

# Air Pollution Monitoring and Spatial-Temporal Hotspot Pattern Analysis of Sensors Based on Sensor Grid for the Industrial Parks in Taiwan

Bing-Si Ni

Department of Geography  
National Taiwan Normal  
University  
Taipei City, Taiwan (R.O.C)  
email: nispc@nisp.c.tw

Yu-Chieh Huang

Environmental Management  
Consultant Technologies Inc.  
Taipei City, Taiwan (R.O.C)  
email: sophia@emct.com.tw

Chun-Ming Huang

Environmental Protection  
Administration  
Taipei City, Taiwan (R.O.C.)  
email: chmihuang@epa.gov.tw

**Abstract**—To identify sources of pollution and predict future pollution events, the Environmental Protection Administration of Taiwan has deployed dense sensor networks in industrial districts. In face of overwhelming real-time data collected from the Internet of Things (IoT) applications for smart environmental sensing, no standard procedure based on the space-time statistical methods, such as Getis-Ord  $G^*$  or Moran's  $I$  exist for defining and analyzing pollution events. We used raw data generated from microsensors as the data source, adopted spatial statistics to perform hotspot analysis, then define the event base on the result of statistical hypothesis and grid connectivity. This approach was effective in distinguishing independent pollution events when two or more events occurred concurrently in the same region. Finally, spatial and temporal descriptive statistical analysis was performed on the targeted pollution events, including the identity of pollution events through spatial-temporal hotspot analysis integrated with data visualization.

**Keywords**—spatial-temporal patterns analysis; air pollution events; Internet of Things; sensors; industrial parks

## I. INTRODUCTION

The media coverage and the government's environmental policies raised environmental awareness among the general population and increased attention to air pollution. Excessive levels of ozone and fine particulate matter which less than 2.5 micrometers in diameter in the air, known as "PM<sub>2.5</sub>", pose a considerable threat to human health. Air quality has become a crucial indicator of people's quality of life, and small-scale air pollution monitoring was seen as increasing demand. In Taiwan, air quality monitoring is typically performed by examining data from the network of national air quality monitor stations. However, the limitations of micro-sensors mean that air quality determined using a single datum cannot be used as evidence for inspection. Therefore, this study organized sensor data into clusters and adopted spatial-temporal statistics to solve problems concerning the processing of microsensor data; in addition, data mining was employed to identify trends in the data clusters. Many industrial districts of various cities in Taiwan have begun to establish microsensor networks embedded in streetlights, which in the future could serve as a source of real-time monitoring for emerging air pollutions or air quality monitoring. Specifically, data regarding local weather

dynamics are integrated into microsensor networks, which can be used not only for tracing but also for predicting short-term pollution events to rapidly identify pollution sources.

This paper is organized as follows: Section II introduces the research methods, including statistical methods and the weight matrix we adopted. Section III presents the results. Section IV concludes the paper.

## II. RESEARCH METHODS

Studies have reached no consensus on the definitions of pollution clusters or events, and spatial analysis has been a bottleneck in Statistics. Since the 1990s, in addition to national air quality monitoring systems, which are 77 stations in Taiwan, newly deployed microsensors (more 3,000) have monitored PM<sub>2.5</sub> as an indicator of pollution caused by suspended particulates, volatile organic compounds (VOCs), and black carbons emitted from factories. For various pollutants in the atmosphere, such as the above-mentioned PM<sub>2.5</sub> and VOCs, there are also nitrogen oxides (NO<sub>x</sub>), sulfides (SO<sub>2</sub>). These pollutants have different production factors and have different effects on human health. Therefore, the sources of these substances are often discussed in previous researches.

Many studies are discussing the relationship between pollution concentrations and other factors, such as the relationship between different human activities and various pollutant concentrations, the covariation between different pollutants [1], and the relationship between weather factors and various pollutants [2]. In the studies mentioned above, researchers describe the pollution event along the concentration of contaminants, duration, return period, etc. They develop the follow-up study on the characteristics of the above pollutants. Additionally, the EU government also controls the contaminants by the average concentration over a period of time as standards [3]. However, spatial autocorrelation has not been widely used in the above studies. Therefore, there is still a lack of useful indicators for the accumulation of pollutants caused by small-scale human activities in industrial parks, helping researchers to identify the development, concentration, and diffusion of pollution clusters in the study area.

This study performed a hotspot analysis based on spatial statistics approaches. However, a small industrial park in Taiwan is typically 10 – 15 km<sup>2</sup>; the occurrence of a severe pollution event easily affects all the devices in the overall industrial park at the same time, so only considering the spatial dimension is insufficient because it will lack relative

reference values. Therefore, we must also incorporate the temporal dimension to determine the temporal continuity and trend of the events hotspots. That implies we should conduct a spatial-temporal extended version of spatial autocorrelation instead of pure spatial autocorrelation to prove the assumption [4].

First, data aggregation was conducted to reduce the time complexity involved in data preprocessing. Second, using “Global Moran’s I” (1) to confirm whether there is a significant autocorrelation on the concentrations of fine particulate matter through time and space [5]. Third, local spatial statistical techniques such as “Local Getis-Ord  $G_i^*$ ” (2) or “Local Moran’s I” were employed to determine the distribution of cold spots and hot spots. [5, 6]

$$I = \frac{\sum_i \sum_j w_{ij} z_i z_j / S_0}{\sum_i z_i^2 / n} \quad (1)$$

$$G_i^* = \frac{\sum_j w_{ij} x_j}{\sum_j x_j} \quad (2)$$

Both global and local spatial autocorrelation methods need to define a weight matrix  $w_{ij}$  to point out the degree of dependency between every two elements. Our weight matrix is based on the spatial-temporal neighbors and generated by the three-dimensional “Queen” rule (Fig.1) as in [7]. The identified spatial-temporal hotspots served as the basis for defining pollution events. Finally, the pollution events were analyzed using descriptive statistics, and the results were applied for subsequent analysis of pollution trends and patterns.

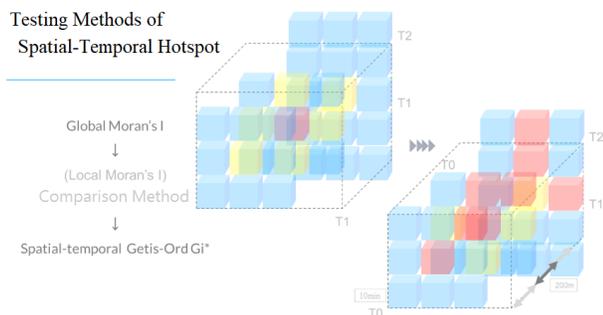


Figure 1. Data processing and testing methods.

### III. EMPIRICAL DATA DISCUSSION

The study examined the Dafa Industrial District in Kaohsiung City, a municipality in Taiwan known for its heavy industry. The industrial district is 374 ha with a trapezoid shape. Sensors in the region were deployed every 200 m, with a total of 150 sensors. The Sensors placed in the Dafa Industrial District began operations in September 2018 and encompasses nearly 700 factories that mostly provide services to the light industry and mixed metal-based heavy industry (see Fig. 2).

The raw data were displayed on the leaflet online map which we developed every 5 minutes. During the first month after sensor installation, we found the sensor readings continually increased in the evenings, and the increases were mostly in the southern and northwestern part of the industrial district. From this data, the locations

of pollution sources were identified manually according to factory locations and wind directions.



Figure 2. Deployment of sensors in the study area.

In this paper, we further propose an automated hot spot identification program. The implementation details and parameters are set as follows:

First, we number the grids according to the spatial-temporal locations of the sensors, so that we created  $200 \times 200$  meters grids on the XY-plane, with units of time is equal to 10 minutes. Then, we use eight days of sensor data, the temporal dimensions of space-time cubes were divided into 1,110 grids. Therefore, all data be divided into approximately 69,709 data cubes.

Second, we calculate grid neighbors. Some modules of python like GeoPandas, Pandas, and PySal were used to obtain the adjacent  $5 \times 3 \times 3 - 1$  spatial-temporal neighbors (excluding the grid itself), and produce a weight matrix based on the spatial-temporal neighbors we calculated above.

Third, Spatial-temporal autocorrelation was calculated using the spatial-temporal weight matrix and sensor data. The “Global Moran’s I” is 0.491067 (values of I usually range from  $-1$  to  $1$ ). The result indicates there is a positive autocorrelation. In this step, we performed a hypothesis testing using the Monte Carlo method, and the one-tailed p is 0.000; therefore,  $H_0$  (random distribution) was rejected, indicating that significant clustering existed. The result implies that we can use the local autocorrelation to identify when and where the hot spot and cold spot appear and disappear.

Fourth, we conduct local autocorrelation analysis. The cold spots and hot spots were verified using Getis-Ord’s  $G_i^*$ . The  $G_i^*$  statistics calculated by statistical software are usually converted into Z-scores, which indicates the level of significance and thus could be explained easily.

Finally, we develop a new visualization method combines Z-score statistics, three-dimensional contour plots, and the concept of space-time cubes to present the significance level of local autocorrelation across space and time. With such a tool, let us initially observe whether we proposed is better than the traditional spatial autocorrelation.

TABLE I. COMPARISON OF RESEARCH METHODS

	Experimental group 1*	Experimental group 2	Control group 1	Control group 2
<b>Spatial-temporal interpolation</b>	No	Yes	No	Yes
<b>Definition of neighbors</b>	Spatially and temporally adjacent		Only spatially adjacent	
<b>Calculation results of spatial-temporal <math>G_i^*</math></b>	Non-interpolation	Data pre-interpolation	Non-interpolation	Data pre-interpolation
	1. Temporal (continuous) clustering can be detected.		1. Temporal clustering cannot be detected.	
	2. Nonexistent edge hotspots can be estimated through spatial-temporal interpolation.		2. When the polluted area was large, the hotspots area shrank.	

\*Research Recommendations

Table 1 shows the comparison between traditional pure spatial autocorrelation which was used as the control group (i.e., the spatial  $G_i^*$ ) and spatial-temporal  $G_i^*$  which was used as the experimental group.

For the control groups, the results show that although the hotspots mostly exhibited a continuous distribution, they were shaking (see Fig 3. (a)). That is because the  $G_i^*$  statistics of each time were calculated only according to the values in the corresponding time slice. Moreover, when the raw values of the sensors throughout the study area increase at the same time, the hotspots in the space-time contour plot will become scattered, and each of them will gradually shrink. That is inconsistent with the considerable increase observed in reality.

As for whether we should interpolate the missing values before the hotspot analysis, we found if the grids are located at the perimeter of the study site, the lack of neighbors easily resulted in the appearance of false hotspots at the place. That is because extrapolation is typically less accurate than interpolation.

Compared with control groups, spatial-temporal autocorrelation defines the adjacent space and time as spatial-temporal neighbors. This method was found to generate satisfactory calculation results, and the identified hotspot areas (e.g., Z-Scores > 1.6) continuous changes with the development of the events could also be presented clearly (see Fig 3. (b)). As a result, we chose Experimental group 1 to do the final analysis and discussion.

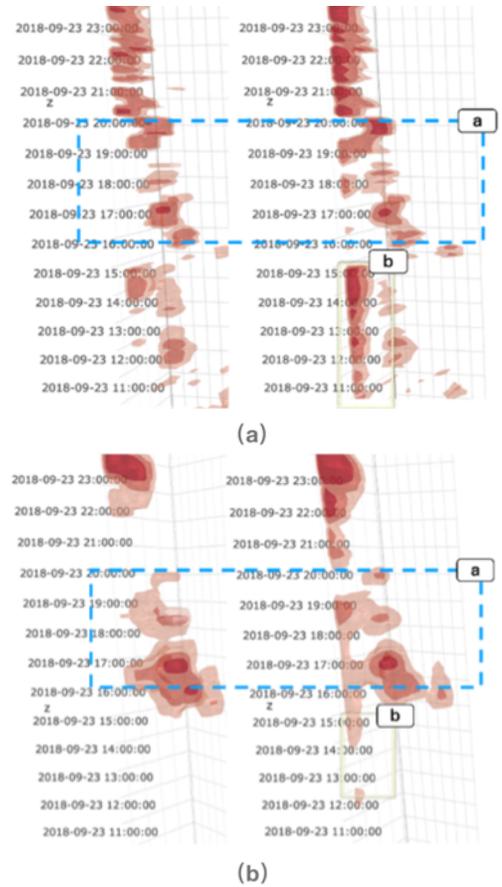


Figure 3. (a) Space-time contour plots for the control groups. (left: non-interpolation; right: interpolation) (b) Space-time contour plots for our work. (left: non-interpolation; right: interpolation)

Hotspot maps were illustrated based on the central points of pollution events and standard distance. Numerous pollution events were observed to be clustered and overlapping in the southeastern region of the study site. This implicated that major pollution sources were located in that region and that therefore further inspection was required (see Fig. 4).



Figure 4. Overlaying all hotspot maps of the study site.

To further demonstrate the potential of using this method to identify events, we first give each event some corresponding descriptive statistics, make simple chart plots, and try to explain the results. In Fig. 5 (a), each point denotes a cluster of individual pollution values. The x-axis indicates the pollution hotspot duration, and the y-axis indicates the standard distance of each pollution events, representing the level of pollution transmission in space. The color of each point indicates the maximum value of the  $G_i^*$  (Z-score) of the hotspots. In Fig. 5 (b), each point denotes a cluster of individual pollution values. The x-axis indicates the pollution hotspot duration. The y-axis indicates the moving distance, which is the distance between the first geometric center of the statistical unit serving as the hotspot and the last geometric center of the statistical unit that was identified as the hotspot, namely the amount of cluster movement. The color of each point indicates the maximum value of the  $G_i^*$  (Z-score) of the hotspots.

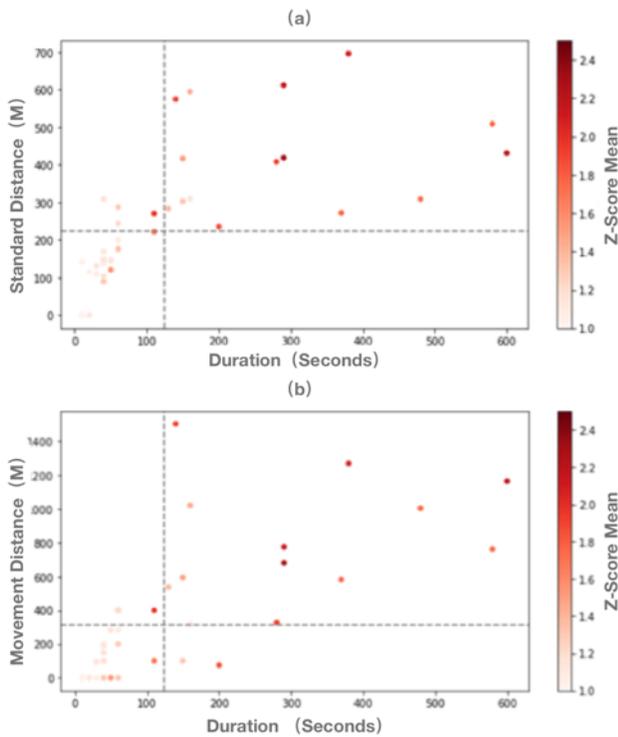


Figure 5. (a) Pollution hotspot duration versus range of the pollution event. (b) Pollution hotspot duration versus cluster movement distance.

In Fig. 5 (a), the data were divided into four quadrants as follows: The first quadrant denotes large-scale pollution events with long durations and large affected areas. Compared with small-scale pollution events, these large-scale pollution events had higher maximal  $G_i^*$  value; that is, the raw data were highly spatial-temporal autocorrelated. The second quadrant comprises pollution events with short durations but large affected areas. Wind direction and speed were inferred as the causes of pollution transmission within a region. The third quadrant contains local events

such as small-scale emission events, which featured short durations and low levels of transmission. The fourth quadrant indicates events featuring long pollution durations but short standard distances. This type of event was not included in our result.

In Fig. 5 (b), the data were divided into four quadrants as follows: The first quadrant represents events featuring long durations and vast moving distances of pollution cluster centers. In this study, this type of event had a relatively high maximum value of  $G_i^*$  (i.e., the sensor data were highly spatial-temporal autocorrelated). The second quadrant contains pollution events with short durations and large moving distances, which were possibly caused by large wind speeds. The third quadrant denotes instantaneous events featuring short pollution durations and small moving distances, including short-term emissions or equipment failure. The fourth quadrant comprises pollution events with long durations and short moving distances. This type of event had a relatively high  $G_i^*$  value, representing the continuously accumulating pollution.

#### IV. CONCLUSION

This study interpolated data points on a 3D plane, and the time dimension was considered to perform raw data interpolation and obtain experimental group 1 (see Fig 6.). When localized pollution became a pollution event for the entire industrial district, spatial-temporal  $G_i^*$  continued to increase, which is consistent with the distribution of the actual data. Our work demonstrates the feasibility of using spatial-temporal  $G_i^*$  to examine this type of data. The present study observed spatial  $G_i^*$  (3D spatial contour plot) and found that the hotspot of each time interval was discontinuous, thus generating the vibration phenomenon, possibly because the statistics for each time section were calculated separately. On the other hand, a 3D spatial-temporal contour plot was drawn to present the results, providing greater continuity to help researchers understand the development of pollution events.

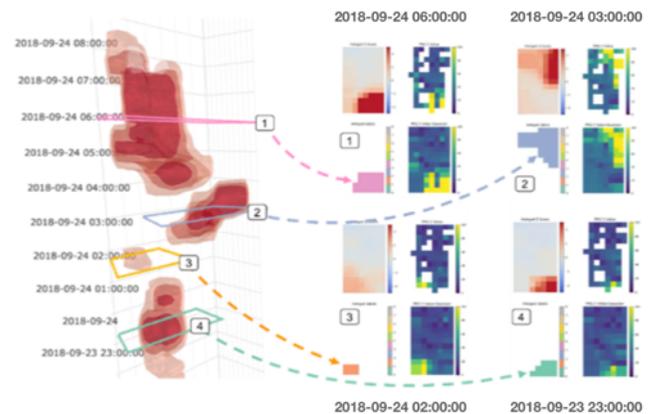


Figure 6. Identification of pollution events through spatial-temporal hotspot analysis integrated with visualized display method.

Moreover, the space-time kernel density was not used because (1) the focus was on clustered sensor data, rather

than the number of clustered sensors, and (2) previous studies have often used  $G_i^*$  to identify hotspots. Therefore, it was most critical to locate the spatial-temporal range.

Continuous transmission of pollutant data in industrial districts requires automatic spatial-temporal monitoring to provide instant warnings regarding excess pollution. In future studies, we will integrate the heterogeneous data of environmental dynamics as the basis for early signs regarding the dispersion of pollutants. Weather dynamics, especially wind-related information, should also be included. Research limitations of the current study included a lack of information about the wind patterns in the industrial district and a lack of building models. Thus, models were fixed by using wind direction and wind speed data from the monitoring station of the Central Weather Bureau to the industrial district.

#### ACKNOWLEDGMENT

This research has been supported by a grant from the Environmental Protection Administration, Executive Yuan, for the project of Smart Environmental Application Development.

#### REFERENCES

- [1] Battista, G., Pagliaroli, T., Mauri, L., Basilicata, C., and De Lieto Vollaro, R.: 'Assessment of the Air Pollution Level in the City of Rome (Italy)', *Sustainability*, 2016, 8, (9), pp. 838
- [2] Wang, J., and Ogawa, S.: 'Effects of meteorological conditions on PM<sub>2.5</sub> concentrations in Nagasaki, Japan', *International Journal of Environmental Research and Public Health*, 2015, 12, (8), pp. 9089-9101
- [3] 'Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe', 2008
- [4] Shen, C., Li, C., and Si, Y.: 'Spatio-temporal autocorrelation measures for nonstationary series: A new temporally detrended spatio-temporal Moran's index', *Physics Letters A*, 2016, 380, (1-2), pp. 106-116
- [5] Anselin, L.: 'Local indicators of spatial association—LISA', *Geographical analysis*, 1995, 27, (2), pp. 93-115
- [6] Getis, A., and Ord, J.K.: 'The analysis of spatial association by use of distance statistics', *Geographical analysis*, 1992, 24, (3), pp. 189-206
- [7] Gao, S., Zhu, R., and Mai, G.: 'Identifying Local Spatiotemporal Autocorrelation Patterns of Taxi Pick-ups and Dropoffs', in Editor (Ed.) (Eds.): 'Book Identifying Local Spatiotemporal Autocorrelation Patterns of Taxi Pick-ups and Dropoffs' (2016, edn.), pp. 109-113