Identifying Leading Nodes of PM_{2.5} Monitoring Network in Taiwan with Big Data-oriented Social Network Analysis

I-Cheng Chang*

Department of Environmental Engineering, National Ilan University, Yilan 26047, Taiwan

ABSTRACT

TEPA (Taiwan Environmental Protection Administration) currently has regulated six types of air pollutants based on the AQI. Among these, the three items most prone to exceeding the standard are PM_{2.5}, PM₁₀, and O₃, in that order. PM_{2.5} pollution episodes in Taiwan mainly occur in winter and spring when the northeast monsoon prevails. In addition to local pollution sources, transboundary air pollution affects Taiwan. Obviously, the existing AQ monitoring data analyzed by the BD-oriented perspective not only simplifies the simulation calculation and verification resources of the AQ model but also assists in real-time insight into the causal relationships between the AQ and important parameters of meteorology, pollution sources, and regions. This study integrates the BD-oriented Social Network Analysis (SNA) approach and data visualization tools to analyze the event co-occurrence and spatial correlation characteristics of two pollution scenarios for AQ monitoring stations based on two severe PM_{2.5} pollution conditions: (1) the Z-value of PM_{2.5} daily average concentration is higher than 1.65 (Scenario I), and (2) the daily average concentration of PM_{2.5} exceeds TEPA's regulation on the AQ standard (Scenario II), to identify the regional leading nodes suitable for different pollution scenarios. Furthermore, Principal Component Analysis (PCA) and time series data are employed to verify the spatial-temporal representation of these leading nodes, which can be regarded as means to the real-time AQ management decision-making as well as instant transboundary pollution precaution in the future. This study contributes to the application of the discrete data-driven approach (SNA) and the continuous data-driven approach (PCA) in an ambient AQ monitoring network, which can clearly explain and analyze the regional high pollution characteristics of PM_{2.5} in Scenarios I and II. The results of this study are consistent with previously relevant findings in Taiwan.

Keywords: PM_{2.5} pollution scenarios; Transboundary air pollution episodes; Social Network Analysis; Betweenness centrality.

INTRODUCTION

The health threats caused by PM_{2.5}, such as allergic rhinitis, asthma, severe chronic diseases, and premature death (Franklin *et al.*, 2007; Kloog *et al.*, 2013; Kim *et al.*, 2018), are critical issues around the world. Pope *et al.* (2002) observed that long-term exposure to PM_{2.5} increases the morbidity rates of general diseases, cardiopulmonary diseases, and lung cancer by 4%, 6%, and 8% respectively for every 10 µg of PM_{2.5}. Xie *et al.* (2015) analyzed the three-year (2010–2012) epidemiological data collected in the Beijing area and concluded that PM_{2.5} affects the morbidity and mortality of angina in Beijing. Hadei *et al.* (2017) applied a AQ model to assess potential public health problems resulted from exposure to air pollution during 2013 to 2016, and

presented that PM_{2.5} had the maximum short term health impacts of ambient air pollution in the Tehran area. Sahu and Kota (2017) found that the PM_{2.5} was the primary air pollutant in Indian capital during 2011 to 2014, and the daily exceeding ratios is around 85%. Zhao *et al.* (2018) indicated the Kaifeng area in China experienced the worst ambient AQ during winter of 2015 to 2017. Thus, to protect human health, in 2005 the WHO (2005) set the recommended annual and daily average standards for PM_{2.5} at 10 and 25 µg m⁻³, respectively. As in many countries, Taiwan's PM_{2.5} pollution level often exceeds WHO standards.

Episodes and scenarios related to air pollution are widely used as essential elements in formulating regulations on air pollution, AQ simulation, and air quality management (AQM). However, simulating secondary air pollutants and related topics requires great time and calculation resources to process the complex physical and chemical reactions involved. Further, the three-dimensional grid AQ simulation model entails high resource costs, making it unsuitable for determining an immediate pollution episode. Yet, it is obvious that above mentioned simulation approaches can be simplified through use of appropriate air pollution episodes.

* Corresponding author.

Tel.: +886-3-9357400 ext. 7586; Fax: +886-3-9367642
E-mail address: icchang@niu.edu.tw;
d2507002@gmail.com

Conventional screening of air pollution episodes and scenarios begins with classifying meteorological patterns and preparing relevant statistics. However, simply using the aforementioned meteorological classifications or AQ models is not feasible for instantly selecting appropriate and statistically representative pollution scenarios. As a result, some typical statistical approaches have factored in the screening of air pollution episodes using AQ monitoring data (Meyer *et al.*, 1997). Fang *et al.* (2017) analyzed concentration characteristics of PM_{2.5} during haze periods in the Changchun area in China, and found the Pearson correlation coefficient existed high values between PM_{2.5} and CO, that were affected by pollutant emissions and meteorological conditions; they also indicated the wind speed, temperature and pressure are low and the RH value is relatively high on haze days, in addition, stable weather during haze days makes the pollution heavier. Multivariate statistical analysis approaches, especially Principal Component Analysis (PCA) and Cluster Analysis (CA), have also been applied (Kuebler *et al.*, 2002; Beaver and Palazoglu, 2006). Zhang *et al.* (2014) used PCA and a nonparametric T2 control chart to predict episodes of over-standard ozone concentrations. Sun *et al.* (2015) predicted over-standard ozone episodes by adopting Generalized Linear Mixed Effects Models (GLMMs) and used BD-oriented and Machine Learning (ML) classification models with lower prediction errors as well.

Currently, data and BD-oriented approaches and their applications have drawn the attention of academic researchers and business professionals. However, a number of potentially useful values remain hidden in large-scale data. For instance, the large volumes of AQ monitoring data provided in official open data might reflect critical pollution patterns and relations. Though researchers have attempted to screen for representative pollution episodes or explore possible classifications of pollution patterns through AQ dataset of monitoring stations, they appear to be unable to identify stations (nodes) that are useful representative indicators of regional pollution patterns. A social network is defined as a set of social entities, such as people, organizations, or countries, with some pattern of relationship between them (Qing *et al.*, 2019; Oliveira and Gama, 2012). In social network analysis (SNA) graphs of the network are used to interpret patterns of social ties, where the nodes represent social entities and the lines represent social interactions between nodes. To provide AQ real-time warnings and evaluate the effectiveness of the AQ regional management strategy, it is necessary to identify leading AQ monitoring stations (nodes) which are representative of their air pollution regions. Thus, the aim of this study is to analyze the spatial-temporal characteristics of two severe PM_{2.5} pollution scenarios and identify the leading nodes affected by transboundary pollution in an ambient AQ monitoring network (TAQMN) deployed by TEPA through harnessing the BD-oriented SNA approach. In addition, the PCA approach is used to compare the results.

LITERATURE REVIEW

Multivariate Analysis on Air Pollution

In the field of AQM research, PCA approach is an

effective and objective data analysis tool that is often used to simplify variables (Dai *et al.*, 2015; Iodice *et al.*, 2016; Mari *et al.*, 2016; Yao *et al.*, 2016; Chen *et al.*, 2017), identify sources of pollution (Song *et al.*, 2006; Viana *et al.*, 2006; Shi *et al.*, 2009; Deka *et al.*, 2014; Huang *et al.*, 2015; Luo *et al.*, 2015; Gao *et al.*, 2016; Arhami *et al.*, 2017; Widiana *et al.*, 2017), classify meteorological patterns (Cheng and Lam, 2000), and evaluate model diagnostics (Dominick, 2012; Eder *et al.*, 2014; Li and Wen, 2014). To summary, the PCA has two advantages: (1) it simplifies the relevant variables of air pollution to achieve economic effectiveness; (2) it usually uses for screening pollution episodes as well as exhibits more objective statistical representativeness than the meteorological classification. For example, Yu (2013) and Yu *et al.* (2005, 2006) have adopted PCA to screen O₃ and PM₁₀ pollution episodes, and AQ meteorological characteristics of these episodes were also analyzed in Taiwan.

The BD-oriented Approach and Analysis on Air Pollution

Big data (BD) is a revolutionary phenomenon, widely explored in scientific and business research. In fact, several definitions for BD have been proposed, but attributive, comparative, and architectural definitions play a significant role in shaping how BD is viewed (Hu *et al.*, 2014; Pankaj *et al.*, 2015; De Mauro *et al.*, 2016; Makrufa and Aybeniz, 2016). In sum, regardless of the amount of data, any new understanding that can be obtained from data, and any new data values that can be created by useful technologies and tools may be regarded as being produced by a BD-oriented approach.

Lausch *et al.* (2015) suggested that the BD-oriented method of building air pollution receptor models can provide real-time predictions and analyses for a large amount of data and perform functions such as correlation (association) analysis, prediction, regression, segmentation (classification), and clustering (grouping). Vardoulakis *et al.* (2003) argued that important factors affecting the urban PM_{2.5} should include meteorology, traffic flow, personnel mobility, road network structure, and concerned receptor points. Therefore, the BD-oriented method may be used to simultaneously analyze the above factors. Zheng *et al.* (2013) adopted BD-oriented analysis techniques, such as Artificial Neural Network (ANN) and Conditional Random Forest (CRF), to combine the history of AQ stations with the real-time AQ data and then archived the parameters of important factors in PM_{2.5} levels previously mentioned by Vardoulakis *et al.* (2003) to estimate the real-time urban PM_{2.5}. Karner *et al.* (2010) collected roadside monitoring data from 37 research papers from 1983 to 2007, including the measurement values of more than 600 air pollutant concentration levels at different road distances, using the BD-oriented method to analyze the impact of roads to the air pollution concentration. Their results indicated that road distances ranging from 80 to 600 m can be attenuated as background values, and the concentration of pollutants at a road distance at 150 m dropped at least 50%.

Social Network Analysis

SNA aims to understand networks and their participants

and has two main focuses: the actors, who form the nodes, and their relationships in a specific social context (Cachia, 2008; Stevan *et al.*, 2019). Zhang and Peng (2017) indicated that SNA-related research is a hot topic and has been applied to substantive problems in many subjects and disciplines in recent years, in particular in the field of management science. SNA is not a formal theory in sociology. Instead, it is a strategy for investigating social structures through networks and graph theory. Typically, SNA is a concept and a method, studying social structures, organizational systems, interpersonal relationships, and group interactions. The SNA methodology accurately quantifies and analyzes complex relationships, revealing their structures and interpreting their phenomena using visual graphs (i.e., sociograms) and quantitative measures of social networks (Brass *et al.*, 2004; Charles, 2011; Grunspan *et al.*, 2014; Anthony, 2017). Furthermore, SNA can be regarded as a discrete data-driven analytical technology useful in BD-oriented approaches for examining hidden and valuable rules and knowledge through the interactive relationships of nodes and ties in a social network.

SNA centrality analysis is a family of concepts of characterizing the structural importance of a node's position in a network. The node-level centrality measure refers to the relationship between the observation node and its surroundings. Degree, closeness, and betweenness centrality are the most crucial and well-known centrality measures, widely used in SNA. Node-level centrality measures may be further standardized as dimensionless or percentage values to avoid being affected by the size of the network (Freeman *et al.*, 1991; Borgatti, 2005; Rong, 2013). According to Borgatti and Everett (2006), Freeman (2004), and Nieminen (1973, 1974), the definitions of the three well-known SNA centrality measures are excerpted and described as follows: (1) Degree centrality of a node is defined as the number of ties a node has (in graph-theoretical terminology, the number of edges adjacent to this node). It is used to calculate and observe the linking number of nodes. (2) The closeness centrality of a node is equal to the total distance (in the graph) of this node from all other nodes. Closeness centrality is used to measure the proximity of a certain node to others in the network. (3) Betweenness centrality may be defined loosely as the number of times a node needs to pass a node to reach another node. It is thus the number of shortest paths that pass through a given node. Evelien and Ronald (2002) and Faust (1997) deemed that betweenness centrality measures the ability of a node to broker the delivery of messages between any two other nodes.

In recent years, SNA has been widely applied in the fields of information management and social science for characteristics classification and identification (Manski, 2013; Hasan and Ukkusuri, 2014; Suthaharan, 2014; García-Palomares *et al.*, 2015; Amerini *et al.*, 2017; Kim and Hastak, 2018). Li *et al.* (2018) applied SNA to provide an up-to-date bibliometric view of current life cycle assessment (LCA). However, SNA has been less used in environmental management, especially in AQM research in Taiwan. For that reason, the application of SNA in this study represents an innovative research application in the field of AQM.

Since this study addresses the crucial characteristics of nodes in a social network, the betweenness centrality measure, which has been successfully applied in many disciplines (Gómez *et al.*, 2013; Wrzus *et al.*, 2013; Grunspan *et al.*, 2014; Tian *et al.*, 2019), plays a key role in this study. Accordingly, we use it to identify the leading nodes in our social network structure. In sum, betweenness centrality is a useful measure of the extent to which the network is dominated by a few central nodes.

METHODS

Research Procedure

The research processes and methods of this study are shown in Fig. 1. They are divided into six sub-phases sequentially based on a systematic CRISP-DM (Cross-industry standard process for data mining) cycle process of the BD-oriented methodology (Shafique *et al.*, 2014). The phases shown in Fig. 1 include: (1) Project understanding: what this study expects to gain from the data, including the objective, relevant references, appropriate methods, scope, and limitations of the research; (2) Data understanding: The data understanding phase of the CRISP-DM involves scrutinizing the data available for analysis. This step is critical in avoiding unexpected problems during the next phase, data preparation, which is typically the longest part of a study or project; (3) Data preparation: this phase is one of the most important and often time-consuming aspects of data analysis. In fact, it is estimated that data processing usually takes 50–70% of a project's time and efforts; (4) Modeling: Determining the most appropriate analytical methods, tools, or models for analyzing the data based on the previous phases of Project and Data understanding; (5) Evaluation: At this stage, the assessment of whether the project's outcomes meet the research requirements or expectations is formalized. This step requires a clear understanding of the stated research objectives and key points to be interpreted or verified in this phase; (6) Deployment: the process of using or discussing the new insights and findings to advance understanding and application in the field.

In the data understanding phase of this study, the information on the synaptic weather pattern and PM_{2.5} monitoring data are collected by ETL (Extract-Transform-Load) rules from the open data cloud service of TEPA. After data collection, it is necessary to prepare and process the data for further research. Data preparation involves data cleaning, summarizing, scaling, and normalizing the data. The processed data is subsequently imported into the SNA and PCA modeling and analysis phase. The results are validated and analyzed using data visualization tools. With these tools, paired nodes' relationships in a network through the number of connections (i.e., tie-strength) will be identified. Further, the map of betweenness centrality measures presented will be used to determine the characteristics of leading nodes for differentiation of pollution areas they cover or link. Furthermore, the duration lag and effect related to PM_{2.5} transboundary pollution is also identified by using the data visualization approach from a set of organized data.

Fig. 1. Diagram of research procedures and methods.

Data Understanding: Ambient AQ Monitoring Data

The AQ monitoring network in Taiwan was planned by TEPA in 1990 and officially launched running in September 1993. To date, there are 76 general AQ monitoring stations (Fig. 2), in which 63 general stations, 4 industrial stations, 2 national park stations (one of which is a general station), 4 background stations (two of which are general stations), and 6 traffic stations. Since August 2005, monitoring of PM_{2.5} pollutants has been added to various stations, in order to make observation and understand the characteristics of

PM_{2.5} pollutions. On May 14, 2012, the PM_{2.5} concentration threshold for the AQ standard was amended as an average of 35 $\mu\text{g m}^{-3}$ for 24 hours or a day. Concerning the aforementioned AQ monitoring data, currently TEPA has provided requesters with data by means of the OPENDATA method, in the way they can download and use the data by themselves. For that reason, this study collected the hourly monitoring data of PM_{2.5} from the aforementioned data source on the basis of the defined critical air pollutant and spatial-temporal scope as well.

Fig. 2. Locations of the 76 TEPA's monitoring stations.

Data Preparation: Data Definitions and Concentration Criteria for both $PM_{2.5}$ Pollution Scenarios

This study selected the concentration conditions and intentions of pollution episodes in the following two severe $PM_{2.5}$ pollution scenarios. (1) The daily average concentration of the $PM_{2.5}$ at any station with a high Z-value (> 1.65). This indicates that the concentration of $PM_{2.5}$ of the AQ monitoring station conforms to a high value of 95%. However, according to historical AQ monitoring data from Taiwan, this phenomenon is most likely to be driven by transboundary air pollution. (2) The daily average concentration of the $PM_{2.5}$ exceeding the standard ($35 \mu g m^{-3}$) regulated by TEPA. This indicates that managing levels exceeding the AQ standard has always been the objective of the authorities. Therefore, the number of days which exceed standard limits can be used as a performance evaluation indicator for an AQM strategy.

According to the temporal scope defined, the AQ monitoring data collected period was designated from 2015 to 2017, and the daily (24-hrs) average concentration of $PM_{2.5}$ of each AQ monitoring station was taken as the sample value. The data dimension of daily concentration of $PM_{2.5}$ could be treated as a collection of time series of 76 vectors (76 stations $\times$ 1096 days). Subsequently, these monitoring data were normalized by using the following equation (Eq. (1)):

$$Z_{ik} = \frac{C_{ik} - \mu_i}{S_i} \quad (1)$$

where Z_{ik} is the k th day Z-value of concentration of $PM_{2.5}$ at station i ; C_{ik} is the k th day concentration of $PM_{2.5}$ at station i ; μ_i is the average concentration of $PM_{2.5}$ at station i ; and S_i is the concentration standard deviation of $PM_{2.5}$ at station i .

The Z_{ik} is greater than 1.65 and C_{ik} exceeds $35 \mu g m^{-3}$ of the k th day were designated as concentration criterion for making definitions of scenario I and II of this study respectively.

Modeling and Evaluation: SNA Approach

The methodological procedure of SNA includes determining the cooperation among nodes and the analysis of hidden relationships (Podobnik and Lovrek, 2015; de-Marcos *et al.*, 2016; Erfanmanesh and Hosseini, 2016; Schlattmann, 2017). Therefore, the SNA methods in this study are summarized and explained as follows:

(1) Generate a 2-mode affiliated matrix with binary expression for each pollution scenario

Prior to performing the analysis of the SNA centrality measure and constructing the sociogram, a 2-mode affiliated matrix (i.e., sociomatrix) (1096 days $\times$ 76 stations) with binary expression should be constructed for each scenario in this study after the $PM_{2.5}$ concentration scale of both scenarios is confirmed. In the SNA, 2-mode data refers to data recording ties between two different sets of entities. The best known example of a 2-mode SNA is the study of class and race by Davis, Gardner, and Gardner (DGG) published

in their classic book *Deep South* (Davis, 1941). Table 1(a) presents a simplified example. We construct a 2-mode day-by-station SNA matrix X with 4 rows (days d1, d2, d3, and d4) and 5 columns (stations A, B, C, D, and E) in which element $x_{ij} = 1$ if a pollution episode tie was observed between day i and station j , and $x_{ij} = 0$ otherwise. For example, the last row has values 1, 0, 0, 0, 0 indicating that day 4 (d4) has ties only with the first station (A). By contrast, the third row of the table shows that day 3 (d3) has ties with stations B and E.

(2) Convert to a 1-mode quantitative square matrix for each scenario

Most social networks are conceived of as relations among a set of nodes, and are therefore represented as a 1-mode matrix or a simple graph or digraph. Typically, an SNA 2-mode (node-by-event) matrix (M) with binary expression can be converted to a 1-mode (node-by-node) quantitative (value) square matrix by matrix multiplication or product MM^T (Borgatti, 2009). Table 1(b) shows a simplified 1-mode (station-by-station) quantitative square matrix example in which element x_{ij} indicating the total number of episodes appears simultaneously in the 4 days between stations i and j . We convert it from Table 1(a) (a 2-mode day-by-station matrix X) by matrix multiplication $X^T X$. For example, the first row has the values 2, 0, 1, 0, 0 indicating that the total number of episodes in 4 days appearing at station A is 2. This station has one pollution episode appearing simultaneously in 4 days with station C only. The elements in an SNA 1-mode square matrix are used as tie-weights or tie-strength between any pair of nodes. Consequently, the value of each element in this 1-mode matrix can be regarded as the total number of days (i.e., tie-strength) of PM_{2.5} pollution episodes occurring simultaneously between paired AQ monitoring stations over the three years (2015 to 2017).

(3) Establish the SNA sociogram between nodes

Typically, the SNA sociogram is used to visually comprehend the tie relationships and topology of all nodes in an SNA network after an SNA 1-mode matrix is developed. As illustrated in Fig. 3, Table 1(b) is a simplified example demonstrating the construction of an SNA

sociogram from a 1-mode matrix. However, the diagonal elements of Table 1(b) are not taken into account while drawing an SNA sociogram. The digits labeled in Fig. 3 are tie-strengths or tie-weights between any pair of nodes, in short the tie-strength indicates the days of PM_{2.5} episode appear simultaneously at any pair of AQ monitoring stations(nodes) in this study.

(4) SNA betweenness centrality analysis

Liu *et al.* (2008) indicated that betweenness centrality is the sum of the shortest paths of a given node and is a measurement of a node's ability to mediate and control resources. Nodes with high betweenness centrality are key to knowledge transfer, influencing both the content and fluidity of information dissemination. High betweenness centrality indicates the node is a broker for the delivery of messages between other nodes in a SNA network.

According to Freeman (2004) and Borgatti *et al.* (2002) the node-level betweenness centrality measure of node i , denoted as $C_B(n_i)$, could be obtained from the mathematical expression in Eq. (2). It indicates the number of shortest paths (geodesic distance or steps) that pass through a given node. It is usually used to gauge the extent to which a node facilitates the flow of information in an SNA network.

$$C_B(n_i) = \sum_{j < k} \frac{g_{jk}(n_i)}{g_{jk}} \quad (2)$$

where g_{jk} is the number of shortest paths from node j to node k ; $g_{jk}(n_i)$ refers to the number of shortest paths from node j to node k passing through node i .

In addition, the standardized node-level betweenness centrality measure, denoted as $C_B^*(n_i)$, is to divide Eq. (2) by the maximum possible linking number of the node-level betweenness centrality in a g -node SNA network, as written in Eq. (3):

$$C_B^* = \frac{C_B(n_i)}{\frac{(g-1) \times (g-2)}{2}} = \frac{2 \times C_B(n_i)}{(g-1) \times (g-2)} \quad (3)$$

Table 1. Framework examples of a 2-mode afflicted matrix with binary expression and a 1-mode quantitative square matrix.

(a) 2-mode day-by-station SNA matrix (X)					
	A	B	C	D	E
d1	0	1	0	1	1
d2	1	0	1	0	1
d3	0	1	0	0	1
d4	1	0	0	0	0
(b) 1-mode station-by-station SNA matrix ($X^T X$)					
	A	B	C	D	E
A	2	0	1	0	1
B	0	2	0	1	2
C	1	0	1	0	1
D	0	1	0	1	1
E	1	2	1	1	3

Fig. 3. Schematic representation of building up a SNA sociogram.

Modeling and Evaluation: PCA Approach

According to Yu *et al.* (2006), the relationships between the standardized Z-value and the nonrotating principal component value are shown as following Eqs. (4) and (5):

$$Z_{ik} = \sum_{j=1}^n L_{ij} P_{jk} \quad (4)$$

where L_{ij} is the factor loading of the j th principal component of station i ; and P_{jk} is the component score of the k th variable (i.e., The k th AQ station) in the j th principal component.

$$P_{jk} = \sum_{i=1}^n (L_{ij} / \lambda_j) Z_{ik} \quad (5)$$

where λ_j is the eigenvalue of the j th principal component and also represents the variance of the j th principal component.

The principal components of these linear combinations exhibit not only the smallest variance in their components, but also the maximum individual differences of the individual components (Jolliffe and Cadima, 2016). Therefore, the first principal component can explain the maximum variation in the concentration field. The factor loading of PCA is not only used to represented the correlation between the principal component and each station, but also to comprehend the spatial characteristics of PM_{2.5} pollution episodes in this study.

Modeling and Evaluation: Data Visualization

In the field of BD, data visualization tools and technologies are essential in analyzing massive amounts of information to make data-driven decisions. A core skill in data science is the presentation of quantitative information and data in a graphical form by turning large and small datasets into visuals that are easier for the human brain to understand and process. Therefore, data visualization tools are used in this study to determine the relationships between nodes and implement the duration lag and effect related to transboundary pollution of PM_{2.5}.

RESULTS AND DISCUSSIONS

Evaluation and Deployment: SNA Approach

Based on the concentration conditions of Scenario I and the established SNA 1-mode square matrix for 76 of TEPA's

AQ monitoring stations (nodes), this study performed a SNA betweenness centrality analysis and then a sociogram (Fig. 4(a)) was drawn. As shown in Fig. 4(a), the offshore island stations ST_073 (Matsu) and ST_075 (Magong) in Scenario I are simultaneously connected to PM_{2.5} pollution episodes (Z-value > 1.65) on the main island of Taiwan. Consequently, according to results of aforementioned analysis, the essential and regional leading nodes of Scenario I are determined and summarized in Table 2 by further using visualization tools.

Based on the concentration conditions of Scenario II and the established SNA 1-mode square matrix of 76 AQ monitoring stations (nodes), similar to Scenario I, this study performed a SNA betweenness centrality analysis and a sociogram (Fig. 6(a)) was drawn. As shown in Fig. 6(a), station ST_074 (Kinmen) is connected to PM_{2.5} pollution episodes of Scenario II on the main island of Taiwan, in particular in the areas south of Taichung in the western part of Taiwan. Therefore, according to aforementioned analysis results, the essential and regional leading nodes of Scenario II are determined and summarized in Table 3.

Evaluation and Deployment: PCA Approach

PCA was performed based on the daily (24-hour) concentration of the 76 AQ monitoring stations in the PM_{2.5} monitoring project managed by TEPA in the defined temporary scope of this study (2015–2017). As shown in Table 4 and Fig. 7, among top five principal components, the explained variance of the fifth principal component was 1.66%, of which the maximum factor loading was 0.299, which belonged to a low correlation, however, it did not have any value of discussion. As a result, the principal component after the fifth principal component could be ignored, and there are four important principal components identified and factored in this study.

Additionally, ST_073 and ST_075 are generally regarded as AQ background stations for the main island of Taiwan. In the past, when transboundary air pollution occurred, the abovementioned AQ at these two stations often appeared high concentration pollution. The geographic locations of stations ST_073 and ST_075 are quite far from Taiwan's main island. The PCA analytical results of this study indicated that the PC1 factor loading values for stations ST_073 and ST_075 are about 0.47 while the PC2 are 0.47 and 0.56, respectively. It is shown that there is a leading correlation of transboundary air pollution among stations

Fig. 4. The sociogram of SNA betweenness centrality analysis and tie-strength contours of leading nodes for scenario I.

Table 2. Results and discussions to summarize for Scenario I.

Linking region	Description
<ul style="list-style-type: none"> Offshore islands and the northern region (the area north of Taoyuan city & Yilan county) 	<ul style="list-style-type: none"> This SNA tie-strength (the total number of PM_{2.5} pollution episodes (days) occurring simultaneously between paired stations in the three-year study period) contours for stations ST_073 and ST_075 in Scenario I, as shown in Figs. 4(b) and 4(c). The visualized Web-chart (Fig. 5(a)) indicates that stations ST_073 and ST_075 are strongly linked with AQ monitoring stations situated on the main island of Taiwan, such as stations ST_001, ST_015, ST_004, and ST_064. For that reason, station ST_073 is identified as an "offshore" leading node linking PM_{2.5} pollution of the main island of Taiwan (especially in northern Taiwan), while ST_075 is an alternative node. Fig. 4(a) shows that station ST_001 (Keelung) has higher SNA betweenness centrality in northern Taiwan and a stronger link with station ST_073. Furthermore, Figs. 4(a) and 5(a), and the tie-strength contour (Fig. 4(d)) indicate that station ST_001 can be identified as an "inland" leading node linking PM_{2.5} pollution in northern Taiwan. Stations ST_001, ST_015 and ST_004 could be grouped as an "inland" leading node group in this region.
<ul style="list-style-type: none"> The Hualien-Taitung region (the area south of Yilan county in the eastern part of the main island of Taiwan) 	<ul style="list-style-type: none"> Fig. 4(a) shows that station ST_064 (Yilan) has higher SNA betweenness centrality. The visualized Web-chart (Fig. 5(b)) indicates that station ST_064 is strongly linked to stations ST_001 and ST_004 in northern Taiwan, and stations ST_061, ST_062, ST_065, ST_076 in Hualien-Taitung region. In addition, Figs. 4(a) and 5(b), and the tie-strength contour (Fig. 4(e)) show that station ST_064 can be identified as an inland leading node linking PM_{2.5} pollution, and meanwhile stations ST_064 and ST_065 could be grouped as a leading node group in this region.
<ul style="list-style-type: none"> The central region (i.e., the area of Taichung-Changhua-Nantou and Hsinchu-Miaoli) 	<ul style="list-style-type: none"> Fig. 4(a) shows that station ST_034 (Xianxi) has the highest SAN betweenness centrality. The visualized Web-chart (Fig. 5(c)) indicates that station ST_034 is strongly linked to stations ST_029, ST_033, ST_022, and ST_067 in central region. In addition, Figs. 4(a) and 5(e), and the tie-strength contour (Fig. 4(f)) show that station ST_034 can be identified as an inland leading node linking PM_{2.5} pollution, and it is linked with stations ST_029 and ST_033 to form a leading node group in this region.
<ul style="list-style-type: none"> The Yunlin-Chiayi-Tainan region (the area of Yunlin county, Chiayi county and Tainan city) 	<ul style="list-style-type: none"> Fig. 4(a) shows that stations ST_038 (Lunbei) and ST_044 (Shanhua) have higher SNA betweenness centrality, while the visualized Web-charts (Figs. 5(c) and 5(d)) indicate that these two stations are strongly linked to several stations, including stations ST_040, ST_039, ST_034, ST_043, ST_072, ST_045, ST_042, ST_046, ST_043, and ST_037 in Yunlin-Chiayi-Tainan region. Furthermore, Figs. 4(a), 5(c) and 5(d), and the tie-strength contour (Fig. 4(g)) show that these two stations can be identified as inland leading nodes linking the PM_{2.5} pollution, and meanwhile they could form a leading node group with stations ST_040, ST_039, ST_045, and ST_042 in this region.
<ul style="list-style-type: none"> The Kaohsiung-Pingtung region (the area of Kaohsiung city and Pingtung county) 	<ul style="list-style-type: none"> Fig. 4(a) shows that station ST_051 (Daliao) has the higher SNA betweenness centrality. The Web-chart (Fig. 5(f)) indicates that station ST_051 is strongly linked to stations ST_054, ST_055, ST_050, ST_057, ST_052, and ST_070 in Kaohsiung-Pingtung region. In addition, Figs. 4(a), and 5(f), and the tie-strength contour (Fig. 4(h)) show station ST_051 can be identified as an inland leading node linking PM_{2.5} pollution, and it is linked with stations ST_054 and ST_055 to form a leading node group in this region.

Fig. 5. The Web-charts of leading nodes for scenario I.

Fig. 6. The sociogram of SNA betweenness centrality analysis and tie-strength contour of station ST_074 (Kinmen) for scenario II.

ST_073, ST_075 and the western part of the main island of Taiwan. Hence, this conclusion is consistent with previous SNA analytical results of this study.

Evaluation and Deployment: The Synoptic Weather Patterns of the Offshore Leading Nodes

Table 5(a) shows the proportion of the synoptic weather patterns of station ST_073 located at offshore of Taiwan under the PM_{2.5} pollution episodes of Scenario I (Z-value >

1.65) of the study. It indicates that the top four patterns are strong northeast monsoon (31.68%), standard northeast monsoon (22.36%), weak northeast monsoon (11.8%), and high pressure reflux (8.7%), respectively. In similar, the Table 5(b) indicates that the strong northeast monsoon (23.31%), high pressure reflux (14.66%), standard northeast monsoon (13.91%), and weak northeast monsoon (13.53%) top four the synoptic weather patterns of station ST_074 located at offshore of Taiwan under the PM_{2.5} pollution

Table 3. Results and discussions to summarize for Scenario II.

Linking region	Description
<ul style="list-style-type: none"> Offshore islands and the main island of Taiwan 	<ul style="list-style-type: none"> The similar analysis approaches of Scenario I are continued to be used in Scenario II, and Figs. 6(a) and 6(b) show that the offshore station ST_074 (Kinmen) has a stronger linkage with stations located in the areas south of Taichung in the western part of the main island of Taiwan, including stations ST_038, ST_054, ST_042, ST_072, ST_043, ST_044, and ST_068. For that reason, station ST_073 is identified as an offshore leading node linking PM_{2.5} pollution occurring simultaneously on the main island of Taiwan, especially the areas south of stations ST_030 (Dal) and ST_031 (Zhongming) in the western part of the main island of Taiwan.
<ul style="list-style-type: none"> The area south of Taichung in the western part of the main island of Taiwan (the central, Yunlin-Chiayi-Tainan, and Kaohsiung-Pingtung regions) 	<ul style="list-style-type: none"> Fig. 6(a) shows that stations ST_068 (Zhushan), ST_038 (Lunbei), ST_037 (Doului), and ST_054 (Zuoying) have much higher SNA betweenness centrality in Scenario II. Therefore, the similar visualization tools (tie-strength contour and Web-chart) are continued to be used for further determining the leading roles of stations ST_068 in the central (Taichung-Changhua-Nantou) region, ST_038 in the Yunlin-Chiay-Tainan region, and ST_054 in the Kaohsiung-Pingtung region.
<ul style="list-style-type: none"> The central region (the area of Taichung-Changhua-Nantou) 	<ul style="list-style-type: none"> According to the similar visualization tools used in Scenario I, station ST_068 is identified as an inland leading node linking PM_{2.5} pollution of Scenario II, and it is linked with its adjacent station ST_037 to form a leading node group in this region.
<ul style="list-style-type: none"> The Yunlin-Chiayi-Tainan region (the area of Yunlin county, Chiayi county and Tainan city) 	<ul style="list-style-type: none"> According to the similar visualization tools used in Scenario I, station ST_038 is identified as an inland leading node linking PM_{2.5} pollution of Scenario II, and it is linked with stations ST_072, ST_43, and ST_044 to form a leading node group in this region.
<ul style="list-style-type: none"> The Kaohsiung-Pingtung region (the area of Kaohsiung city and Pingtung county) 	<ul style="list-style-type: none"> According to the similar visualization tools used in Scenario I, station ST_054 is identified as an inland leading node linking PM_{2.5} pollution of Scenario II, and it is linked with stations ST_057, ST_049, and ST_70 to form a leading node group in this region.

Table 4. The results of PCA based on rotated principal components.

Principal Component	Eigenvalues	Explained variances (%)
PC1	25.74	33.86
PC2	20.86	27.46
PC3	13.97	18.38
PC4	4.13	5.44

Fig. 7. The loading factor contours of rotated principal components**Table 5.** The top five synoptic weather patterns of Scenarios I and II (Source: Soong et al., 2005).

(a) Synoptic weather patterns of Scenario I	Days	Episodes	Ratio (%)
Strong northeast monsoon	150	51	31.68
Standard northeast monsoon	142	36	22.36
Weak northeast monsoon	127	19	11.8
High pressure reflux	62	14	8.7
Off-shored high pressure I	23	10	6.21
(b) Synoptic weather patterns of Scenario II	Days	Episodes	Ratio(%)
Strong northeast monsoon	150	62	23.31
High pressure reflux	62	39	14.66
Standard northeast monsoon	142	37	13.91
Weak northeast monsoon	127	36	13.53
South flow	155	21	7.89

episodes of Scenario II (the daily average concentration exceeds $35 \mu\text{g m}^{-3}$) of the study. Figs. 8(a) and 8(b) show connections between the abovementioned four main weather patterns and seasons occurred in these two offshore leading nodes identified separately in Scenario I and Scenario II of this study by means of Web-chart (the comparison among tie-strength in a Web-chart is in a relative measure). Thus, the strong northeast monsoon, the standard northeast monsoon, the weak northeast monsoon, and the high pressure reflux are the top four common weather patterns in the two scenarios of $\text{PM}_{2.5}$ pollution episodes defined in this

study, often occurring in the winter and spring seasons when the northeast monsoon prevails.

Evaluation and Deployment: Analysis on the Duration Tag and Effect Related to Transboundary Air Pollution

To understand the impact of transboundary air pollution of $\text{PM}_{2.5}$ in Taiwan, this study adopted the external $\text{PM}_{2.5}$ pollution episode of October 29, 2017, as reported by TEPA, for further analysis. The synoptic weather pattern of that day was a strong northeast monsoon in winter. On the day (2017/10/29) at station ST_073 on an offshore island, the

Fig. 8. The Web-charts of weather conditions and seasons (Source: Soong *et al.*, 2005).

average hourly concentration of $\text{PM}_{2.5}$ at 13:00 in the afternoon was about $20 \mu\text{g m}^{-3}$, while the concentration was less than $30 \mu\text{g m}^{-3}$ between 14:00 and 17:00. However, it had been affected by transboundary air pollution since 18:00. Therefore, the average concentration began to increase and exceeded $30 \mu\text{g m}^{-3}$, reaching $35 \mu\text{g m}^{-3}$ at 19:00, and increased hourly to its highest value of approximately $91 \mu\text{g m}^{-3}$ at 03:00 the next day (2017/10/30). Then, it dropped to $21 \mu\text{g m}^{-3}$ at 16:00. As a result, the duration of the impact of the transboundary air pollution on ST_073 may be estimated at approximately 22 hours, from 18:00 on 2017/10/29 to 16:00 the following day (2017/10/30), with a maximum concentration of approximately $91 \mu\text{g m}^{-3}$. According to TEPA monitoring data, on 2017/10/29, station ST_003 (Wanli) on the main island of Taiwan nearest to Matsu (108km south of Matsu), the average hourly concentration of $\text{PM}_{2.5}$ was about $27 \mu\text{g m}^{-3}$ at 13:00 in the afternoon and less than $30 \mu\text{g m}^{-3}$ between 14:00 and 19:00. However, it began to be affected by transboundary air pollution at 20:00. The average concentration began to increase and exceeded $30 \mu\text{g m}^{-3}$, reaching $45 \mu\text{g m}^{-3}$

at 21:00, and climbed hourly to its highest value of about $73 \mu\text{g m}^{-3}$ at 23:00. It then fell hourly to $23 \mu\text{g m}^{-3}$ at 06:00 the following day (2017/10/30).

To interpret the data from stations in different regions of the main island of Taiwan and how they were affected by the aforementioned $\text{PM}_{2.5}$ transboundary pollution episodes, this study employed the $\text{PM}_{2.5}$ concentration contour of time (X-axis) and distance from the south of Matsu (Y-axis). As shown in Fig. 9, for station ST_003, it will be affected by the $\text{PM}_{2.5}$ transboundary pollution episodes about 2 hours later than station ST_073. Accordingly, we estimate that the areas from northern Taiwan to the central and Yunlin-Chiayi-Tainan regions (around 100 km, 200 km, 300 km away from Matsu, respectively) will be affected by the $\text{PM}_{2.5}$ transboundary pollution episodes approximate 2 to 6 hours later than Matsu. Because the Kaohsiung-Pingtung area in the south has a higher level of $\text{PM}_{2.5}$ concentration, the impact and duration lag of the transboundary pollution episodes are not easily identified. Based on these results, we estimate the degree (interval) of duration lag caused by the

Fig. 9. Analysis of the duration lag and effect related to $\text{PM}_{2.5}$ transboundary air pollution (Source: TEPA's AQ monitoring data).

PM_{2.5} transboundary pollution episodes at the leading nodes identified in different AQ management areas, in order to provide recommendations for generating real-time warnings in the future.

Evaluation and Deployment: A Short Summary

A short summary on results and discussions of this section is organized as follows.

(1) SNA

Scenario I: Figs. 9 and 10(a) show that station ST_073 (Matsu) is a useful offshore leading node linking to the PM_{2.5} pollution episodes of Scenario I in the northern region of the main island of Taiwan. Furthermore, the northern region is estimated to be affected by the PM_{2.5} transboundary pollution episodes approximate 2 hours later than Matsu in particular during northeast monsoon prevailing in the spring and winter. Station ST_001 (Keelung) is a feasible inland leading node in northern Taiwan, while station ST_064 (Yilan) is a leading node in the Hualien-Taitung region. Stations ST_034 (Xianxi) and ST_038 (Lunbei) are respectively identified as useful inland leading nodes in the central (Taichung-Changhua-Nantou and Hsinchu-Miaoli areas) and Yunlin-Chiayi-Tainan regions. In addition, both regions are estimated to be separately affected by the PM_{2.5} transboundary pollution episodes around 4 and 6 hours later than Matsu. Station ST_051 (Daliao) is a feasible leading node in the Kaohsiung-Pingtung region. Nevertheless, the possible interval of the duration lag of this region with the PM_{2.5} transboundary pollution episodes of station ST_073 is difficult to determine from Fig. 9.

Scenario II: Fig. 10(b) indicates that station ST_074 (Kinmen) is a useful offshore leading node linking PM_{2.5} pollution episodes of Scenario II in the areas south of stations ST_030 (Dali) and ST_031 (Zhongming) in the western Taiwan. Stations ST_068 (Zhushan), ST_038 (Lunbei), and ST_054 (Zuoying) are appropriate inland leading nodes in Taichung-Changhua-Nantou, Yunlin-Chiayi-Tainan, and Kaohsiung-Pingtung, respectively. However, the daily PM_{2.5} rarely exceeds TEPA's AQ standard value of 35 $\mu\text{g m}^{-3}$ in the areas north of stations ST_030 and ST_031 in the western and eastern part of the main island of Taiwan, none of the characteristics of the PM_{2.5} pollution episodes and leading nodes could be identified or screened in this Scenario. In addition, the number of PM_{2.5} pollution episodes exceeding the AQ regulatory standard value of offshore station ST_073 (approximately 14.7%) is lower than station ST_074 (approximately 24.3%) within the period of this study (2015–2017). In this case, offshore station ST_073 could not establish a stronger linkage of PM_{2.5} pollution episodes in this Scenario with most stations situated on the main island of Taiwan. Similarly, according to Fig. 9 stations ST_074, ST_068, ST_038, and other stations are located about 200km south of Matsu, if PM_{2.5} transboundary pollution episodes occur in station ST_073, it is estimated that they will affect the areas south of the central region in the western part of Taiwan about 4 hours to 6 hours later in particular during the northeast monsoon prevailing in spring and winter. Yet, station ST_054 in the Kaohsiung-Pingtung region faces

the same problem as in Scenario I. It is difficult to clearly determine the possible interval of the duration lag in the area affected by the PM_{2.5} transboundary pollution episodes occurring in station ST_073.

(2) PCA

The PCA results revealed that the high factor loadings of the first principal component (PC1) were concentrated in the area north of Hsinchu and Yilan counties, consistent with the areas linked by the leading node group in the northern region of SNA Scenario I (since few pollution episodes appeared in this region in Scenario II). The high factor loadings of the second principal component (PC2) were concentrated in the western area south of Chiayi county, consistent with the areas linked by the leading node groups of the Yunlin-Chiayi-Tainan and Kaohsiung-Pingtung regions in the SNA Scenario I and II. The high factor loadings of the third principal component (PC3) were concentrated in the areas of Taichung city and Changhua as well as Yunlin counties, consistent with the areas linked by the leading node group of the central region in SNA Scenario I. Except for the areas north of Miaoli county, PC3 and the areas linked by the leading node group of the central region in SNA Scenario II may be regarded as consistent. The high factor loadings of the fourth principal component (PC4) were concentrated in Hualien and Taitung counties, consistent with the areas linked by the leading node group of the Hualien-Taitung region in SNA Scenario I (since few pollution episodes appeared in this region in Scenario II). As a result, the SNA approach can be used in this study to analyze the results of these two scenarios, which can be reasonably matched with the analytical results of the PCA approach, especially in Scenario I.

(3) Summary

In sum, the analytical results of Scenario I indicate that the offshore leading node, station ST_073 (Matsu), links simultaneously with the PM_{2.5} pollution episodes of the northern region on the main island of Taiwan. Stations ST_001 (Keelung), ST_064 (Yilan), ST_034 (Xianxi), ST_038 (Lunbei), and ST_051 (Daliao) can be separately identified as inland leading nodes of PM_{2.5} pollution episodes situated at certain AQ zones of TAQMN in Scenario I. In addition, the analytical results of Scenario II show that the offshore leading node, station ST_074 (Kinmen), links simultaneously with the PM_{2.5} pollution episodes between the areas south of Taichung in the western part of Taiwan. Stations ST_071 (Zhushan), ST_038 (Lunbei), and ST_054 (Zuoying) are separately identified as inland leading nodes linking the PM_{2.5} pollution episodes of Scenario II occurring in the central and southern AQ zones of the TAQMN. The areas from northern Taiwan to the central and Yunlin-Chiayi-Tainan regions are estimated that will be affected by the PM_{2.5} transboundary pollution episodes approximate 2 to 6 hours later than Matsu.

CONCLUSIONS

This study adopted PM_{2.5} concentrations of TEPA's 76

Fig. 10. Locations of leading nodes and groups for scenario I and II.

ambient AQ monitoring stations between 2015 and 2017 and synoptic weather patterns as data sources. Meanwhile, the BD-oriented SNA and PCA methods were used to analyze two severe $PM_{2.5}$ pollution scenarios selected by this study. Scenario I represents a daily $PM_{2.5}$ concentrations in which the monitoring station reaches a high value, 95%, for three years. Since the concentration distribution and baseline

concentration for each station are different, the Z-values of the $PM_{2.5}$ concentration for some stations are higher than 1.65 but do not exceed TEPA's regulatory standard value ($35 \mu g m^{-3}$). Most of the $PM_{2.5}$ pollution episodes of Scenario I appear in the winter and spring, and transmit transboundary pollutants to Taiwan through strong or standard northeast monsoons. Therefore, Scenario I is a useful and meaningful

measure for zoning emergency responses to areas affected by transboundary air pollution and protection of residents' health. Scenario II is the daily $PM_{2.5}$ concentration higher than TEPA's regulatory standard. The reasons for its excess include not only transboundary air pollution episodes caused by strong and standard northeast monsoons but also severe air pollution episodes caused by nearby and local emission sources and unfavorable diffusion weather conditions. Consequently, Scenario II is useful and meaningful as a measure for zoning improvement for periods when air pollution exceeds the TEPA standard.

The BD-oriented and discrete data-driven SNA approach has been proven to be a reasonable and alternative means for determining $PM_{2.5}$ leading nodes and pollution regional classification in TAQMN by this study. For Scenarios I and II in the Yunlin-Chiayi-Tainan area, the $PM_{2.5}$ concentration level of station ST_038 (Lunbei) has a high leading association with the $PM_{2.5}$ pollution episodes of other nearby stations. Hence, it is a useful leading node for two different scenarios in the region. Stations ST_34 (Xianxi) and ST_68 (Zhushan) in Taichung-Changhua-Nantou region are respectively identified as inland leading nodes in Scenario I and II. Scenario I represents $PM_{2.5}$ pollution episodes in which transboundary pollutants reach Taiwan through the strong or standard northeast monsoon. Since there are no obvious fixed or mobile pollution sources close to station ST_34, in this case station ST_34 is a reasonable leading node for Scenario I. Scenario II represents a $PM_{2.5}$ concentration level higher than the AQ regulatory value. Since station ST_68 located at the site of major fixed pollution sources and leeward of mobile pollution sources in the central region of Taiwan, its $PM_{2.5}$ concentration easily exceeds the regulatory standard. For that reason, it is a rational leading node identified by Scenario II.

We adopted the continuous data-driven PCA method to analyze the daily $PM_{2.5}$ concentrations, in order to compare the SNA analytical results. Our findings showed that the PCA pollution regional classification were consistent with the regional classification of SNA in Scenario I and the area south of Taichung in Scenario II. Although the PCA method can explain the spatial characteristics of specific air pollution scenarios, it cannot comprehend the spatial-temporal leading characteristics among AQ monitoring stations. Nevertheless, the SNA method can identify leading nodes during the period when the contaminated areas are influenced by the $PM_{2.5}$ transboundary pollution episodes and help us understand the strength of the spatial-temporal correlation between AQ monitoring stations. Consequently, Big data-oriented SNA could supply effective and visualized information of leading nodes for identifying possible sources, distinguishing spatial correlation of monitoring stations, evaluating the effectiveness of control strategies on sources for distinct $PM_{2.5}$ pollution scenarios.

This study suggests other novel BD-oriented methods (such as deep learning for time series analysis) or visualization techniques can be introduced to understand the time series and spatial pollution characteristics of pollutants. In addition, this study suggests that appropriate scenarios (such as domestically contaminated scenarios or clean and

pollution-free scenarios) can be selected in the future for follow-up research and analysis to further comprehend and estimate the possible interval of duration lag in Kaohsiung-Pingtung region affected by the $PM_{2.5}$ transboundary pollution of Matsu.

ACKNOWLEDGMENTS

We express our gratitude to the Ministry of Science and Technology of Taiwan (MOST 107-2622-E-197-010 -CC3) for funding this study.

SUPPLEMENTARY MATERIAL

Supplementary data associated with this article can be found in the online version at <http://www.aaqr.org>.

REFERENCES

- Amerini, I., Caldelli, R., Del Mastio, A., Di Fuccia, A., Molinari, C. and Rizzo, A. P. (2017). Dealing with video source identification in social networks. *Signal Process. Image Commun.* 57: 1–7.
- Anthony, G. (2017). *Sociology* (eighth ed). Polity Press, Oxford, UK.
- Arhami, M., Hosseini, V., Shahne, M.Z., Bigdeli, M., Lai, A. and Schauer, J.J. (2017). Seasonal trends, chemical speciation and source apportionment of fine PM in Tehran. *Atmos. Environ.* 153: 70–82.
- Beaver, S. and Palazoğlu, A. (2006). A cluster aggregation scheme for ozone episode selection in the San Francisco, CA Bay Area. *Atmos. Environ.* 40: 713–725.
- Borgatti, S.P. (2005). Centrality and network flow. *Social Networks* 27: 55–71.
- Borgatti, S.P. (2009). Social Network Analysis, Two-Mode Concepts in. In *Computational Complexity*, Meyers R. (Ed.), Springer. New York, US.
- Borgatti, S.P., Everett, M., Freeman, L.C. (2002). *UCINET 6 for windows: Software for social network analysis USER'S GUIDE*. Analytic Technologies, Harvard, MA, US.
- Borgatti, S.P. and Everett M. (2006). A graph-theoretic perspective on centrality. *Social Networks*. 28: 466–484.
- Brass, D.J., Galaskiewicz, J., Greve, H.R. and Tsai, W. (2004). Taking stock of networks and organizations: A multilevel perspective. *Acad. Manage. J.* 47: 795–817.
- Cachia, R. (2008). *Study on the use and impact of online social networking*. Office for Official Publications of the European Communities, European Communities, Luxembourg.
- Charles, K. (2011). *Understanding social networks: Theories, concepts and findings*. Oxford University Press, Oxford, UK.
- Chen, Y., Li, X., Zhu, T., Han, Y. and Lv, D. (2017). $PM_{2.5}$ -bound PAHs in three indoor and one outdoor air in Beijing: Concentration, source and health risk assessment. *Sci. Total Environ.* 586: 255–264.
- Cheng, S. and Lam, K.C. (2000). Synoptic typing and its application to the assessment of climatic impact on

- concentrations of sulfur dioxide and nitrogen oxides in Hong Kong. *Atmos. Environ.* 34: 585–594.
- Dai, Q.L., Bi, X.H., Wu, J.H., Zhang, Y.F., Wang, J., Xu, H. and Feng, Y.C. (2015). Characterization and source identification of heavy metals in ambient PM₁₀ and PM_{2.5} in an integrated iron and steel industry zone compared with a background site. *Aerosol Air Qual. Res.* 15: 875–887.
- Davis, A., Gardner, B. and Gardner, M. (1941). *Deep south: A social anthropological study of caste and class*. University of Chicago Press, Chicago, US.
- De Mauro, A., Andrea, G., Greco, M., and Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Libr. Rev.* 65: 122–135.
- Deka, P. and Hoque, R.R. (2014). Diwali fireworks: Early signs of impact on PM₁₀ properties of rural Brahmaputra valley. *Aerosol Air Qual. Res.* 14: 1752–1762.
- de-Marcos, L., García-López, E., García-Cabot A., Medina-Merodio, J., Domínguez, A., Martínez-Herráiz J. and Díez-Folledo, T. (2016). Social network analysis of agamified e-learning course: Small-world phenomenon and network metrics as predictors of academic performance. *Comp. Hum. Behav.* 60: 312–321.
- Dominick, D., Juahir, H., Latif, M.T., Zain, S.M. and Aris, A.Z. (2012). Spatial assessment of air quality patterns in Malaysia using multivariate analysis. *Atmos. Environ.* 60: 172–181.
- Eder, B., Bash, J., Foley, K. and Pleim, J. (2014). Incorporating principal component analysis into air quality model evaluation. *Atmos. Environ.* 82: 307–315.
- Erfanmanesh, M. and Hosseini, E. (2016). Using social network analysis method to visualize library & information science research. *J. Adv. Inf. Technol.* 7: 177–182.
- Evelien, O. and Ronald R. (2002). Social network analysis: A powerful strategy, also for the information sciences. *J. Inf. Sci.* 28: 441–453.
- Fang, C., Zhang, Z., Jin, M., Zou, P. and Wang, J. (2017). Pollution characteristics of PM_{2.5} aerosol during haze periods in Changchun, China. *Aerosol Air Qual. Res.* 17: 888–895.
- Faust, K. (1997). Centrality in affiliation networks. *Social Network* 19: 157–191.
- Franklin, M., Zeka, A., and Schwartz, J. (2007). Association between PM_{2.5} and all-cause and specific-cause mortality in 27 US communities. *J. Exposure Sci. Environ. Epidemiol.* 17: 279–287.
- Freeman, L.C., Borgatti, S.P., and Douglas, R.W. (1991). Centrality in valued graphs: A measure of betweenness based on network flow. *Social Networks* 13: 141–154.
- Freeman, L.C. (2004). *The development of social network analysis: A study in the sociology of science*. Empirical Press, Vancouver, BC, Canada.
- Gao, J., Peng, X., Chen, G., Xu, J., Shi, G.L., Zhang, Y.C. and Feng, Y.C. (2016). Insights into the chemical characterization and sources of PM_{2.5} in Beijing at a 1-h time resolution. *Sci. Total Environ.* 542: 162–171.
- García-Palomares, J.C., Gutiérrez, J., and Mínguez, C. (2015). Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. *Appl. Geogr.* 63: 408–417.
- Gómez, D., Figueira, J.R., and Eusébio, A. (2013). Modeling centrality measures in social network analysis using bi-criteria network flow optimization problems. *Eur. J. Oper. Res.* 226: 354–365.
- Grunspan, D.Z., Wiggins, B.L. and Goodreau, S.M. (2014). Understanding classrooms through social network analysis: A primer for social network analysis in education research. *Life Sci. Educ.* 13: 159–359.
- Hadei, M., Hopke, P.K., Nazari, S. S.H., Yarahmadi, M., Shahsavani, A. and Alipour, M.R. (2017). Estimation of mortality and hospital admissions attributed to criteria air pollutants in Tehran Metropolis, Iran (2013–2016). *Aerosol Air Qual. Res.* 17: 2474–2481.
- Hasan, S. and Ukkusuri, S.V. (2014). Urban activity pattern classification using topic models from online geo-location data. *Transp. Res. C* 44: 363–381.
- Hu, H. and Wen, Y., Chua, T.S. and Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. *IEEE Access* 2: 652–689.
- Huang, Y., Li, T., Wu, C., He, Z., Japenga, J., Deng, M. and Yang, X. (2015). An integrated approach to assess heavy metal source apportionment in peri-urban agricultural soils. *J. Hazard. Mater.* 299: 540–549.
- Iodice, P., Adamo, P., Capozzi, F., Di Palma, A., Senatore, A., Spagnuolo, V. and Giordano, S. (2016). Air pollution monitoring using emission inventories combined with the moss bag approach. *Sci. Total Environ.* 541: 1410–1419.
- Jolliffe, I.T. and Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philos. Trans. A Math. Phys. Eng. Sci.* 374:20150202.
- Karner, A.A., Eisinger, D.S. and Niemeier, D.A. (2010). Near-roadway air quality: synthesizing the findings from real-world data. *Environ. Sci. Technol.* 44: 5334–5344.
- Kim, J. and Hastak, M. (2018). Social network analysis: Characteristics of online social networks after a disaster. *Int. J. Inf. Manage.* 38: 86–96.
- Kim, T.Y., Kim, H., Yi, S.M., Cheong, J.P. and Heo, J. (2018). Short-term effects of ambient PM_{2.5} and PM_{2.5-10} on mortality in major cities of Korea. *Aerosol Air Qual. Res.* 18: 1853–1862.
- Kloog, I., Ridgway, B., Koutrakis, P., Coull, B.A. and Schwartz, J.D. (2013). Long- and short-term exposure to PM_{2.5} and mortality: Using novel exposure models. *Epidemiology* 24: 555–561.
- Kuebler, J., Russell, A.G., Hakami, A., Clappier, A. and Van den Bergh, H. (2002). Episode selection for ozone modelling and control strategies analysis on the Swiss Plateau. *Atmos. Environ.* 36: 2817–2830.
- Lausch, A., Schmidt, A. and Tischendorf, L. (2015). Data mining and linked open data—New perspectives for data analysis in environmental research. *Ecol. Model.* 295: 5–17.
- Li, J., Wang, Y. and Yan, B. (2018). The hotspots of life cycle assessment for bioenergy: A review by social network analysis. *Sci. Total Environ.* 625: 1301–1308.
- Li, S. and Wen, J. (2014). A model-based fault detection and diagnostic methodology based on PCA method and

- wavelet transform. *Energy Build.* 68: 63–71.
- Liu, Z., Chen, Y., Hou, H.Y., Ying, L.C., Yang, Z.K., and Liang, Y.X. (2008). *Mapping knowledge domains methods and application*. People's Publishing House, Beijing, China.
- Luo, X.S., Xue, Y., Wang, Y.L., Cang, L., Xu, B. and Ding, J. (2015). Source identification and apportionment of heavy metals in urban soil profiles. *Chemosphere* 127: 152–157.
- Makrufa, H. and Aybeniz A. (2016). Some indicators of Big Data. *IOSR J. Eng. (IOSR-JEN)*. 6: 2278–8719.
- Manski, C.F. (2013). Identification of treatment response with social interactions. *Econom. J.* 16: S1–S23.
- Mari, M., Sánchez-Soberón, F., Audí-Miró, C., van Drooge, B.L., Soler, A., Grimalt, J.O. and Schuhmacher, M. (2016). Source apportionment of inorganic and organic PM in the ambient air around a cement plant: Assessment of complementary tools. *Aerosol Air Qual. Res.* 16: 3230–3242.
- Meyer, E.L., Baldrige, K.W., Chu, S. and Cox, W.M. (1997). Choice of episodes to model: Considering effects of control strategies on ranked severity of prospective episodes days. Air & Waste Management Association's 90th Annual Meeting & Exhibition 97-MP112.01.
- Niemenen, J. (1973). On the centrality in a directed graph. *Soc. Sci. Res.* 2: 371–378.
- Niemenen, J. (1974). On the centrality in a graph. *Scand. J. Psychol.* 15: 332–336.
- Oliveira, M. and Gama, J. (2012). An overview of social network analysis. *WIREs Data Min. Knowl. Discovery* 2: 99–115.
- Pankaj, D.K., Amneet, K. and Sandeep, K. (2015). Performance analysis in big data. *Int. J. Inf. Technol. Comput. Sci. (IJITCS)* 7: 55–61.
- Podobnik V. and Lovrek, I. (2015). Implicit social networking: Discovery of hidden relationships, roles and communities among consumers. *J. Procedia Comput. Sci.* 60: 583–592.
- Pope, III, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K. and Thurston, G.D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA* 287: 1132–1141.
- Qing, C., Chao, L., Mark, A.H. and Zhiguang, D. (2019). Research characteristics on health law in China: Social network analysis. *J. Acad. Librariansh.* 45: 126–136.
- Rong, T.S. (2013). *Application of UCINET in Social Network Analysis (SNA)*. Wu-nan Book Inc., Taipei, Taiwan.
- Sahu, S.K. and Kota, S.H. (2017). Significance of PM_{2.5} air quality at the Indian capital. *Aerosol Air Qual. Res.* 17: 588–597.
- Schlattmann, S. (2017). Capturing the collaboration intensity of research institutions using social network analysis. *Procedia Comput. Sci.* 106: 25–31.
- Shafique, U. and Haseeb, Q. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *Int. J. Innovation Sci. Res.* 12: 217–222.
- Shi, G.L., Feng, Y.C., Wu, J.H., Li, X., Wang, Y.Q., Xue, Y.H. and Zhu, T. (2009). Source identification of polycyclic aromatic hydrocarbons in urban particulate matter of Tangshan, China. *Aerosol Air Qual. Res.* 9: 309–315.
- Song, Y., Xie, S., Zhang, Y., Zeng, L., Salmon, L.G., and Zheng, M. (2006). Source apportionment of PM_{2.5} in Beijing using principal component analysis/absolute principal component scores and UNMIX. *Sci. Total Environ.* 372: 278–286.
- Soong, W.K., Hung, C.H., Yuan, C.S. and Yang, H.Y. (2005). Effects of local atmospheric circulation on the formation of high ozone concentration in confined area—A case study in southern Taiwan. *J. Chin. Inst. Environ. Eng.* 15: 163–172.
- Stevan, M., Zorica, B., Aleksandra, L., and Dušan, B. (2019). An approach to identify user preferences based on social network analysis. *Future Gener. Comput. Syst.* 93: 121–129.
- Sun, W., Palazoglu, A., Singh, A., Zhang, H., Wang, Q., Zhao, Z. and Cao, D. (2015). Prediction of surface ozone episodes using clusters based generalized linear mixed effects models in Houston–Galveston–Brazoria area, Texas. *Atmos. Pollut. Res.* 6: 245–253.
- Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Perform. Eval. Rev.* 41: 70–73.
- Tian, Z.P., Nie, R.X. and Wang, J.Q. (2019). Social network analysis-based consensus-supporting framework for large-scale group decision-making with incomplete interval type-2 fuzzy information. *Inf. Sci.* 502: 446–471.
- U.S. Environmental Protection Agency (U.S. EPA) (1985). User's Manual, EPA-600/8-85/007a,b. U.S. Environmental Protection Agency, Atmospheric Sciences, Research Laboratory, Research Triangle Park, NC, US.
- Vardoulakis, S., Fisher, B. E., Pericleous, K., and Gonzalez-Flesca, N. (2003). Modelling air quality in street canyons: a review. *Atmos. Environ.* 37: 155–182.
- Viana, M., Querol, X., Alastuey, A., Gil, J.I. and Menéndez, M. (2006). Identification of PM sources by principal component analysis (PCA) coupled with wind direction data. *Chemosphere* 65: 2411–2418.
- WHO (2005). *Ambient (outdoor) air quality and health*, World Health Organization (WHO), [https://www.who.int/en/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/en/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health), Last Access: 2018.
- Widiana, D.R., You, S.J., Yang, H.H., Tsai, J.H. and Wang, Y.F. (2017). Source apportionment of air pollution and characteristics of volatile organic compounds in a municipal wastewater treatment plant, North Taiwan. *Aerosol Air Qual. Res.* 17: 2878–2890.
- Wrzus, C., Hänel, M., Wagner, J. and Neyer, F.J. (2013). Social network changes and life events across the life span: a meta-analysis. *Psychol. Bull.* 139: 53–80.
- Xie, W., Li, G., Zhao, D., Xie, X., Wei, Z., Wang, W., Wang, M., Li, G., Liu, W., Sun, J., Jia, Z., Zhang, Q. and Liu, J. (2015). Relationship between fine particulate air pollution and ischaemic heart disease morbidity and mortality. *Heart* 101: 257–263.
- Yao, L., Yang, L., Yuan, Q., Yan, C., Dong, C., Meng, C. and Wang, W. (2016). Sources apportionment of PM_{2.5} in

- a background site in the North China Plain. *Sci. Total Environ.* 541: 590–598.
- Yu, T.Y., and Chang, L.F.W. (2005). Screening air pollution episodes by principal component analysis. *J. Chin. Inst. Environ. Eng.* 15: 57–64.
- Yu, T.Y. and Chang, I.C. (2006). Spatiotemporal features of severe air pollution in northern Taiwan. *Environ. Sci. Pollut. Res. Int.* 13: 268–275.
- Yu, T.Y. (2013). Identification of source regions of PM₁₀ with backward trajectory-based statistical models during PM₁₀ episodes. *Environ. Monit. Assess.* 185: 6465–6475.
- Zhang, H., Palazoglu, A., Zhang, X., Zhang, W., Zhao, Z., Sun, W. and Liu, S. (2014). Prediction of surface ozone exceedance days using PCA with a non-parametric T2 control limit. *Chemom. Intell. Lab. Syst.* 133: 42–48.
- Zhang, J. and Peng, D. (2017). A study of centrality measures in SNA. 2017 International Conference on Advanced Education and Management Science (AEMS 2017), ISBN: 978-1-60595-438-7, pp. 268–273.
- Zhao, R., Cui, K.P., Wang, W.W. and Chao, H. R. (2018). Characterization of air quality index for both Handan and Kaifeng cities, China. *Aerosol Air Qual. Res.* 18: 2086–2107.
- Zheng, Y., Liu, F. and Hsieh, H.P. (2013). U-air: When urban air quality inference meets big data. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (ACM 2013), pp. 1436–1444.

Received for review, November 2, 2019

Revised, November 23, 2019

Accepted, November 24, 2019