ADF: An Anomaly Detection Framework for Large-Scale PM2.5 Sensing Systems

Ling-Jyh Chen, *Senior Member, IEEE*, Yao-Hua Ho, Hsin-Hung Hsieh, Shih-Ting Huang, Hu-Cheng Lee, and Sachit Mahajan, *Graduate Student Member, IEEE*

***Abstract*—As the population density continues to grow in the urban settings, air quality is degrading and becoming a serious issue. Air pollution, especially fine particulate matter (PM2.5), has raised a series of concerns for public health. As a result, a number of large-scale, low cost PM2.5 monitoring systems have been deployed in several international smart city projects. One of the major challenges for such environmental sensing systems is ensuring the data quality. In this paper, we propose an anomaly detection framework (ADF) for large-scale, real-world environmental sensing systems. The framework is composed of four modules: 1) time-sliced anomaly detection (TSAD), which detects spatial, temporal, and spatio-temporal anomalies in the real-time sensor measurement data stream; 2) real-time emission detection, which detects potential regional emission sources; 3) device ranking, which provides a ranking for each sensing device; and 4) malfunction detection, which identifies malfunctioning devices. Using real world measurement data from the AirBox project, we demonstrate that the proposed framework can effectively identify outliers in the raw measurement data as well as infer anomalous events that are perceivable by the general public and government authorities. Because of its simple design, ADF is highly extensible to other advanced applications, and it can be exploited to support various large-scale environmental sensing systems.**

***Index Terms*—Anomaly detection, data analysis, PM2.5, smart city.**

I. Introduction

A SMART city is a developed city that leverages the advances in information and communication technology (ICT) to improve the city's quality of life, promote sustainable development, and raise the levels of public "happiness" [1]. The concept requires the installation of ICT infrastructures throughout the city, as well as embedding the concept of "smartness" across all city functions. Smart city systems have a wide range of applications, including economy, mobility, environment, people, living, and governance [2].

Manuscript received December 31, 2016; revised August 30, 2017 and October 12, 2017; accepted October 14, 2017. Date of publication October 24, 2017; date of current version April 10, 2018. This work was supported in part by the Ministry of Science and Technology of Taiwan under Grant 105-2221-E-001-016-MY3, Grant 106-3011-F-001-001, and Grant 106-3114-E-001-004, and in part by Academia Sinica under Grant AS-104-SS-A02. *(Corresponding author: Ling-Jyh Chen.)*

L.-J. Chen, H.-H. Hsieh, S.-T. Huang, H.-C. Lee, and S. Mahajan are with the Institute of Information Science, Academia Sinica, Taipei 11529, Taiwan (e-mail: ccljj@gmail.com).

Y.-H. Ho is with the Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei 116, Taiwan.

Digital Object Identifier 10.1109/JIOT.2017.2766085

Among all smart city systems, increasing attention is being paid to the large-scale deployment of applications for monitoring finer-grained air pollution [3]. The reason is that air quality is a global issue and it is being degraded by economic activity, rapid urbanization, and increased energy consumption. Air pollution raises a number of concerns ranging from public health to the social economy [4], [5]. Among all pollutants, fine particulate matter (PM2.5) comprises particles that are less than 2.5 μm in diameter. It has been shown that such particles are harmful to human health as they can penetrate the alveoli (the gas exchange regions of the lungs) and even pass through the lungs to affect other organs. Recent studies have shown that PM2.5 pollution is directly related to various serious health disorders, such as asthma, cardiovascular disease, respiratory diseases, lung cancer, and premature death [6], [7].

A number of outdoor PM2.5 monitoring projects have been launched worldwide [8]–[10]. Conventional approaches usually rely on professional air quality monitoring stations that are deployed at strategic locations and generally operated by national, state, or local environmental protection agencies (EPAs or their equivalent). However, the monitoring stations are extremely large and expensive to operate, and they cannot be deployed in a high density. As a result, sophisticated air pollution dispersion models must be used to estimate PM2.5 concentrations in the areas between different stations [11]. However, the accuracy of the estimations depends on wind conditions, the terrain, and the distance to the nearest stations. Moreover, the measurement results of conventional stationary systems are only effective in representing well-mixed atmospheric pollution. They cannot indicate the air quality in our living surroundings [12], [13].

With advances in sensing and computing technology, several recent works have applied low-cost sensors in microscale PM2.5 sensing [14]–[20]. Meanwhile, some international smart cities have deployed large-scale, low-cost particle sensors for real-time air quality monitoring (e.g., Chicago [21], Taipei [22], and Zürich [23]). In addition, there are several commercial products and nonprofit communities that encourage people to monitor PM2.5 concentrations in their surroundings and send the measurement results to the cloud (e.g., AirCasting [24], Clarity [25], Laser Egg [26], location aware sensing system (LASS) [27], and uHoo [28]).

Data quality is a major challenge in the deployment of large-scale environmental sensing systems for several reasons [29]. First, because monitoring devices are deployed in uncontrolled

environments, unexpected measurement results are normal and have to be carefully managed. Second, it is difficult to use a single model for all devices that are deployed in non-homogeneous environments. Third, the interpretation of the measurement results must be based on domain knowledge, rather than pure computational models. Although a number of approaches address the anomaly detection issue in large-scale sensing systems [30]–[36], the following two aspects have not been investigated by existing studies.

1) *Effectiveness of Anomaly Detection:* Existing anomaly detection techniques rely on long periods of anomaly free observations, *a priori* data profiles, or a large number of labeled observations. However, these prerequisites are infeasible for real-world, large-scale systems because data collection, analysis, and modeling require a great deal of time. Moreover, existing detection techniques cannot deal with unexpected measurement results, which are inevitable in real-world systems and must be handled properly.
2) *Inference of Anomalous Events:* Most existing techniques focus on detecting outliers in the raw measurement results rather than identifying anomalous events, which is essential for real-world environmental sensing systems. Although some approaches [36] can be extended to infer anomalous events, however, they are delayed indicators due to their sophisticated models. They also fail to identify events not shown in the training dataset.

To address the above issues, we propose an anomaly detection framework (ADF) for large-scale, real-world environmental sensing systems. The ADF can effectively identify outliers in the raw measurement data, as well as infer anomalous events that are perceivable by the general public and government authorities. The framework is based on a core module called time-sliced anomaly detection (TSAD), which identifies spatial, temporal, and spatio-temporal anomalous instances in each segment of time-sliced measurement data derived by a large-scale sensing system. Next, the sequences of TSAD results are fed into the real-time emission detection (RED), device ranking (DR), and malfunction detection (MD) modules to, respectively, infer potential regional emission sources, rank the data consistency of each device, and assess the properties of each device.

Using a real-world dataset from a large-scale PM2.5 monitoring system called AirBox [37], we conduct a comprehensive analysis of the dataset, and fine tune the parameters of the TSAD module to ensure its effectiveness. Meanwhile, the RED, DR, and MD modules are implemented as online services. The real-time results are publicly available in an open data format. Our research outcomes have been used by third parties in their data visualization services. Furthermore, the analysis results are sent to our city and national EPA authorities for on-demand responses and consideration in formulating environmental policy.

The contributions of this paper are as follows.

1) We propose an ADF for large-scale environmental sensing systems. The proposed framework is simple, effective, and capable of inferring anomalous events.
2) We discuss the decision criteria of the ADF parameters and conducted a comprehensive analysis based on a real-world dataset to identify the optimal parameter settings.
3) We implement the proposed framework incorporated with a large-scale PM2.5 monitoring system, called AirBox. As a result the system becomes a regular service providing open data service for further applications.
4) We demonstrate that the proposed system is highly extensible and can be used in large-scale environmental sensing systems for real-time anomaly detection.

The remainder of this paper is organized as follows. Section II provides a review of related work on anomaly detection for networked sensing systems. In Section III, we introduce the AirBox project, a large-scale PM2.5 monitoring system that is the platform for this paper. In Section IV, we consider the proposed ADF and discuss its components in detail; and in Section V, we describe the implementation of ADF and report the analysis outcomes of ADF in the AirBox PM2.5 monitoring system. Section VI contains some concluding remarks.

II. Related Work

Data quality is the key to the success of a large-scale networked sensing system. However, the issue is extremely challenging, especially when the sensing devices are deployed in uncontrolled environments [29]. *Anomaly detection* is an essential factor in the implementation of such mechanisms. There have been several studies of anomaly detection in large systems. They can be classified into three categories based on their fundamental algorithms.

1) *Statistics-Based Techniques:* Wu *et al.* [36] proposed a spatial mining-based approach to detect outlying and boundary data in sensor networks. Their approach considers the spatial correlation of the readings between a sensor and its neighboring nodes. If the absolute value of the sensor's deviation degree is greater than a preselected threshold, it is regarded as an outlier. However, the selection of the threshold is very subjective and it is made without evaluation using realistic datasets. Subsequently, Paschalidis and Chen [35] developed a statistical framework that considers both spatial and temporal correlations between the sensor readings in anomaly detection. Based on Markov models, the framework relies on long periods of anomaly free observations, but this may not be possible for real-world environmental monitoring applications [38]. Moreover, the detection process is quite long and cannot be performed in a timely manner.
2) *Cluster-Based Techniques:* Cluster-based techniques do not require prior knowledge of the data distribution in the sensor readings; and they can support incremental models, i.e., the system can adapt to new data instances over time [29]. Such techniques detect an anomalous event if: a) the centroid of its closest cluster is a known anomaly event or b) the distance to the closest cluster that is a normal event is greater than the threshold. The drawbacks of these techniques are: a) determining

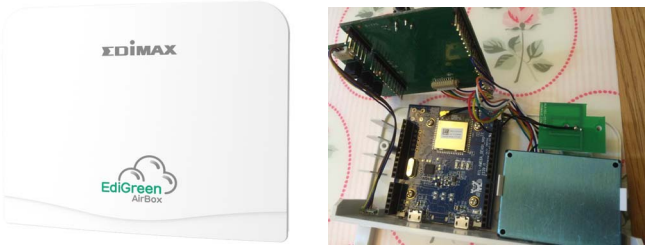
the distance between multivariate measurement data is challenging; b) defining the threshold of a cluster width for normal events is complicated; and c) the measurement data of nonhomogeneous environments cannot be accommodated easily [33]. As a result, these types of techniques are generally used as quick filters for outlier detection (e.g., reducing the number of false alarms in healthcare systems [31] and contextual anomaly detection based on sensor profiles [32]). They are not suitable for real-world anomalous event detection.

3) *Machine Learning-Based Techniques:* Machine learning-based techniques are able to learn anomaly detection from labeled data, and they can be implemented using state-of-the-art machine learning tools directly. The drawbacks of such approaches are: a) a large amount of labeled data is needed as training data in order to cover all possible types of sensor readings in the system and b) it is difficult to achieve accurate anomaly detection unless a sophisticated set of tuning and optimization processes are investigated and implemented. For instance, Murphree [34] proposed a sparse autoencoder that uses neural networks for anomaly detection in large systems. When a large error is observed between the test data and the reconstructed data using the sparse autoencoder, the test data is near the distribution described by the healthy system, and is therefore considered an anomaly. Ayadi *et al.* [30] compared three machine learning-based algorithms, namely, outlier detection by active learning, identifying the density-based local outliers, and feature bagging for outlier detection. Based on the evaluation results, the authors concluded that: a) none of the algorithms is suitable for every case, and the performance of each one depends on the properties of the application datasets and b) anomaly detection in dynamic systems is still challenging.

III. AirBox Deployment

The AirBox project is a pilot deployment of IoT systems for large-scale PM2.5 monitoring in Taiwan. The project was inspired by the work of the LASS community, which engages citizens to participate in the PM2.5 sensing project and enables them to make low-cost PM2.5 sensing devices on their own. It also facilitates PM2.5 monitoring at a finer spatio-temporal granularity and enriches environmental data analysis by making all measurement data freely available to everyone [39]. The AirBox project is more mature than the LASS project because: 1) the number of AirBox devices is much higher than the number of LASS devices; 2) the sensing devices are made by a professional company with industrial product level consistency and reliability; and 3) the devices are deployed in public buildings with a reliable Internet connection and power supply.

The AirBox project started with the deployment of 150 devices in Taipei city on March 22, 2016. Since then the project has been expanded to four other major cities in Taiwan. A total of 990 devices were deployed in elementary schools in the four cities as follows: 242 devices in Kaohsiung on July 26, 2016; 230 devices in Taichung on August 17, 2016; 298 devices in New Taipei on August 19, 2016; and 220 devices in Tainan on August 30, 2016. All of the devices used in the AirBox project were donated to the respective city governments by the manufacturer.



(a) (b)

Fig. 1. AirBox device: (a) overview and (b) internal components.

In addition to the smart city deployment, the manufacturer donated 75 devices to the LASS community to promote AirBox usage in other Taiwanese cities, as well as in other countries. The AirBox device is available online at the retailers' website. So far, more than 300 devices have been purchased, including 120 devices for the Taichung Origin Association's citizen science project. By the end of 2016, there were more than 1500 AirBox devices deployed in 20 cities in Taiwan and 24 countries worldwide.

A. AirBox: Technical Details

Fig. 1 shows the AirBox device in detail. The device is based on the Realtek Ameba RTL8195 board, an open hardware that is Arduino compatible. It is powered by a 5 V micro-USB and contains two environmental sensors—ST HTS221 (i.e., sensing for temperature and relative humidity) and Plantower PMS5003 (G5) (i.e., sensing for fine particle matter). For deployment in outdoor environments, the device is enclosed in a water and UV resistant case.

Connecting to the Internet with a built-in Wi-Fi interface, the AirBox device updates its system time periodically using the network time protocol [40]. Samples are taken every five minutes approximately, and the obtained measurements (temperature, relative humidity, PM2.5, and PM10) are transmitted to the manufacturer's backend database directly. The data is also forwarded to the LASS open data platform for data visualization, analysis, and other applications [39].

The temperature and humidity sensor (ST HTS221) used in the AirBox has proved effective in a wide range of outdoor environmental monitoring applications. The particle sensor (Plantower PMS5003) is a recently developed product based on the laser light scattering principle [41], [42]. It is accurate and robust, especially when the PM2.5 concentration is in the range 30 to100 [43]. Moreover, as the AirBox is based on low-power and open hardware, it can be easily extended to support emerging low-power wide-area networks (LPWANS, e.g., LoRa and SigFox), as well as to support the production of green energy (e.g., solar power).

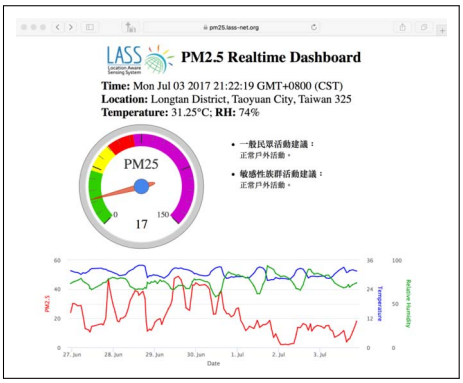


Fig. 2. Dashboard page, which visualizes the real-time monitoring results of a specific device.

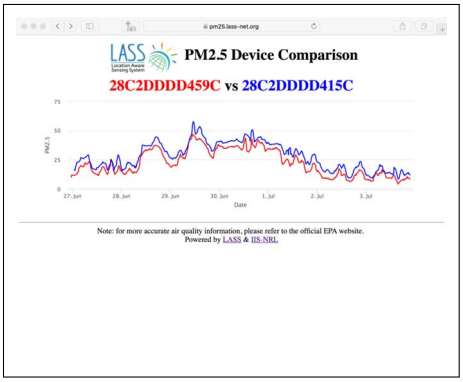


Fig. 3. Device comparison page, which compares the historical PM2.5 monitoring results of two devices.

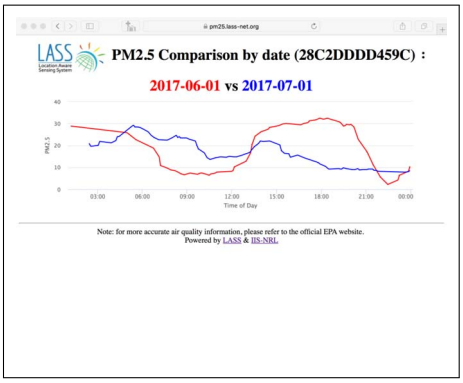


Fig. 4. Device comparison page used to compare the PM2.5 monitoring results of the same device on two different dates.

B. AirBox Applications

Several applications have been developed using data provided by the AirBox deployment. They can be divided into two categories.

1) *Data Visualization:* For AirBox measurement data, two types of data visualization systems have been developed by different parties. The first type is a data-oriented visualization system that focuses on displaying the detailed information of each AirBox device (Fig. 2). It analyses and compares data from different devices on the same day (Fig. 3) and from the same device on different days (Fig. 4). The second type is a geographic information (GIS)-based visualization system that provides comprehensive visual results on a map by mashing up the measurement data and other location-based data. For instance, the *g0v* community combines the measurement data and wind information on the same map to facilitate data analysis and interpretation (Fig. 5). The LASS community applies the Voronoi diagram algorithm to partition the map into regions based on the Euclidean distance between the AirBox devices. Each partitioned region is color-coded based on the PM2.5 measurement data and the danger level advised by the EPA of Taiwan [44] (Fig. 6). The community also utilizes the inverse distance weighting (IDW) algorithm to interpolate the colored gradient between every two AirBox devices nearby. The visual result is a real-time heatmap of PM2.5 distribution in Taiwan (Fig. 7).

2) *Open Data Service:* The measurement data of the AirBox deployment is freely available and can be obtained in two ways: a) by applying for an access token and downloading the data from the AirBox manufacturer's backend database or b) by accessing the open

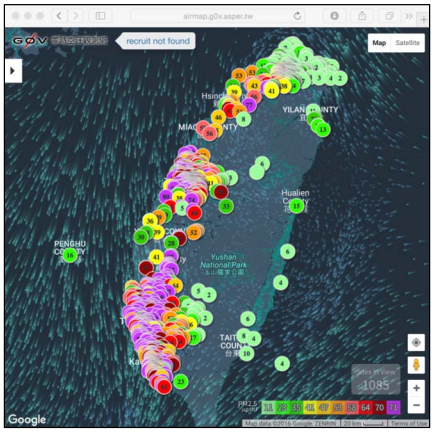


Fig. 5. GIS visualization page provided by the *g0v* community.

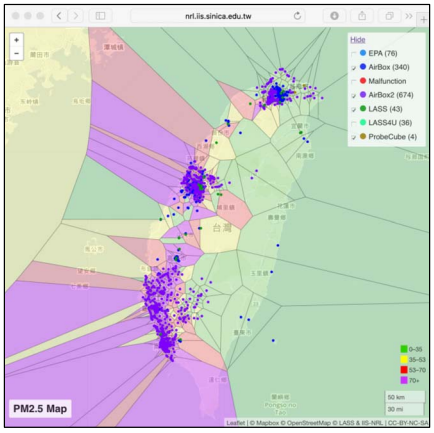


Fig. 6. Voronoi diagram page presents the real-time PM2.5 monitoring results by partitioning the map into color-coded regions.

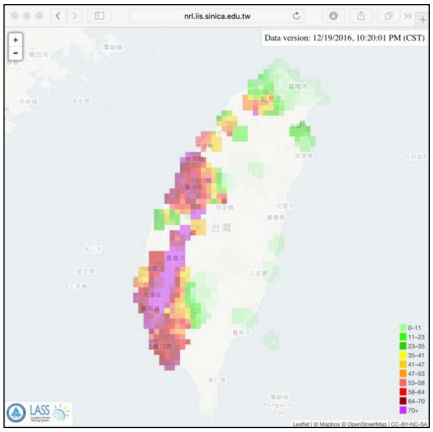


Fig. 7. IDW diagram page that visualizes real-time PM2.5 monitoring results using different colored gradients.

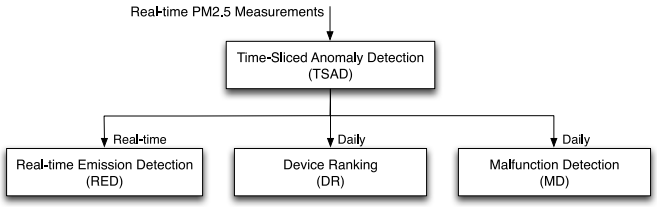


Fig. 8. System architecture of the proposed ADF.

data API provided by the LASS community to download the measurement data directly. The open data API (in the JSON data format) allows people to access the latest measurement data of a specific AirBox device, the last 1000 pieces of measurement data of a particular device, the measurement data of one device on a specific date, and the measurement data of the nearest AirBox device [45].

IV. Anomaly Detection Framework

The system architecture of the proposed ADF is shown in Fig. 8. First, the real time sensor measurement data is fed into the TSAD module to detect any abnormal data. Based on different applications, the data is then processed by the RED module, DR module, and MD module. We discuss each module in the following sections.

A. Time-Sliced Anomaly Detection

The TSAD module addresses the anomaly detection problem by using the time-slicing technique to divide the continuous real-time sensor measurement data into a sequence of fixed-length discrete representations. The latest measurement data provided by each distinct sensor is considered for each time slice. In the kth time slice, D_i^k is the measurement data of the ith sensor.

Three types of anomalies may occur in sensing devices in a real-time sensor measurement data stream, namely, *Spatial Anomalies*, *Temporal Anomalies*, and *Spatio-temporal Anomalies*. They can be assessed by: 1) the similarity of the measurement data in the sensor's neighboring area and 2) the self-consistency in the measurement data of contiguous time slices. We assume that the PM2.5 concentrations are well-mixed and change gradually over time in ordinary atmospheric conditions. The formal definitions of the three types of anomaly are as follows.

Definition 1: A sensing device is a spatial anomaly if its measurement data is far greater (or lower) than the mean of that of its neighbors in the same time-sliced sensor measurement data.

Two sensing devices are regarded as neighbors if they are within d kilometers of each other, where the set of sensing devices N_i are neighbors of the ith device. In the kth time-slice, $Q^{i,k}_{\text{lower}}$ and $Q^{i,k}_{\text{upper}}$ are the lower and upper quartiles, respectively, of the measurement data derived by N_i and the ith device. Using Tukey's test [46], the ith device is deemed a *spatial anomaly* in the kth time-sliced sensor measurement data if

$$D_i^k < Q^{i,k}_{\text{lower}} - r\left(Q^{i,k}_{\text{upper}} - Q^{i,k}_{\text{lower}}\right) \tag{1}$$

or

$$D_i^k > Q^{i,k}_{\text{upper}} + r\left(Q^{i,k}_{\text{upper}} - Q^{i,k}_{\text{lower}}\right) \tag{2}$$

where r is a non-negative constant that is fixed at 1.5 (as suggested in [46]) in this paper.

There are several reasons that a PM2.5 sensing device may yield a *spatial anomaly* assessment. For instance, the sensor may be placed in an indoor location with good air purification [i.e., (2)]; or it may be located near emission sources [i.e., (1)], such as temples (e.g., burning incense) and barbecue restaurants (e.g., cooking by charbroiler or grills). It is also possible that the sensor device becomes a *spatial anomaly* due to incorrect installation or malfunction.

Definition 2: A sensing device is a temporal anomaly if its measurement data increases significantly in contiguous time slices.

More precisely, let δ be the threshold of the reasonable measurement drift in two contiguous time slices. Then, the ith device is deemed a *temporal anomaly* in the kth time-sliced sensor measurement data if

$$\begin{cases} D_i^k - D_i^{k-1} > \delta, \text{ if } D_i^{k-1} \text{ exists} \quad \text{or} \\ D_i^k - D_i^{k-2} > \delta, \text{ if } D_i^{k-1} \text{ is missing.} \end{cases} \tag{3}$$

The possible reasons for *temporal anomaly* assessments by PM2.5 sensing devices are: 1) PM2.5 particles are being emitted near the device; 2) the device is not installed properly; or 3) the device may have malfunctioned.

Definition 3: A sensing device is a spatio-temporal anomaly if it is both a spatial anomaly and temporal anomaly in the same time-sliced sensor measurement data.

The *Spatio-temporal anomaly* combines the properties of the *spatial anomaly* and the *temporal anomaly*. To identify the cause of an anomaly, the behavior of neighboring devices must be taken into account. We discuss this issue in the following subsections.

B. Real-Time Emission Detection

The RED module detects small areas of PM2.5 emissions in real time based on the results obtained from the

TSAD module. The rationale behind RED is that, once an emission occurs, there will be a dramatic increase in the measurements of the closest PM2.5 sensing device. Depending on the volume of the emission and the atmospheric conditions the pollution will gradually spread out for a variable distance.

Using the TSAD module, let $\Omega(D_i^k)$ be an anomaly detection function that determines the anomaly type of D_i^k (i.e., the measurement data of the ith device in the kth time-sliced sensor measurement data). The detected anomaly type is "S" (i.e., *spatial anomaly*), "T" (i.e., *temporal anomaly*), "A" (*spatio-temporal anomaly*), "O" (an ordinary case), or "M" (missing data). The RED module deems that the ith device in the kth time-sliced sensor measurement data is an emission source if both of the following conditions are satisfied.

1) The measurement data of the ith device shows a significant increase in the $(k-1)$th time slice [i.e., $\Omega(D_i^{k-1}) = T$ or A].
2) The measurement data of the ith device is still increasing or it is significantly different to that of the device's neighbors in the kth time slice [i.e., $\Omega(D_i^k) = S$ or A].

Note that the RED module is only effective in detecting small areas of PM2.5 emissions. It cannot be used for cases of large-scale emissions, where PM2.5 particles have been dispersed from the emission source over a wide area for a long period. Moreover, the RED module requires a continuous sensor measurement data stream without missing data. In the worst case, its response time is about double the time slice length.

C. Device Ranking

Based on the results obtained by the TSAD module, the DR module provides a ranking for each sensing device by aggregating the results of the RED module on a daily basis. The ranking is based on the sensing device's "anomaly ratio" in all the time-sliced sensor measurement data collected in one day.

Let $\Delta(a, b)$ be a comparison function that returns 1 if the anomaly type a is equal to b; otherwise, it returns 0. Thus, the ranking of the ith device on day X can be obtained by

$$R_i^X = 1 - \frac{\sum_{x\in X} \Delta\big(\Omega\big(D_i^x\big), S\big) + \Delta\big(\Omega\big(D_i^x\big), T\big) + \Delta\big(\Omega\big(D_i^x\big), A\big)}{|X| - \sum_{x\in X} \Delta\big(\Omega\big(D_i^x\big), M\big)} \tag{4}$$

where $|X|$ is the number of time slices for fixed-length discrete representation of the sensor measurement data in a day.

The ranking is effective in indicating the *consistency level* of a sensing device in its neighboring area and across contiguous time slices. Generally, the higher the ranking, the more confidence there will be in the device's measurement results. However, a lower ranking does not necessarily mean the device's performance is poor. It only indicates that the device must be examined further to determine if it is working properly, or whether additional devices should be deployed to obtain more information about its neighboring area.

D. Malfunction Detection

Finally, the MD module identifies malfunctioning devices based on the daily results of the TSAD module. To implement the MD module, we extend the TSAD module to divide the anomaly type S into "S_L" and "S_H." Specifically, S_L indicates that the device is a spatial anomaly with a much lower sensing value than its neighbors; while S_H indicates that the device is a spatial anomaly with a much higher sensing value than its neighbors. The MD module provides the following four types of outcomes.

1) *Indoor Devices:* For PM2.5 monitoring applications, indoor sensing devices tend to yield much lower PM2.5 measurement results due to good air purification of the indoor HVAC system. As a result, the analysis of the sensing data for different applications may be misleading. Although all the participants of this PM2.5 monitoring project were asked to deploy their sensing devices outdoors with good air circulation, we found that some devices were still deployed indoors. To identify devices that are located indoors, we compute the ratio of the anomaly type S_L among all RED outcomes for the ith device on day X by (5). We determine that the ith device is indoors if $\Phi_i^{S_L} > 1/3$, i.e., one third of the PM2.5 measurements of the ith device are significantly lower than those of its neighbors

$$\Phi_i^{S_L} = \frac{\sum_{x\in X} \Delta\big(\Omega\big(D_i^k x\big), S_L\big)}{|X| - \sum_{x\in X} \Delta\big(\Omega\big(D_i^x\big), M\big)}. \tag{5}$$

2) *Devices Near Emission Sources:* Similar to the detection of indoor devices, we compute the ratio of the anomaly type S_H among all RED outcomes for the ith device on day X by (6). We determine that the ith device is close to an emission source (e.g., a temple or a BBQ restaurant) if $\Phi_i^{S_H} > 1/3$

$$\Phi_i^{S_H} = \frac{\sum_{x\in X} \Delta\big(\Omega\big(D_i^x\big), S_H\big)}{|X| - \sum_{x\in X} \Delta\big(\Omega\big(D_i^x\big), M\big)}. \tag{6}$$

3) *Malfunctioning Devices:* We consider cases, where the sensing device may have been installed incorrectly, or the sensor is polluted or too old. In each scenario, the sensing device yields extreme values (either very high or very low) continuously. Therefore, we determine the ith device is malfunctioning if $\Phi_i^{S_H} > 2/3$ or $\Phi_i^{S_L} > 2/3$.
4) *Undetectable Devices:* The detection of malfunctioning devices relies to a great extent on the spatial anomaly detection in the TSAD module, so the MD module may not function well for sensing devices that do not have a sufficient number of neighbors. Thus, when $|N_i| < 3$, the MD module also reports that the ith device is undetectable.

V. Implementation

In this section, we describe the implementation of the proposed ADF and analyze the anomaly detection results.

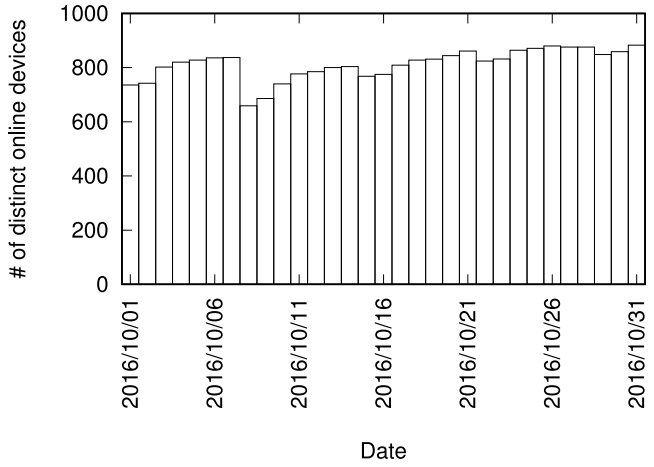


Fig. 9. Number of distinct AirBox devices online each day in the dataset.

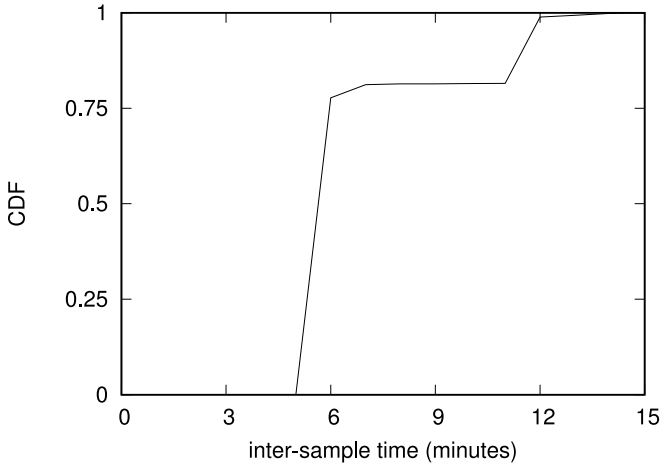


Fig. 10. CDF of the time interval between every two contiguous samples for each PM2.5 sensing device in the dataset.

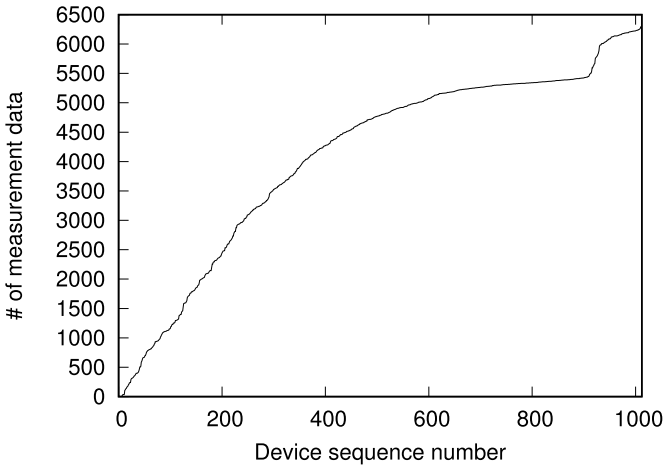


Fig. 11. Distribution of the amount of measurement data for each AirBox device in the dataset.

A. AirBox Dataset

We performed an extensive analysis of the AirBox dataset, which is based on the measurement data collected by the AirBox devices from October 1, 2016 to October 31, 2016. As the devices donated to the elementary schools are maintained by dedicated people with reliable wireless connections and power sources, it was assumed that the quality of the measurement data would be better. Therefore, we only considered those datasets for the analysis.

Fig. 9 shows the number of distinct devices online during the above data collection period. There is a distinct weekly pattern, which demonstrates that more devices are online on weekdays than on the weekends. The reason is that some schools turn off all electronic devices on weekends to save power. Moreover, the number of devices online varies substantially on different days because the devices may be ON or OFF depending on the installation environment and the wireless connection may not be reliable all the time.

Fig. 10 shows the cumulative distribution function (CDF) of the intersample time between every two contiguous samples collected by an AirBox device. Cases with an intersample time greater than 15 min were not considered because the manufacturer claims that the sampling frequency is every five minutes. The results show that the intersample time is approximately 6 min in 80% of the analyzed cases, and about 12 min in 20% of the cases. This is because the AirBox device stays in the standby mode for five minutes between samplings and takes about one minute to complete a sampling action (including waking up the device, reconnecting to the Internet, and polling all sensors to collect measurement data). Thus, the intersample time is about six minutes; however, if the transmission of the first measurement fails, the time increases to 12 min.

Based on the intersample time analysis, we obtain the expected amount of measurement data per device in the dataset by (60 (min/h)/6 (min)) $\times$ 24 (h/day) $\times$31(day) $=$ 7440. However, as shown in Fig. 11, about 40% of the devices contribute less than 60% of the expected amount of data (i.e., 7440 $\times 0.6 = 4464$); and only about 10% of the devices contribute more than 80% of the expected amount (i.e., 7440 $\times 0.8 = 5952$). There are several reasons for such low participation rates. For instance, some elementary schools shut down electrical power in nonworking hours; and some AirBox devices are offline due to unstable Internet connection at the deployment sites. Although the results indicate that participation in the current AirBox deployment could be improved, the amount of measurement data collected is sufficient for further analysis and use in applications.

B. TSAD Parameter Tuning

Using the AirBox dataset, the impact of the d values (i.e., the threshold used to define the neighboring area of an AirBox device) is evaluated in terms of spatial anomaly detection in the TSAD module. Fig. 12 shows the CDF of the number of neighboring AirBox devices in the dataset under different d values. The results indicate that the greater the d value, the larger the number of neighboring devices that must be considered in the TSAD module. Moreover, the offset between two contiguous curves decreases as the value of d increases.

The selected d value has to strike a good balance between two factors: 1) the value of d should be as large as possible to minimize the number of *undetectable* devices, i.e., the number

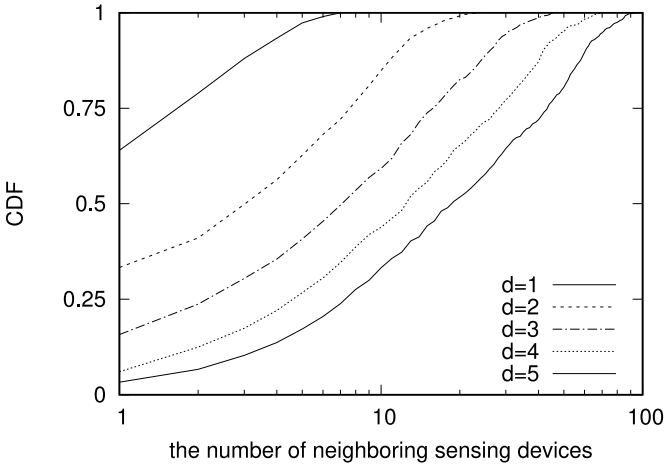


Fig. 12. CDF of the number of neighboring PM2.5 sensing devices under different d settings in the dataset.

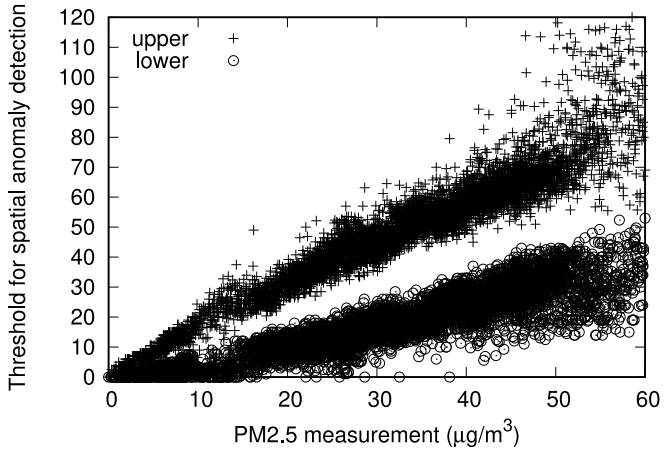


Fig. 13. Distribution of the upper and lower thresholds in the dataset under different PM2.5 measurement values for spatial anomaly detection when $d = 3$ km.

of the neighboring devices is less than three and 2) the value of d should be as small as possible to ensure the consistency in the measurement of the real PM2.5 concentration in the neighborhood area. Thus, based on the results in Fig. 12, d is set at 3 km for the TSAD module in this paper.

Using the selected d setting, Fig. 13 shows the correlations between the PM2.5 measurement result of each AirBox device and the upper/lower thresholds determined by its neighboring devices [ref. (1) and (2)]. The results demonstrate that both the upper and lower thresholds increase with the measurement results of the corresponding AirBox device. Among the 6757 effective instances (i.e., after removing the cases of undetectable devices), there are 404 instances, where the measurement data is greater than the upper threshold. In addition, there are 651 instances, where the measurement data is smaller than the lower threshold. Overall, the number of spatial anomalies detected is 1055, which is approximately 15% of all the effective instances.

Next, we analyze the distribution of the measurement offsets (absolute values) between every two contiguous samples of the same device in the dataset. Fig. 14 shows that the mean and the median of the offsets are at a consistently low level when

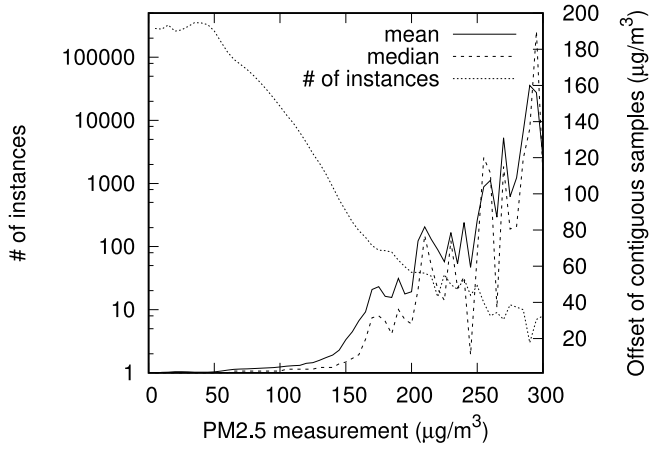


Fig. 14. Distribution of the absolute offsets between every two contiguous samples of the same AirBox device in the dataset.

the PM2.5 measurement is less than 100 μg/m^3. However, the values increase dramatically when the PM2.5 measurement is greater than 100 μg/m^3. The reasons are as follows. First, the particle sensor used in the AirBox device is accurate when the PM2.5 concentration is in the range 30 to 100 μg/m^3. When the concentration is greater than 100 μg/m^3 [43], the sensor's measurement error may be as high as 50% of its true value. Second, the air pollutant mixture is unstable and unpredictable when the PM2.5 concentration is higher, so there is a greater offset between two contiguous samples of the same device.

Let ρ_p be the mean offset of every two contiguous measurement results when the later PM2.5 measurement is p μg/m^3, and let σ_p be its standard deviation. Then, we let $\omega_p = \rho_p + 2\sigma_p$ represent the threshold of a temporal anomaly when the PM2.5 concentration is k. Applying the linear regression technique by fixing the constant factor at zero, the correlation is obtained with the R-square value of 0.96 when $30 \leq p \leq 100$

$$\omega_p = 0.1794p \approx \frac{p}{6}. \tag{7}$$

In addition, the threshold of a temporal anomaly should be greater than a constant that represents potential PM2.5 emissions near the AirBox device, where the constant is set at $\omega_{30} = 5$. Therefore, when the PM2.5 concentration is p, the threshold of a temporal anomaly can be obtained by

$$\delta_p = \max\left(5, \frac{p}{6}\right). \tag{8}$$

We note that the parameters of the TSAD module need to be periodically updated, so that the proposed AFD scheme can adapt to more up-to-date environment scenarios and behaves more effectively. The update process can be done incrementally without interrupting the AFD scheme, as all the calculations can be conducted offline and the results can be applied directly.

C. Anomaly Detection Results

Using the suggested parameters for the TSAD module [i.e., $d = 3$ km and $\delta_p = \max(5, p/6)$], we implemented the RED, DR, and MD modules. The results of the three modules are released regularly as open data services for the AirBox

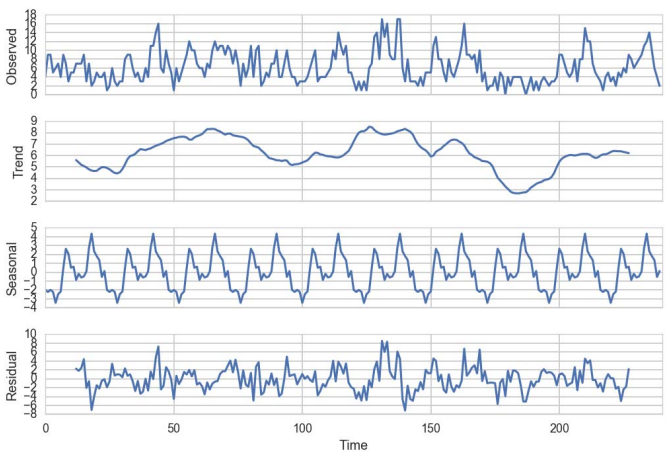


Fig. 15. Time series decomposition of the number of emission events detected by the RED module in the AirBox system from December 16, 2016 to December 25, 2016.

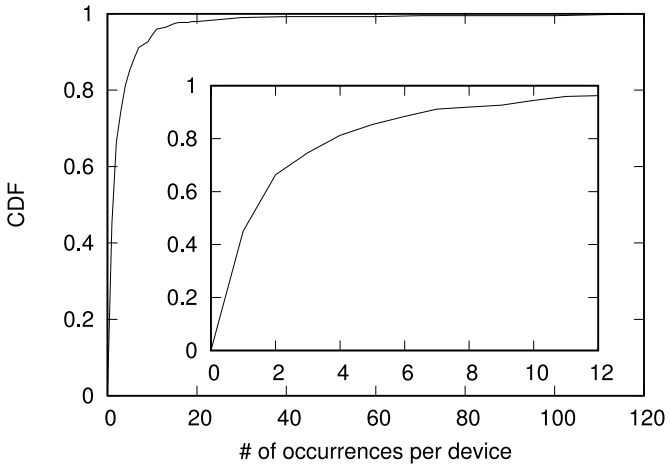


Fig. 16. CDF of the number of emission events detected by each device in the AirBox system from December 16, 2016 to December 25, 2016.

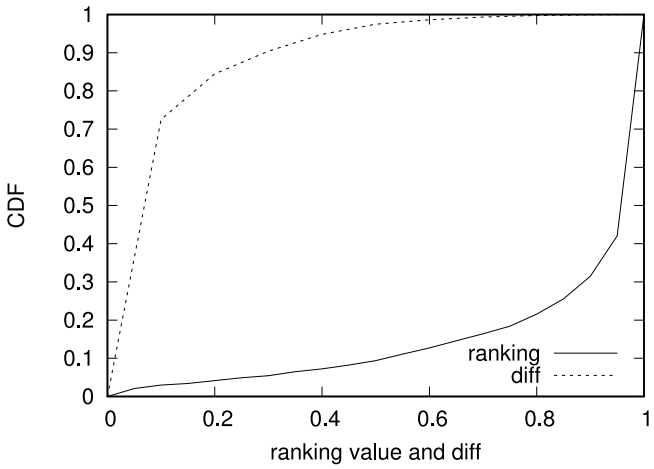


Fig. 17. CDF of the ranking results for all devices on each day of the 10-day observation, and the CDF of the difference between two ranking results of the same device on two contiguous days.

project [47]–[50]. In the following sections, we report the statistics of the RED, DR, and MD outcomes based on a 10-day observation period (from December 16, 2016 00:00 to December 25, 2016 23:59 UTC time) from 1272 distinct AirBox devices located in five cities across Taiwan, namely, Taipei, New Taipei, Taichung, Tainan, and Kaohsiung. Note that we decided to use two datasets in this paper because the original dataset (October 1, 2016 to October 31, 2016) was used for parameter tuning, while the second one (December 16, 2016 to December 25, 2016) was used for anomaly detection experiments based on the parameter tuning results.

1) Results of the RED Module: Fig. 15 shows the time series decomposition of the number of potential emission events detected per hour in the 10-day dataset. There are three findings.

1) The *Observed* figure shows that the number of emission events detected each hour with a lot of oscillations ranged between 0 and 17 over time.
2) The *Trend* figure shows that the emission events per hour also oscillated over time, and the minimum occurred on the morning of December 23rd. That was a rainy day between two smog episodes, resulting in a fairly good air quality in all five cities.
3) The *Seasonal* figure shows that there is a distinct pattern in a daily cycle with peaks occurring in the pattern between 8:00 and 18:00, which corresponds to the time period that people are at work.

Moreover, the RED module detected that 376 AirBox devices recorded regional emission events in the 10-day dataset. Among them, 94% had less than ten occurrences (i.e., once a day on average), and only 2% had more than 20 occurrences, as shown in Fig. 16. The results demonstrate that the RED module is effective in detecting occasional emission events that were difficult to detect previously. With a densely deployed environmental sensing system and the RED module implemented, such occasional emissions can now be identified in almost real time. This enables government authorities to react to emission events on-demand accordingly

2) Results of the DR Module: Using the daily results of the DR module in the 10-day dataset, we calculate the CDF of the raw ranking values for all the devices, as well as the CDF of the difference between two ranking values of the same device on every two contiguous days. Fig. 17 shows that only 21% of the ranking values are below 0.8. The result indicates that the measurement data of most AirBox devices is consistent both temporally and spatially. Moreover, we observe that 84.5% of the ranking difference on two contiguous days is below 0.2, which demonstrates that the data quality of the AirBox devices is stable with only a small number of oscillations in their ranking values.

3) Results of the MD Module: Fig. 18 shows the results of the MD module for the 10-day dataset. Among a total of 327 devices in the dataset that were deemed undetectable, 203 were judged as always undetectable. The MD module regarded the other 124 devices as only occasionally undetectable. The reason is that those devices had a small number of neighboring devices, some of which were turned off during the 10-day observation period. Because the assessment of undetectable devices is based on the number of neighboring devices within

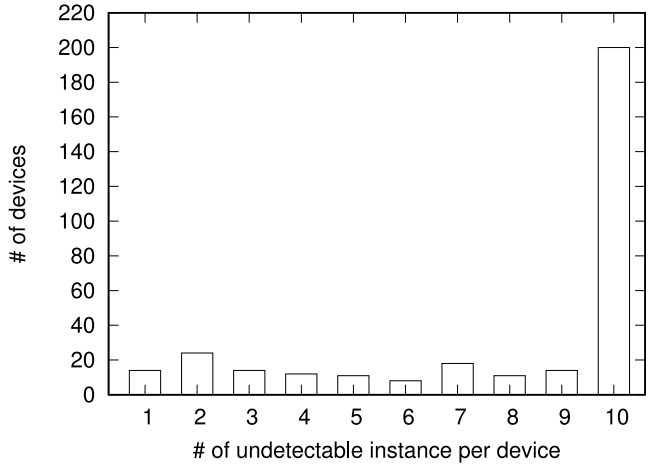


Fig. 18. Histogram of the number of devices with different numbers of occurrences of undetectable results in the AirBox system from December 16, 2016 to December 25, 2016.

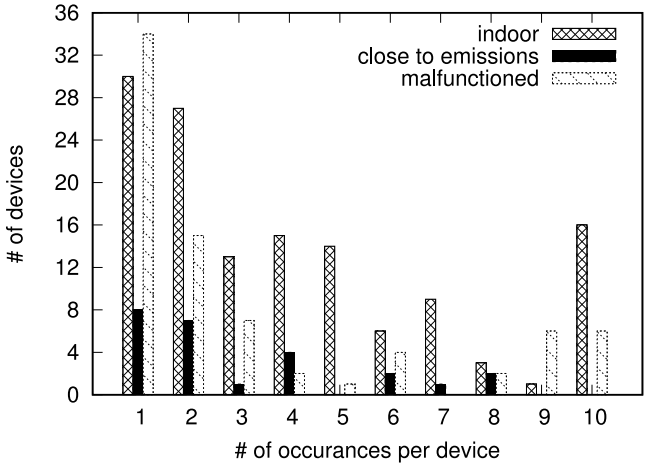


Fig. 20. Histogram of the number of devices with different numbers of occurrences of indoor, close to emission, and malfunctioning results in the AirBox system from December 16, 2016 to December 25, 2016.

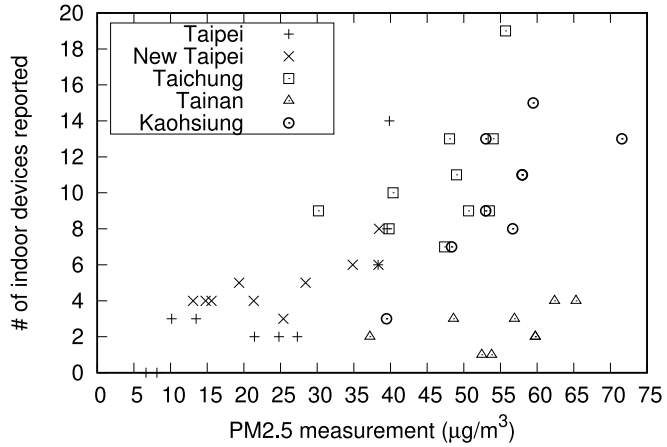


Fig. 19. Correlation between the number of indoor devices detected daily and the daily PM2.5 emissions in the five deployment cities from December 16, 2016 to December 25, 2016.

a geometric distance of d km, devices are deemed undetectable when the number of their online neighboring devices is less than three.

Fig. 19 shows the correlation between the number of daily indoor devices reported by the MD module and the daily PM2.5 average in the five deployment sites in the 10-day dataset. We observe that: 1) the distribution of PM2.5 concentrations in each city is different; 2) even in the same city, the daily PM2.5 concentration varied substantially in the dataset; and 3) generally, the greater the PM2.5 concentration, the larger the number of indoor devices reported by the MD module. The reason is that people tend to turn on air purifiers when the atmospheric PM2.5 concentration is high. This increases the difference between the indoor and outdoor measurement results, so it is easier for the MD module to identify indoor devices.

Fig. 20 shows the number of devices that the MD module identified as indoors, malfunctioning, or close to an emission source on each day of the 10-day observation period. Only 16 devices were deemed indoor devices and six devices were listed as malfunctioning. In addition, a few devices were identified as indoors, close to an emission source, or malfunctioning several times in the 10-day dataset. The results contradict our intuition that the number of these devices should be constant over a short period. The reasons are: 1) when the atmospheric PM2.5 concentration is low, it is difficult for the MD module to detect indoor devices and 2) when the atmospheric PM2.5 concentration is high, the MD module has difficulty distinguishing between a malfunctioning device and a device close to an emission source. Because PM2.5 concentrations vary a great deal in the 10-day dataset, it is normal that there are oscillations in the detection of different types of devices in Fig. 20. The results also indicate that to improve the identification of the sensing devices, we should consider the atmospheric PM2.5 concentration in the MD module and identify each device based on a continuous sequence of MD outcomes.

D. Discussion

In this paper, we do not include a benchmark of the proposed framework in this paper because benchmarking an AirBox-like IoT system is still an open challenge for two reasons: 1) the system is grassroots in nature and there is no way to control each participating node and 2) the dataset has 1272 distinct devices covering more than 1000 km^2 area that makes it impossible to verify outcomes of the proposed framework in practice. Instead of conducting benchmark experiments, we tackle the performance evaluation issue in another way that: 1) we publicize the computation results of the RED, DR, and MD modules on webpages and ask for comments; 2) we push the computation results of the RED module to local EPA and ask for their feedbacks; and 3) we verify some of the computation results of the MD module by directly talking to those elementary schools with AirBox devices that are frequently identified as indoor devices or closing to emission sources. So far (since December 2016 and till October 2017), we have not received complaints about the computation results of the proposed framework, and the outcome of the RED module has been considered a reliable data source for smart inspection

and smart governance by local EPA and governments. Thus, we are confident of the computation results of the proposed framework, and we also release the dataset of this paper for future benchmark comparison research [51].

VI. Conclusion

We have proposed an ADF to ensure the data quality for large-scale environmental sensing systems. The framework comprises four modules, namely, TSAD, RED, DR, and MD. Using the AirBox PM2.5 monitoring system as an example, we analyzed the data to identify the intrinsic properties of the measurement datasets. First, TSAD divide the continuous real-time sensor measurement data into a sequence of fixed-length discrete representations using the time-slicing technique. Once the parameters to be used in the TSAD module are determined, the RED module reports potential regional emission sources every hour using the real-time AirBox data streams. Meanwhile, the DR module ranks the consistency of each sensing device over time and the cross-device consistency with its neighboring devices. Finally, the MU module identifies malfunctioning devices based on the daily results of the TSAD module. Using the daily summary of TSAD, the attributes and status of each device are evaluated by the system; that is, whether a device is deployed indoors, or close to an emission source. The analyzed results of the proposed approach are available to the public in open data format. They have been used by various parties for data visualization, on-demand responses, and future policy-making. Because of its simple design, ADF is highly extensible to other advanced applications, and it can be exploited to support other large-scale environmental sensing systems.

Acknowledgment

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

References

[1] D. Ballas, "What makes a 'happy city'?" *Cities*, vol. 32, pp. S39–S50, Jul. 2013.
[2] R. Giffinger, "Smart cities: Ranking of European medium-sized cities," Centre Regional Sci., Vienna Univ. Technol., Vienna, Austria, Tech. Rep., Oct. 2007.
[3] *US EPA Smart City Air Challenge*. Accessed: Dec. 20, 2016. [Online]. Available: https://www.challenge.gov/challenge/smart-city-air-challenge/
[4] X. Tang, "An overview of air pollution problem in megacities and city clusters in China," in *Proc. AGU Spring Meeting Abstracts*, May 2007, Art. no. A23C-03.
[5] VUFO—NGO Resource Centre Vietnam. (Sep. 19, 2013). *Vietnam Named Among Top Ten Nations With Worst Air Pollution*. [Online]. Available: http://www.ngocentre.org.vn/news/vietnam-named-among-top-ten-nations-worst-air-pollution
[6] B. Ostro, "WHO environmental burden of disease series," in *Outdoor Air Pollution: Assessing the Environmental Burden of Disease at National and Local Levels*. Geneva, Switzerland: World Health Org., 2004.
[7] Y.-F. Xing, Y.-H. Xu, M.-H. Shi, and Y.-X. Lian, "The impact of PM2.5 on the human respiratory system," *J. Thorac. Disease*, vol. 8, no. 1, pp. E69–E74, Jan. 2016.
[8] *AirNow*. Accessed: Dec. 20, 2016. [Online]. Available: https://airnow.gov
[9] *Air Quality Data—Central Pollution Control Board*. Accessed: Dec. 20, 2016. [Online]. Available: http://cpcb.nic.in/RealTimeAirQualityData.php
[10] *Air Pollution—European Environment Agency*. Accessed: Dec. 20, 2016. [Online]. Available: https://www.eea.europa.eu/themes/air/intro
[11] M. Markiewicz, "A review of mathematical models for the atmospheric dispersion of heavy gases. Part I. A classification of models," *Ecol. Chem. Eng. S*, vol. 19, no. 3, pp. 297–314, Jul. 2012.
[12] S.-C. C. Lung, I.-F. Maod, and L.-J. S. Liu, "Residents' particle exposures in six different communities in Taiwan," *Sci. Total Environ.*, vol. 377, no. 1, pp. 81–92, May 2007.
[13] S.-C. C. Lung *et al.*, "Variability of intra-urban exposure to particulate matter and CO from Asian-type community pollution sources," *Atmos. Environ.*, vol. 83, pp. 6–13, Feb. 2014.
[14] M. Alvarado, F. Gonzalez, A. Fletcher, and A. Doshi, "Towards the development of a low cost airborne sensing system to monitor dust particles after blasting at open-pit mine sites," *Sensors*, vol. 15, no. 8, pp. 19667–19687, 2015.
[15] M. Budde, R. E. Masri, T. Riedel, and M. Beigl, "Enabling low-cost particulate matter measurement for participatory sensing scenarios," in *Proc. Int. Conf. Mobile Ubiquitous Multimedia*, 2013, Art. no. 19.
[16] Y. Cheng *et al.*, "AirCloud: A cloud-based air-quality monitoring system for everyone," in *Proc. ACM SenSys*, 2014, pp. 251–265.
[17] S. Devarakonda *et al.*, "Real-time air quality monitoring through mobile sensing in metropolitan areas," in *Proc. ACM SIGKDD Int. Workshop Urban Comput.*, 2013, Art. no. 15.
[18] Y. Gao *et al.*, "Mosaic: A low-cost mobile sensing system for urban air quality monitoring," in *Proc. IEEE INFOCOM*, San Francisco, CA, USA, 2016, pp. 1–9.
[19] K. Weekly *et al.*, "Low-cost coarse airborne particulate matter sensing for indoor occupancy detection," in *Proc. IEEE Int. Conf. Autom. Sci. Eng.*, Madison, WI, USA, 2013, pp. 32–37.
[20] Y. Zhuang, F. Lin, E.-H. Yoo, and W. Xu, "AirSense: A portable context-sensing device for personal air quality monitoring," in *Proc. ACM MobileHealth*, 2015, pp. 17–22.
[21] *Array of Things*. Accessed: Dec. 20, 2016. [Online]. Available: https://arrayofthings.github.io
[22] *Taipei AirBox*. Accessed: Dec. 20, 2016. [Online]. Available: http://pm2.5.taipei/
[23] *OpenSense at ETH Zurich*. Accessed: Dec. 20, 2016. [Online]. Available: http://www.opensense.ethz.ch/
[24] *AirCasting*. Accessed: Dec. 20, 2016. [Online]. Available: http://aircasting.org
[25] *Clarity*. Accessed: Dec. 20, 2016. [Online]. Available: http://joinclarity.io
[26] *Laser Egg*. Accessed: Dec. 20, 2016. [Online]. Available: http://laseregg.origins-china.com
[27] L.-J. Chen, W. Hsu, M. Cheng, and H.-C. Lee, "Demo: LASS: A location-aware sensing system for participatory PM2.5 monitoring," in *Proc. ACM MobiSys*, 2016, p. 98.
[28] *uHoo*. Accessed: Dec. 20, 2016. [Online]. Available: http://uhooair.com
[29] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 12, no. 2, pp. 159–170, 2nd Quart., 2010.
[30] H. Ayadi, A. Zouinkhi, B. Boussaid, and M. N. Abdelkrim, "A machine learning methods: Outlier detection in WSN," in *Proc. IEEE Int. Conf. Sci. Tech. Autom. Control Comput. Eng.*, 2015, pp. 722–727.
[31] S. A. Haque, M. Rahman, and S. M. Aziz, "Sensor anomaly detection in wireless sensor networks for healthcare," *Sensors*, vol. 15, no. 4, pp. 8764–8786, Apr. 2015.
[32] M. A. Hayes and M. A. Capretz, "Contextual anomaly detection framework for big sensor data," *J. Big Data*, vol. 2, no. 2, p. 22, 2015.
[33] M. Moshtaghi, S. Rajasegarar, C. Leckie, and S. Karunasekera, "Anomaly detection by clustering ellipsoids in wireless sensor networks," in *Proc. IEEE Int. Conf. Intell. Sensors Sensor Netw. Inf. Process.*, Melbourne, VIC, Australia, 2009, pp. 331–336.
[34] J. Murphree, "Machine learning anomaly detection in large systems," in *Proc. IEEE AUTOTESTCON*, Anaheim, CA, USA, 2016, pp. 1–9.
[35] I. C. Paschalidis and Y. Chen, "Statistical anomaly detection with sensor networks," *ACM Trans. Sensor Netw.*, vol. 7, no. 2, p. 17, Aug. 2010.
[36] W. Wu *et al.*, "Localized outlying and boundary data detection in sensor networks," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 8, pp. 1145–1157, Aug. 2007.
[37] Edimax Inc. *AirBox: PM2.5 Sensing for Smart Cities*. Accessed: Dec. 20, 2016. [Online]. Available: https://airbox.edimaxcloud.com
[38] J. N. R. Jeffers, *Practitioner's Handbook on the Modelling of Dynamic Change in Ecosystems* (SCOPE Report). Chichester, U.K.: Wiley, 1988.

[39] L.-J. Chen *et al.*, "An open framework for participatory PM2.5 monitoring in smart cities," *IEEE Access*, vol. 5, pp. 14441–14454, 2017.
[40] D. L. Mills, "Network time protocol specification, implementation and analysis," Internet Eng. Task Force, Fremont, CA, USA, RFC 1305, Mar. 1992.
[41] H. Grimm and D. J. Eatough, "Aerosol measurement: The use of optical light scattering for the determination of particulate size distribution, and particulate mass, including the semi-volatile fraction," *J. Air Waste Manag. Assoc.*, vol. 59, no. 1, pp. 101–107, Jan. 2009.
[42] A. Morpurgo, F. Pedersini, and A. Reina, "A low-cost instrument for environmental particulate analysis based on optical scattering," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf.*, Graz, Austria, 2012, pp. 2646–2650.
[43] AQICN.org. *The Plantower PMS5003 and PMS7003 Air Quality Sensor Experiment*. [Online]. Accessed: Dec. 20, 2016. Available: http://aqicn.org/sensor/pms5003-7003/hk/
[44] *PM2.5 Concentration Indexes and Activity Advices*. Accessed: Dec. 20, 2016. [Online]. Available: http://taqm.epa.gov.tw/taqm/tw/fpmi.aspx
[45] *PM2.5 Open Data Portal*. Accessed: Dec. 20, 2016. [Online]. Available: https://pm25.lass-net.org
[46] J. W. Tukey, *Exploratory Data Analysis*. Reading, MA, USA: Addison-Wesley, 1977.
[47] *The API for Detecting Potential Regional Emission Sources Detected (Hourly)*. Accessed: Dec. 20, 2016. [Online]. Available: https://data.lass-net.org/ data/device_pollution.json
[48] *The API for the Ranking Results of the AirBox Devices (Daily)*. Accessed: Dec. 20, 2016. [Online]. Available: https://data.lass-net.org/data/device_ranking.json
[49] *The API for Potential Indoor AirBox Devices (Daily)*. Accessed: Dec. 20, 2016. [Online]. Available: https://data.lass-net.org/data/device_indoor.json
[50] *The API for Malfunctioning AirBox Devices (Daily)*. Accessed: Dec. 20, 2016. [Online]. Available: https://data.lass-net.org/data/device_malfunction_daily.json
[51] *AirBox Dataset*. Accessed: Dec. 20, 2016. [Online]. Available: https://sites.google.com/site/ccljj/dataset-airbox

Ling-Jyh Chen (S'03–M'05–SM'12) received the B.Ed. degree in information and computer education from National Taiwan Normal University, Taipei, Taiwan, in 1998, and the M.S. and Ph.D. degrees in computer science from the University of California at Los Angeles, Los Angeles, CA, USA, in 2002 and 2005, respectively.

His current research interests include wireless networks, mobile computing, network measurements, and social computing.

Yao-Hua Ho received the B.S., M.S., and Ph.D. degrees in computer science from the University of Central Florida, Orlando, FL, USA, in 2001, 2002, and 2009, respectively.

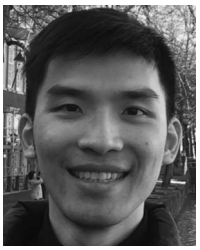
He joined the Department of Computer Science and Information Engineering, University of Central Florida, in 2012. He has authored or co-authored many papers, including several that have been recognized as best/top papers at various international conferences. His current research interests include social networks and computing, mobile, and wireless networks (WLAN/WSN/MANET/VNET), network protocols, mobile applications, location-aware service, and network measurements.

Hsin-Hung Hsieh received the B.S. and M.S. degrees in computer science and information engineering from National Taiwan Normal University, Taipei, Taiwan, in 2011 and 2013, respectively.

She is currently a full-time Research Assistant with the Network Research Laboratory, Academia Sinica, Taipei. Her current research interests include image processing, data analysis, and Internet of Things.

Shih-Ting Huang received the B.S. degree in computer science from National Chengchi University, Taipei, Taiwan, in 2016. She is currently pursuing the master's degree in computer science at New York University, New York, NY, USA.

She was a full-time Research Assistant with the Network Research Laboratory, Academia Sinica, Taipei. Her current research interests include computer graphics and data analysis.



Hu-Cheng Lee received the B.S. and M.S. degrees in computer science and information engineering from National Taiwan Normal University, Taipei, Taiwan, in 2011 and 2013, respectively.

He is currently a full-time Research Assistant with the Network Research Laboratory, Academia Sinica, Taipei. His current research interests include Internet of Things, sensor systems, and mobile computing.

Sachit Mahajan (GS'17) received the B.Tech. degree in electrical and computer engineering from Punjab Technical University, Kapurthala, India, in 2012, and the M.S. degree in communication engineering from the University of Manchester, Manchester, U.K., in 2013. He is currently pursuing the Ph.D. degree in social networks and human centered computing at the Network Research Laboratory, Academia Sinica, Taipei, Taiwan.

His current research interests include machine learning, data science, and the Internet of Things.