
	EPA-Airbox	15.2742	49.6974	25.3301	< 0.0001	8.8158	0.6937
	Airbox-Fusion	−2.9570	101.2849	−4.6359	< 0.0001	28.5790	0.5157
2016-10-20 03:00	EPA-Fusion	10.0367	61.1571	10.4218	< 0.0001	16.2009	0.7308
	EPA-Airbox	31.7141	59.7726	31.7851	< 0.0001	16.7659	0.6995
	Airbox-Fusion	−21.6774	50.9087	−20.7777	< 0.0001	18.6654	0.7124

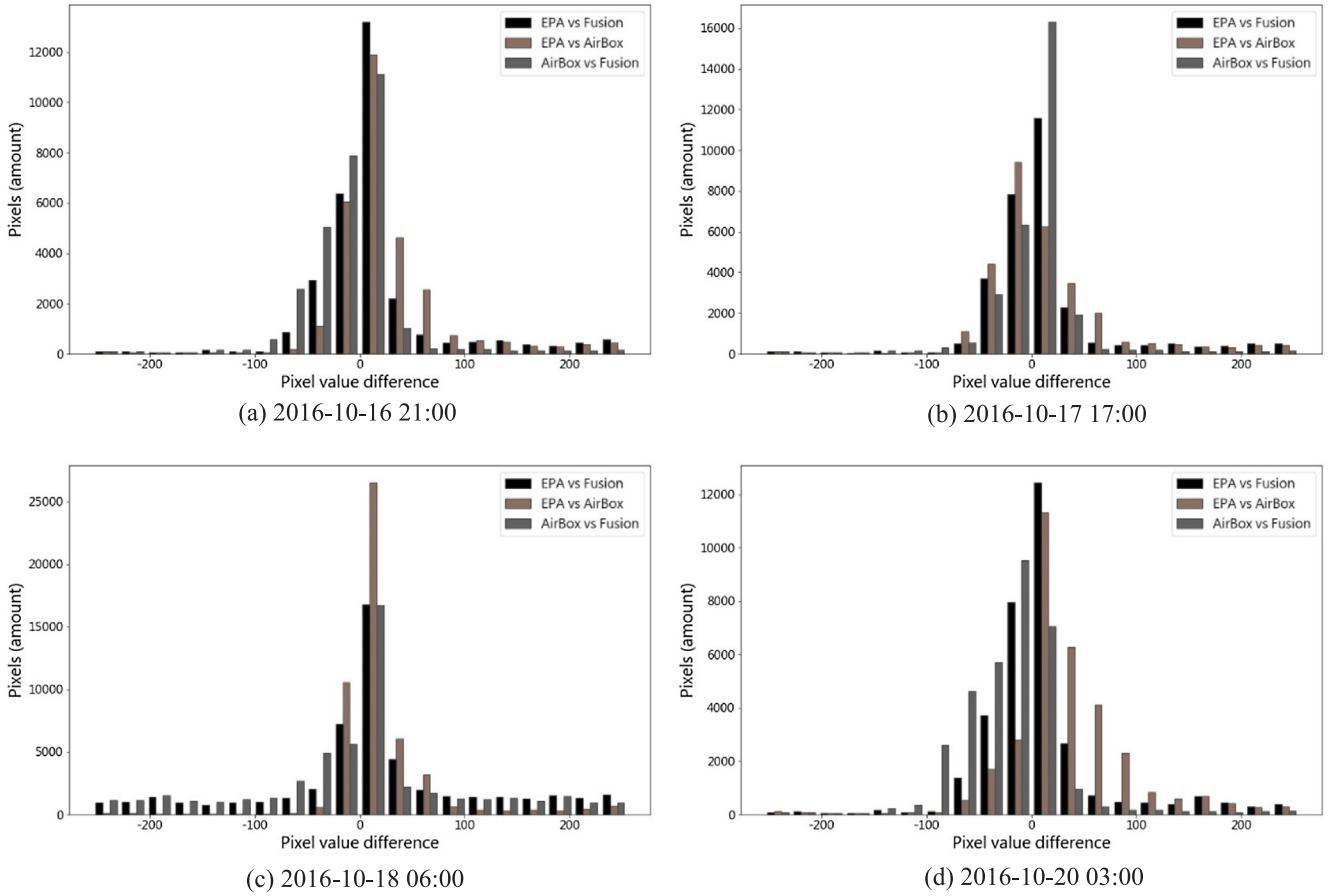

Fig. 12. The distribution of pixel score differences between before fusion and after fusion to test the significance of spatial pattern differences for Fig. 11.

example, the suburbs of Tainan City, Kaohsiung City, and New Taipei City are the majority.

Corresponding to the representative stations shown in Figs. 6–8, which are demonstrating the fusion effects under different mean weights, the boxplots are shown in Fig. 10. Boxplots show the difference and performance in the distribution of data before and after data fusion quantitatively. Compared with the Taiwan EPA data, AirBox raw data has a higher mean, variance, and more outliers. After the fusion process, the mean and variance are between the two time series, and fewer outliers, especially for the median weight cases.

Furthermore, in Table 2, we exam the statistical significance of the difference between EPA data and fusion results for the representative stations shown in Fig. 10 based on *t*-test. From the result of *p*-values, it shows that several stations have very significant differences from the original EPA data spatial estimation after the fusion process, especially for the stations with the high weight of the Airbox, and also for some stations with median and low weight. For example, for the two-tailed significance level of $p = 0.05$, Chiautou, Yangming, Wujia, Fuhsing, Oil_R, Yujing, Jiasian, Duona, Nanhua, Gouping, and Beimen stations are very significant. ShuangXi and KungLiao with high weights are no significant due to there are no significant differences between EPA and Airbox, therefore, the fusion results for these two stations present no significant difference from the EPA. Overall, for all stations, the testing result shows significant differences between each other. The testing statistical significance results are as follow: EPA and Fusion: the t statistic = −37.021, *p*-value = 0.0000; EPA and AirBox: the t

(a) The first event: 2016-10-15 06:00 to 2016-10-15 14:00

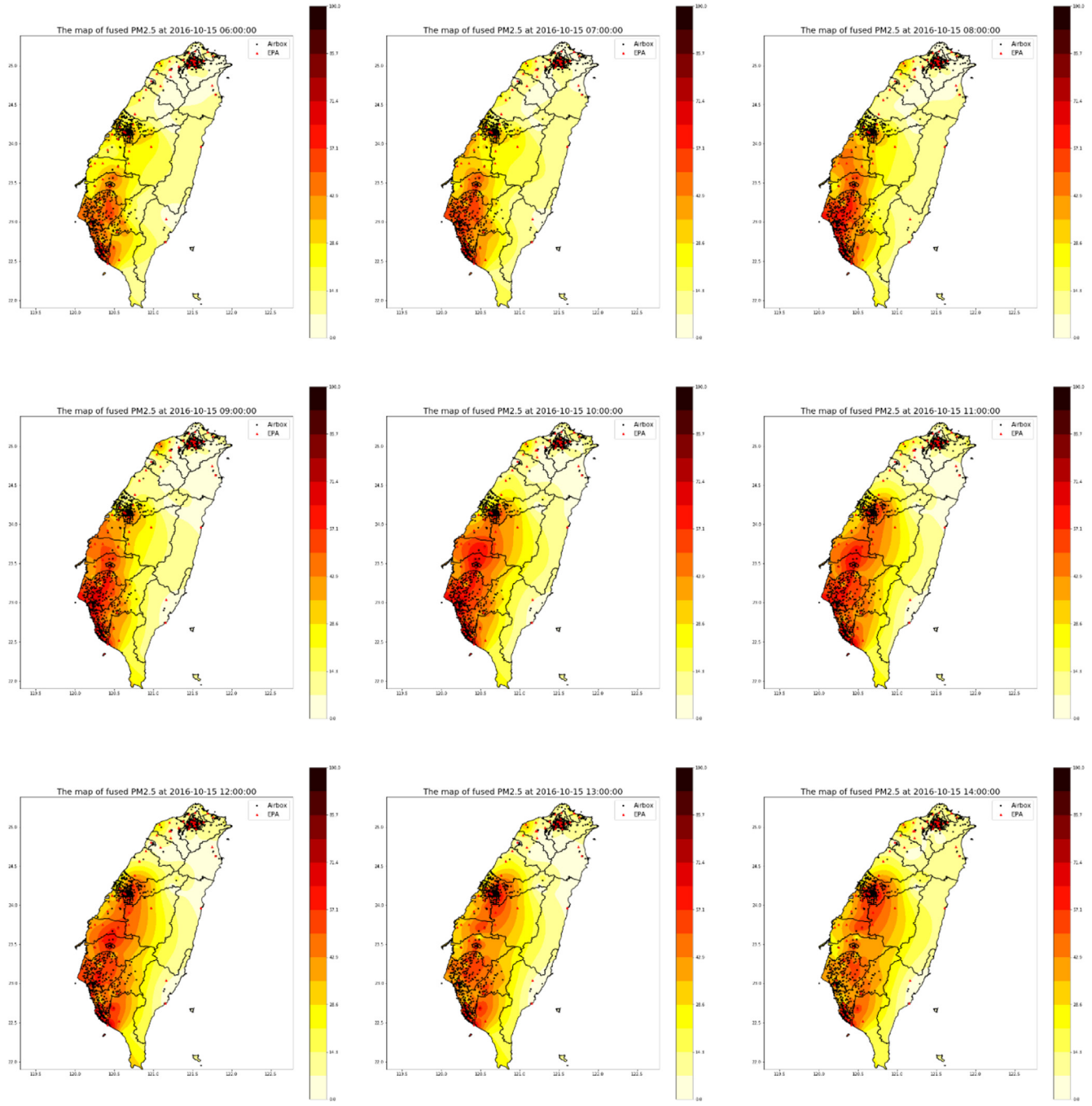

Fig. 13. The high-resolution spatial estimation map of $PM_{2.5}$ concentration (μg/m^3) after using multi-sensor space-time data fusion framework at two different events: (a) 2016-10-15 06:00 to 2016-10-15 14:00 (b) 2016-10-27 06:00 to 2016-10-27 14:00.

statistic = −194.552, *p*-value = 0.0000; AirBox and Fusion: the t statistic = 159.135, *p*-value = 0.0000.

For the spatial viewpoint, we first exam the spatial estimation map of $PM_{2.5}$ concentration before and after using multi-sensor space-time data fusion framework at four different time-slices: (a) 2016-10-16 21:00 (b) 2016-10-17 17:00 (c) 2016-10-18 06:00 (d) 2016-10-20 03:00. All the time-slices maps are shown by gif animation displayed in the Appendix of the online version. As the spatial distribution of $PM_{2.5}$ shown in Fig. 11, without the fusion process, only Taiwan EPA data used and only AirBox data were used for generating the spatial estimation map, respectively, and compare the results of the fusion in the third column. They all present similar large-scale spatial patterns. This is because the targets of them are all $PM_{2.5}$ concentrations, and the air has the characteristics of circulation and diffusion. The overall spatial distribution should be similar. However, if we zoom in, only EPA data shows low spatial resolution, but revealing large-scale and gradually changing spatial characteristics. On the other hand, the estimation result of only AirBox data has very high spatial variability, although it has a much higher spatial resolution than EPA data. AirBox data can reveal some local air quality information. However, it also brings higher uncertainty due to some unclear monitoring process and data quality. The spatial variations in AirBox $PM_{2.5}$ concentration often change suddenly,

(b) The second event: 2016-10-27 06:00 to 2016-10-27 14:00

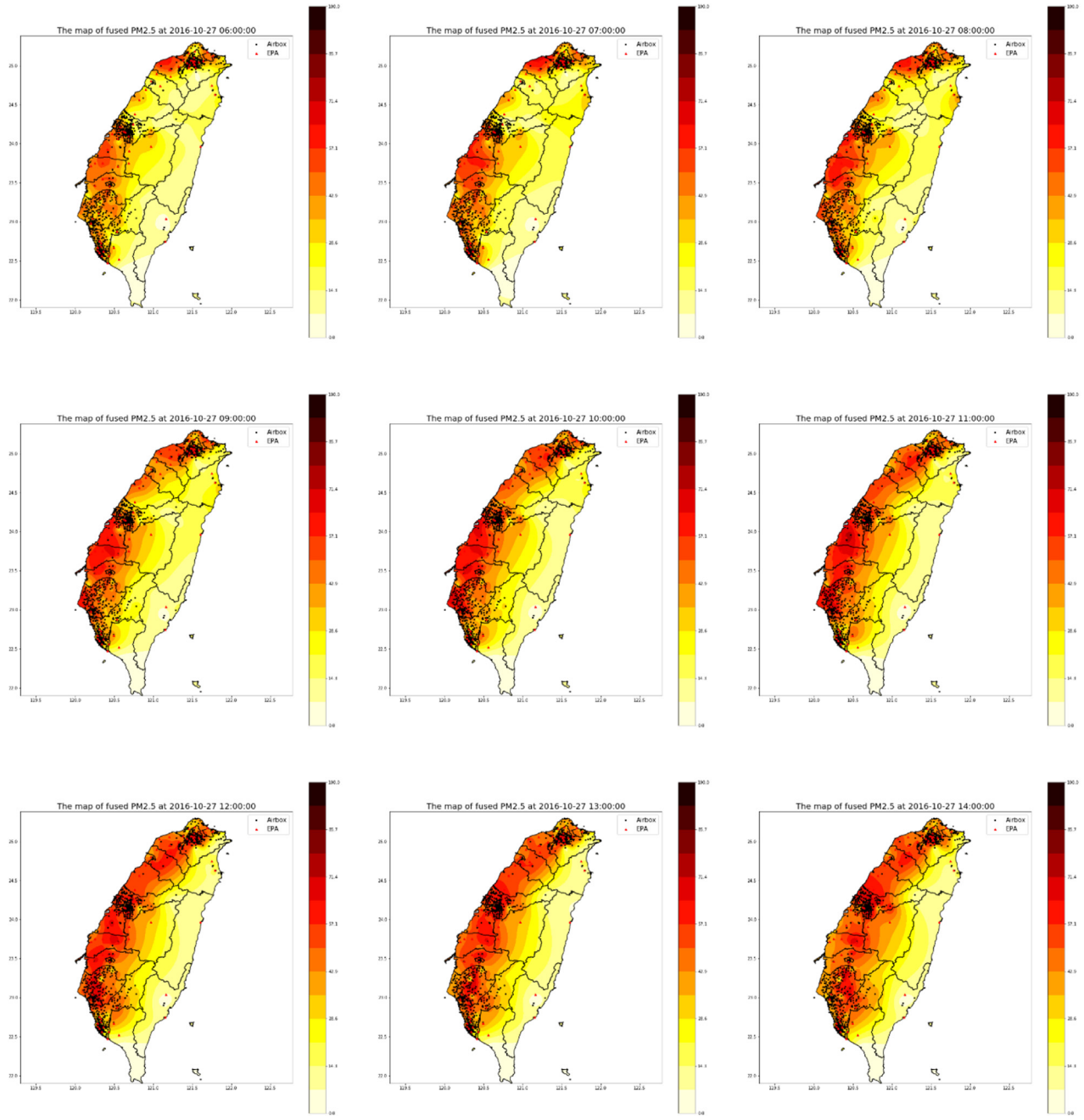

Fig. 13. (*continued*)

even though they are very close neighbors. This may be less consistent with the natural physical properties of air circulation and diffusion.

For instance, in Fig. 11a and b, the overall spatial distributions of EPA are similar to AirBox, so the spatial distribution of the fusion results are not significantly different from the original maps. It is a relatively simple example. However, in Fig. 11c and d, the spatial distribution of AirBox have high variability and also have different patterns with the distribution of EPA. These areas of the AirBox during these periods may be affected by local air pollution events in very small areas (such as someone smoking near the AirBox, roasting, firing, etc.) or the sensor abnormality of the AirBox itself. Without the data fusion analysis, it is very unreasonable only through the spatial distribution map of the original AirBox data. For example, the high concentration value of PM2.5 in the mountainous area in Fig. 11d is very high in the mountains, which is very unreasonable. Therefore, through this research framework, we could find a reasonable compromise between EPA and AirBox, keeping the large-scale spatial pattern and important local information, such as the abnormal points in Fig. 11c and d are all have been corrected after the fusion process. In fact, Fig. 11a and b also have similar corrections, but they are not so obvious. From Fig. 11, we can find that it does improve the estimation resolution of the space from 74 to 1176 spatial data points after combining AirBox and Taiwan EPA information, producing a more reasonable spatial estimation for the unknown $PM_{2.5}$ concentration area without monitoring stations. It not only retains the spatial characteristics of the large-scale view but also shows the spatial variation of local air quality information.

In order to verify whether there are significant spatial pattern differences in Fig. 11 between EPA, AirBox, and Fusion, we use the

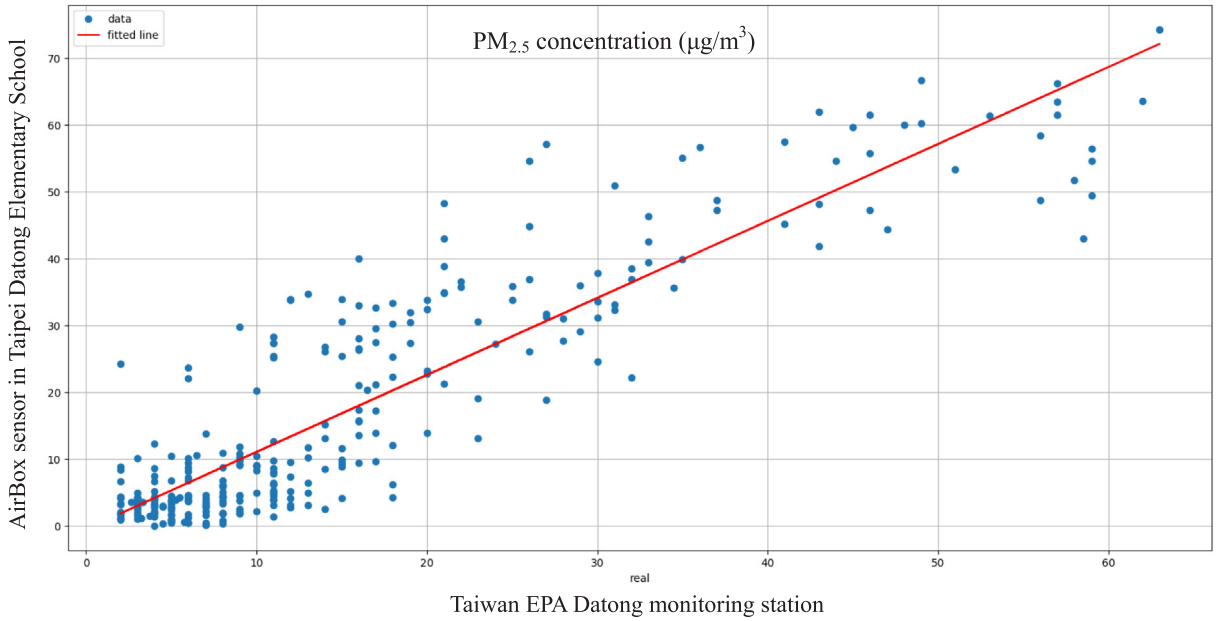

Fig. 14. The scatter plot of $PM_{2.5}$ concentration between Taiwan EPA Datong monitoring station (x: 25.0632, y: 121.513311) and 300 m away AirBox sensor in Taipei Datong Elementary School (Sensor ID: 28C2DDDD4568, x: 25.0653, y: 121.516). The linear regression $r^2 = 0.89$, p-value = 5.98e-106, std. err = 0.036.

statistical test method proposed by (Levine et al., 2009) to compare the spatial distribution maps. We performed pixel-based statistical comparisons between the results in Fig. 11 and calculated the mean difference in pixel scores, standard deviation, *t*-value, *p*-value for the *t*-test, percent difference, and Kappa value. The results are shown in Table 3. The results show that in the maps in Fig. 11, almost all before fusion and after fusion maps show significant differences in the spatial pattern, except for the result between Airbox only and fusion at 2016-10-17 17:00. The distributions of pixel difference scores for Fig. 11 are shown in Fig. 12. The histograms show the distribution that many pixels are not with values of or close to zero (difference = 0). It also indicates that there are significant differences between before fusion and after fusion. Compare with other time slices, the distribution of pixel difference is more concentrated at zero at 2016-10-17 17:00, which is also no significant difference from the statistical analysis result. Since the main objective of this study is not to find the results of all the fusion results that are significantly different from the original data, but to combine the advantages of the two sources to find the best fusion in the time and space, some of the time slices have no significant difference is quite a normal phenomenon.

Fig. 13 shows the representative spatial estimation maps of $PM_{2.5}$ concentration by using multi-sensor space-time data fusion framework at several time-slices based on different events: (a) 2016-10-15 06:00 to 2016-10-15 14:00 (b) 2016-10-27 06:00 to 2016-10-27 14:00. Similarly, due to the article space limitations, all the other time-slices maps are also shown by gif animation displayed in the Appendix of the online version. We can find that during the first event, there was a higher concentration of $PM_{2.5}$ in central and south-west of Taiwan, which is a common distribution of air pollution in autumn and winter in Taiwan. High $PM_{2.5}$ concentration starts from the south-west of Taiwan at 2016-10-15 06:00 and gradually spread to central Taiwan. The second event begins with local pollution in small areas in the north and central of Taiwan, and finally spread throughout western Taiwan. Due to the improved spatial resolution, we can understand the boundaries of concentration changes in more detail, the large-scale characteristics, the local concentration changes in the urban and suburban areas, and more detailed pollution transmission and diffusion with time. Fig. 14 shows the scatter plot of $PM_{2.5}$ concentration between Taiwan EPA Datong monitoring station (x: 25.0632, y: 121.513311) and 300 m away AirBox sensor in Taipei Datong Elementary School (Sensor ID: 28C2DDDD4568, x: 25.0653, y: 121.516). The linear regression $r^2 = 0.89$, p-value = 5.98e-106, std. err = 0.036, showing a good validation result with a linear relationship.

5. Discussions

Data fusion technology was first used in the military, mainly for the detection and identification of military targets. The definition of data fusion currently accepted by most researchers was proposed by Joint Directors of Laboratories in 1991: “Data fusion is a process dealing with the association, correlation, and combination of data and information from single and multiple sources to achieve refined position and identity estimates, and complete and timely assessments of situations and threats as well as their significance.“(White, 1991) In recent years, with the advancement of sensor technology, the number of sensors is increasing, and the technical performance of various sensors is different. The observation data obtained by a single sensor often cannot meet the requirements in terms of accuracy and resolution, while the multi-sensor system mainly has the advantages including enhances the system's survivability, extend space coverage, extend time coverage, improve credibility, improve sensing performance, reduce information ambiguity, and improve spatial-temporal resolution. Therefore, the use of multi-sensors for complementary data fusion techniques in the field of atmospheric science, sonar, remote sensing, and other fields are also attracting more attention in recent years. The technology of fusion the multi-source generated data by two or more sensors to obtain the best estimate has become a hot topic in various fields.

Commonly used methods for data fusion including weighted averaging, Wavelet transform, Kalman Filter, Bayesian estimation, and Neural Networks. Among them, the data fusion method based on the weighted average method is the earliest application, which can be applied to static or dynamic raw data (Yan et al., 2011). With people's attention to air quality, data fusion technology is also used in areas such as estimating air pollutant concentration. (Garcia et al., 2010) developed a comparison of statistical techniques for combining modeled data and observed concentrations to create high-resolution ozone air quality surfaces, using the weighted average method combined with Kriging estimation results and CMAQ model (The Community Multiscale Air Quality Modeling System) to obtain the best ozone concentration estimation. Cross-validation results indicate that the simpler techniques

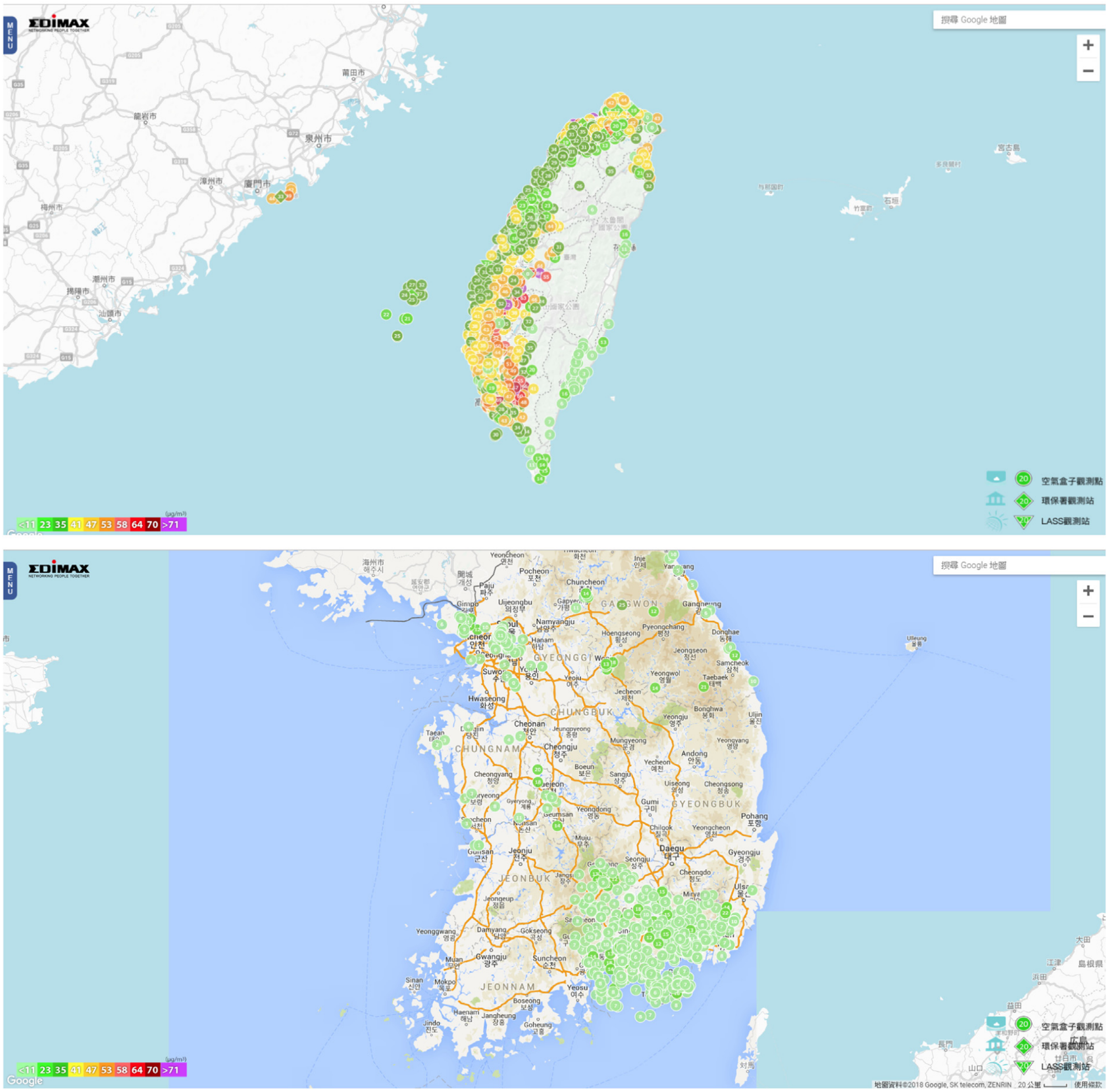

Fig. 15. The AirBox website of EDIMAX Technology Co., Ltd.: (source: https://airbox.edimaxcloud.com/). Upper: Taiwan; Lower: Korea. Screenshot on 2018/10/10.

perform as well as or better than the more complex techniques. Furthermore, some other researchers have also used Kalman Filter technique and CMAQ model data fusion methods to improve the spatial-temporal resolution and forecasting of air pollution (Alexis et al., 2008; Djalalova et al., 2015). However, these methods mostly integrate the numerical model estimation results with the observation data, rather than two different types of observation data from different monitoring sensors. In the field of information retrieval and sonar signal processing, some scholars have proposed an Optimal Linear Data Fusion method based on the weighted average method and the least squares method. This method combines the independent observations of different sensors to obtain the best state estimation and suitable weights for all the component retrieval systems by observing the error (Li, 2012; Wu et al., 2011).

Therefore, in this study, a novel application based on the Optimal Linear Data Fusion method was applied combing with the Kriging estimation technique to the data fusion between different types of $PM_{2.5}$ sensors including Taiwan EPA and AirBox. The Optimal Linear Data Fusion depends on the variance of the observed data. The larger the variance, the smaller the weight. Therefore, the choice of the variance is the point of the method, and it is also difficult. If the selection of the variance is not good, it is easy to cause the fusion result to be biased toward a certain type of sensor. In this study, the variance was calculated from the $PM_{2.5}$ concentration values of the two sensors per day, i.e., daily variance. The weight does not only depend on the distance from EPA stations, but also on the variance of EPA and AirBox data. For example, in Fig. 11d, though the distance between the center of Taiwan and EPA stations is large, the Airbox data at that moment was apparently an outlier due to some measuring problems, the variance was higher. Therefore, in the fusion process, this is the advantage of this framework to find the best compromise between EPA data and AirBox data, and fix the spatial-temporal outlier problems automatically. The weight is dynamic changes and will update by time. In this study, the monitoring results of the two sources have similar large-scale patterns but different in many local details and spatial distribution, so it is very suitable for fusion. In particular, the AirBox can measure $PM_{2.5}$ in many areas without any national stations, although the accuracy is not enough, the framework of this study provides AirBox as a good source of soft information to combine with EPA hard information.

With the rapid development of low-cost sensors and IoT technology,

the monitoring of air quality enters a new era. In recent years, AirBoxes have been rapidly deployed around the world, e.g., Korea, China, USA, etc. The basic framework is similar to this case shown in the study. The methodology and framework of this study can be applied around the world due to the air quality monitoring framework are similar. An AirBox costs only about $105 USD, easy to buy and easy for installation. Through the spontaneous participation of the people and the open data framework, it can be quickly and extensively distributed and become a part of a smart city framework. Currently, most of the low-cost $PM_{2.5}$ AirBox micro-sensors are only used for the display of real-time air quality (The AirBox website of EDIMAX Technology Co., Ltd.: https://airbox.edimaxcloud.com/) (Fig. 15). In addition, compared with the previous studies, only the information of less than 77 Taiwan EPA station is used (Chen et al., 2012; Chu et al., 2012; Wu et al., 2018). The spatial resolution of $PM_{2.5}$ estimation is too low. It existing a certain degree of uncertainty while performing the spatial-temporal estimation. However, in this study, combined with high-cost monitoring data and the fusion method proposed by this research, the value of low-cost observation data are improved, time and space resolution are also improved, and more detailed $PM_{2.5}$ concentration values in unknown regions can be estimated and provided for subsequent health risks analysis. Furthermore, produce the $PM_{2.5}$ concentration map and build the prediction model of the village level's air quality will become more accurate.

6. Conclusions

Public participation in monitoring air quality will be the future trend, especially with the development of the smart city. This study making low-cost micro-sensor AirBox information becomes more valuable. Based on this framework, the new information after the estimation has retained the information of the Taiwan EPA national monitoring station and the micro-sensor AirBox to the greatest extent. It not only corrects the high variability of the time series but also combines the information of the Taiwan EPA station and low-cost micro-sensor AirBox, providing better spatial-temporal resolution for $PM_{2.5}$ concentration distribution across Taiwan. This study proves that the low-cost micro-sensor Airbox does provide useful soft information for spatial-temporal estimation, and the fusion data can be used for further prediction, analysis, health risk assessment, real-time online display, and environmental management, such as air pollution source localization, health risk assessment, and micro-scale air pollution analysis.

AirBox has the advantages of small, low cost, portable, and easy to install, while the advantages of Taiwan EPA national stations are relatively sophisticated measure method and high accuracy with great data quality. By the framework proposed in this study, we maintain the advantages of $PM_{2.5}$ observation data from both Taiwan EPA and AirBox. Moreover, the Ordinary Kriging method was modified for spatial-temporal estimation. This is the first study to combine the low-cost micro-sensor Airbox data and EPA national monitoring stations data for air quality estimation, providing good extensions for future applications.

Acknowledgments

We are grateful to the support of the projects of the Taiwan Environmental Protection Administration (Taiwan EPA) and the Ministry of Science and Technology (Taiwan). Projects No. EPA-106-L103-02-A022, EPA-106-L102-02-A142, and MOST 107-2119-M-008-006. We are also thankful for the cooperation with the following companies: Supergeo Technologies Inc. (https://www.supergeotek.com/), Environmental Management Consultants Technologies (EMCT) Inc. (http://www.emct.com.tw/), Chunghwa Telecom Co., Ltd. (https://www.cht.com.tw/), and EDIMAX Technology Co., Ltd. (https://www.edimax.com/edimax/global/). We also thank Prof. Hwa-Lung Yu of National Taiwan University, and Dr. Ling-Jyh Chen of the Institution of Information Science, Academia Sinica, as well as LASS (Location Aware Sensing System) community establishing and promoting the $PM_{2.5}$ micro-sensor monitoring network. We also thank the National Science and Technology Center for Disaster Reduction (NCDR), and the Python programing language and its data analysis modules as a powerful tool in our data analysis.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2019.105305.

References

Alexis, Z., Li, C., Kotamarthi, V.R., 2008. EAKF-CMAQ: Introduction and evaluation of a data assimilation for CMAQ based on the ensemble adjustment Kalman filter. J. Geophys. Res. Atmos. 113 (D9). https://doi.org/10.1029/2007JD009267.

Anselin, L., Le Gallo, J., 2006. Interpolation of air quality measures in hedonic house price models: spatial aspects. Spatial Econ. Anal. 1 (1), 31–52. https://doi.org/10.1080/17421770600661337.

Atkinson, R.W., Kang, S., Anderson, H.R., Mills, I.C., Walton, H.A., 2014. Epidemiological time series studies of PM2.5 and daily mortality and hospital admissions: a systematic review and meta-analysis. Thorax. https://doi.org/10.1136/thoraxjnl-2013-204492.

Atzori, L., Iera, A., Morabito, G., 2010. The internet of things: a survey. Comp. Networks 54 (15), 2787–2805. https://doi.org/10.1016/j.comnet.2010.05.010.

Bayraktar, H., Turalioglu, F.S., 2005. A Kriging-based approach for locating a sampling site—in the assessment of air quality. Stochastic Environ. Res. Risk Assess. 19 (4), 301–305. https://doi.org/10.1007/s00477-005-0234-8.

Chen, C.-C., Wu, C.-F., Yu, H.-L., Chan, C.-C., Cheng, T.-J., 2012. Spatiotemporal modeling with temporal-invariant variogram subgroups to estimate fine particulate matter PM2.5 concentrations. Atmos. Environ. 54, 1–8. https://doi.org/10.1016/j.atmosenv.2012.02.015.

Chen, L.J., Ho, Y.H., Lee, H.C., Wu, H.C., Liu, H.M., Hsieh, H.H., Lung, S.C.C., 2017. An open framework for participatory PM2.5 monitoring in smart cities. IEEE Access 5, 14441–14454.

Chen, M., Mao, S., Liu, Y., 2014. Big data: a survey. Mobile Networks Appl. 19 (2), 171–209. https://doi.org/10.1007/s11036-013-0489-0.

Chu, H.-J., Yu, H.-L., Kuo, Y.-M., 2012. Identifying spatial mixture distributions of PM2.5 and PM10 in Taiwan during and after a dust storm. Atmos. Environ. 54, 728–737. https://doi.org/10.1016/j.atmosenv.2012.01.022.

Djalalova, I., Delle Monache, L., Wilczak, J., 2015. PM2.5 analog forecast and Kalman filter post-processing for the Community Multiscale Air Quality (CMAQ) model. Atmos. Environ. 108, 76–87. https://doi.org/10.1016/j.atmosenv.2015.02.021.

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., Gnanzou, D., 2015. How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. Int. J. Prod. Econ. 165, 234–246. https://doi.org/10.1016/j.ijpe.2014.12.031.

Garcia, V.C., Foley, K.M., Gego, E., Holland, D.M., Rao, S.T., 2010. A comparison of statistical techniques for combining modeled and observed concentrations to create high-resolution ozone air quality surfaces. J. Air Waste Manag. Assoc. 60 (5), 586–595.

Jacob, D.J., Winner, D.A., 2009. Effect of climate change on air quality. Atmos. Environ. 43 (1), 51–63. https://doi.org/10.1016/j.atmosenv.2008.09.051.

Kan, H.D., Chen, R.J., Tong, S.L., 2012. Ambient air pollution, climate change, and population health in China. Environ. Int. 42, 10–19.

Krige, D.G., 1952. A statistical approach to some basic mine valuation problems on the Witwatersrand. J. South Afr. Inst. Min. Metall. 52 (9), 201–203.

Labrinidis, A., Jagadish, H.V., 2012. Challenges and opportunities with big data. Proc. VLDB Endow. 5 (12), 2032–2033. https://doi.org/10.14778/2367502.2367572.

Leiva, G.M.A., Santibañez, D.A., Ibarra, E.S., Matus, C.P., Seguel, R., 2013. A five-year study of particulate matter (PM2.5) and cerebrovascular diseases. Environ. Pollut. 181, 1–6. https://doi.org/10.1016/j.envpol.2013.05.057.

Lelieveld, J., Evans, J.S., Fnais, M., Giannadaki, D., Pozzer, A., 2015. The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature 525 (7569), 367–371. https://doi.org/10.1038/nature15371.

Levine, R.S., Yorita, K.L., Walsh, M.C., Reynolds, M.G., 2009. A method for statistically comparing spatial distribution maps. Int. J. Health Geogr. 8, 7. https://doi.org/10.1186/1476-072X-8-7.

Li, J., Heap, A.D., 2014. Spatial interpolation methods applied in the environmental sciences: a review. Environ. Modell. Software 53, 173–189. https://doi.org/10.1016/j.envsoft.2013.12.008.

Li, Q. (2012). Digital Sonar Design in Underwater Acoustics: Principles and Applications. Hangzhou; Berlin; New York: Zhejiang University Press; Springer.

Madakam, S., Ramaswamy, R., Tripathi, S., 2015. Internet of things (IoT): a literature review. J. Comput. Commun. 3 (05), 164.

Matheron, G., 1963. Principles of geostatistics. Econ. Geol. 58 (8), 1246–1266. https://doi.org/10.2113/gsecongeo.58.8.1246.

Mayer-Schönberger, V., Cukier, K., 2013. Big Data: A Revolution that will Transform how we Live, Work, and Think. Houghton Mifflin Harcourt.

McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D., Barton, D., 2012. Big data. Manage. Revol. Harvard Bus Rev. 90 (10), 61–67.

Niu, Z., Chen, J., Xu, L., Yin, L., Zhang, F., 2013. Application of the environmental

internet of things on monitoring PM2.5 at a coastal site in the urbanizing region of southeast China. Int. J. Sust. Develop. World Ecol. 20 (3), 231–237. https://doi.org/10.1080/13504509.2013.782904.

Sagiroglu, S., Sinanc, D. (2013, 20-24 May 2013). Big data: A review. Paper presented at the Collaboration Technologies and Systems (CTS), 2013 International Conference on.

Tecer, L.H., Alagha, O., Karaca, F., Tuncel, G., Eldes, N., 2008. Particulate matter (PM2.5, PM10-2.5, and PM10) and Children's Hospital Admissions for asthma and respiratory diseases: a bidirectional case-crossover study. J. Toxicol. Environ. Health, Part A 71 (8), 512–520. https://doi.org/10.1080/15287390801907459.

White, F. E. (1991). Data fusion lexicon. Retrieved from.

Wong, D.W., Yuan, L., Perlin, S.A., 2004. Comparison of spatial interpolation methods for the estimation of air quality data. J. Exposure Anal. Environ. Epidemiol. 14, 404. https://doi.org/10.1038/sj.jea.7500338.

Wu, C.D., Zeng, Y.T., Lung, S.C., 2018. A hybrid kriging/land-use regression model to assess PM2.5 spatial-temporal variability. Sci. Total Environ. 645, 1456–1464. https://doi.org/10.1016/j.scitotenv.2018.07.073.

Wu, S., Bi, Y., Zeng, X. (2011). The Linear Combination Data Fusion Method in Information Retrieval, Berlin, Heidelberg.

Wu, Y.-C., Lin, Y.-C., Yu, H.-L., Chen, J.-H., Chen, T.-F., Sun, Y., Chen, Y.-C., 2015. Association between air pollutants and dementia risk in the elderly. Alzheimer's Dementia: Diagn. Assess. Disease Monit. 1 (2), 220–228. https://doi.org/10.1016/j.dadm.2014.11.015.

Yan, Z.-Z., Yan, X.-P., Xie, L., Wang, Z. (2011). The Research of Weighted-Average Fusion Method in Inland Traffic Flow Detection, Berlin, Heidelberg.

Yu, H.-L., Chien, L.-C., Yang, C.-H., 2012. Asian dust storm elevates children's respiratory health risks: a spatiotemporal analysis of children's clinic visits across Taipei (Taiwan). PLoS One 7 (7), e41317. https://doi.org/10.1371/journal.pone.0041317.

Yu, H.-L., Lin, Y.-C., Sivakumar, B., Kuo, Y.-M., 2013. A study of the temporal dynamics of ambient particulate matter using stochastic and chaotic techniques. Atmos. Environ. 69, 37–45. https://doi.org/10.1016/j.atmosenv.2012.10.067.

Yu, H.L., Lin, Y.C., Kuo, Y.M., 2015. A time series analysis of multiple ambient pollutants to investigate the underlying air pollution dynamics and interactions. Chemosphere 134, 571–580. https://doi.org/10.1016/j.chemosphere.2014.12.007.

Zanella, A., Bui, N., Castellani, A., Vangelista, L., Zorzi, M., 2014. Internet of things for smart cities. IEEE Internet Things J. 1 (1), 22–32. https://doi.org/10.1109/JIOT.2014.2306328.

Zheng, Y., Liu, F., Hsieh, H.-P. (2013). U-Air: when urban air quality inference meets big data. Paper presented at the Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Chicago, Illinois, USA.