Particulate Matter Monitoring and Forecast with Integrated Low-cost Sensor Networks and Air-quality Monitoring Stations

Huynh A. D. Nguyen[1,*], *Trung H.* Le[1], *Quang P.* Ha[1], *Hiep* Duc[2], and *Merched* Azzi[2]

[1]Faculty of Engineering and IT, University of Technology Sydney, Ultimo NSW 2007, Australia.
[2]Department of Planning and Environment of New South Wales, Lidcombe NSW 2141, Australia.

Abstract.
The fusion of low-cost sensor networks with air quality stations has become prominent, offering a cost-effective approach to gathering fine-scaled spatial data. However, effective integration of diverse data sources while maintaining reliable information remains challenging. This paper presents an extended clustering method based on the Girvan-Newman algorithm to identify spatially correlated clusters of sensors and nearby observatories. The proposed approach enables localized monitoring within each cluster by partitioning the network into communities, optimizing resource allocation and reducing redundancy. Through our simulations with real-world data collected from the state-run air quality monitoring stations and the low-cost sensor network in Sydney's suburbs, we demonstrate the effectiveness of this approach in enhancing localized monitoring compared to other clustering methods, namely K-Means Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Agglomerative Clustering. Experimental results illustrate the potential for this method to facilitate comprehensive and high-resolution air quality monitoring systems, advocating the advantages of integrating low-cost sensor networks with conventional monitoring infrastructure.

Keywords: particulate matter, monitoring, clustering, low-cost sensors, air-quality stations.

1 Introduction

Particulate matter (i.e., $PM_{1.0}$, $PM_{2.5}$, and PM_{10}) is associated with severe health problems even at low concentration levels [1]. An analysis conducted by the World Health Organization (WHO) indicates that reducing current air pollution levels to below the standards could potentially prevent nearly 80% of deaths worldwide related to $PM_{2.5}$ [2]. Consequently, monitoring real-time air pollution concentrations, especially fine particles in both regional and local areas, is imperative for effective air quality control and sustainable development in smart cities.

In Australia, the state-run air-quality monitoring stations (AQMSs) provide high-quality observations at sparse locations [3]. This limitation constrains the precise monitoring and analysis of microclimatic conditions and local-scale air pollution emissions. The advent of Internet-of-Thing (IoT) technology, dominated by low-cost wireless sensor networks (LWSNs), has recently offered a promising solution for localized observations and contributed to enhancing our ability to assess the potential risks associated with air pollution [4]. For instance, the AirU pollution monitor network was designed and deployed in street-level locations in Salt Lake City, Utah, USA to evaluate the trapping pollution on the valley floor [5]. A dependable wireless sensing framework was proposed to enhance the reliability of an LWSN for local monitoring of dust emission from construction sites at Melrose Park, in the state of New South Wales (NSW), Australia [4].

In air quality monitoring, IoT-enabled sensor networks often encounter difficulties related to the reliability, accuracy, and sustainability of the sensory data due to the volatility of the operational environment. These issues arise from various error sources, including signal interference, cross-sensitivity, missing values, and measurement drift [6]. Consequently, substantial efforts are needed to process data obtained from these affordable sensors, involving tasks such as noise and outlier removal, as well as data imputation to recover missing information. Therefore, there is a growing interest in improving the monitoring and forecasting of air particle levels across various regional and local scales through the integration of data from wireless sensors and regular observatories. For example, an IoT-enabled sensor network has been deployed and integrated with nearby AQMSs to augment the capacity of monitoring and forecasting local emissions from construction activities [7].

Data fusion from both AQMSs and LWSNs requires an efficient clustering method to capture the distinctive local characteristics of air pollution [6]. Current clustering approaches applied to data analysis and predictions of air pollution include K-Means clustering [8], Hierarchical Clustering [9], Density-Based Spatial Clustering of Applications with Noise

*E-mail: huynhanhduy.nguyen@uts.edu.au

(DBSCAN) [10], and Agglomerative Clustering [11]. Each of them has its own strengths and limitations which may either affect the representation of spatiotemporal features or incur further latency due to redundancy in air-quality data obtained from diverse sources. For example, K-Means' performance can be affected by the initial placement of cluster centroids, DBSCAN encounters issues dealing with data featuring varying densities, while Hierarchical and Agglomerative clustering methods tend to be less efficient when applied to extensive datasets and facing difficulty in determining the optimal number of clusters [9].

In LWSNs, each sensor can be seen as a node within a network, and the interactions between sensors form the edges of a graph, providing a structured way to analyze the complex relationships between the sensors and their centroid in a cluster. As such, the community structure, a technique proposed for social and biological networks for detecting communities having representative nodes that can be joined together by using centrality indices to find community boundaries [12], could be useful to identify groups of sensors exhibiting similar behaviors or common features within the network. Motivated by the capability of extending the Girvan-Newman structure to large-scale networks [13, 14], here we propose a new formula for calculating the edge betweenness centrality based on distances and correlations between low-cost sensors and AQMSs for air quality data fusion. The contributions of this study are as follows:

- A correlation-based Girvan-Newman (CGN) algorithm proposed for enhancing the clustering capacity by incorporating spatiotemporal intercorrelations of LWSNs and AQMSs.
- Comprehensive comparison conducted for experimental validation of CGN with respect to widely-recognized clustering methods such as K-Means, DBSCAN, and Agglomerative Clustering, using the real-world datasets obtained from 14 regular monitoring stations and 47 air-quality sensors located throughout the Sydney basin.

2 Monitoring stations and low-cost wireless sensor networks

2.1 Air-quality monitoring stations

In NSW, there are over 50 air-quality monitoring stations (AQMS) operated by the Department of Planning and Environment (NSW-DPE). These stations are strategically located to provide representative data including the six types of pollutants (i.e., $PM_{2.5}$, PM_{10}, O_3, NO, NO_2, CO, SO_2 and NH_3) along with visibility and meteorological variables (i.e., wind speed, wind direction, air temperature, relative humidity and rainfall) [3]. This study will focus on stations located in the Sydney suburbs with a high population density.

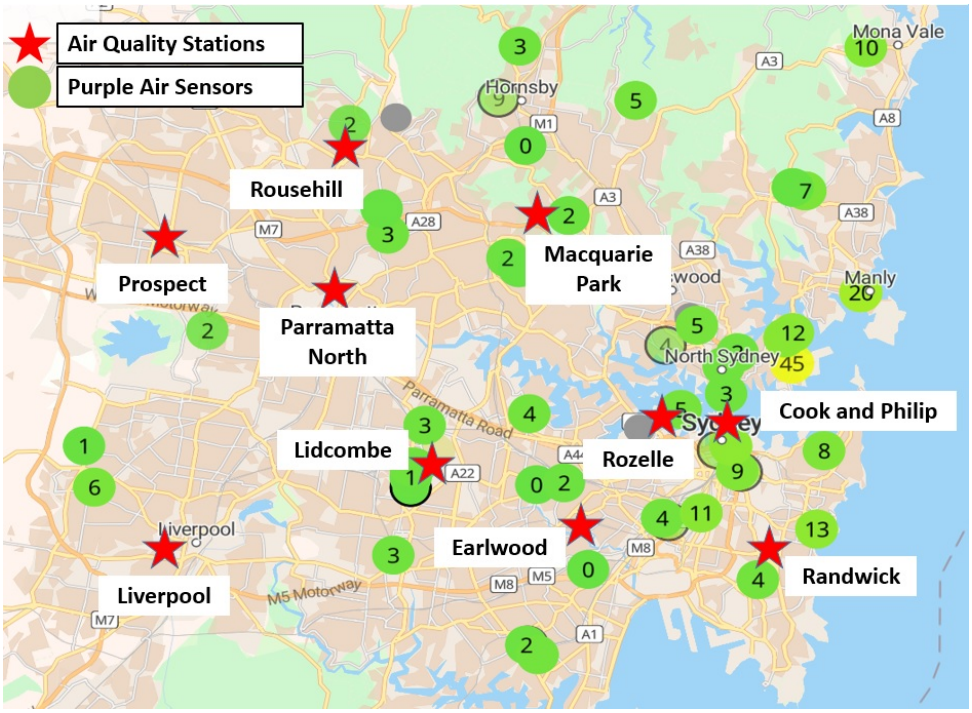


Figure 1. Spatial locations of monitoring stations (red stars) and low-cost sensors (green circles) in Sydney basin

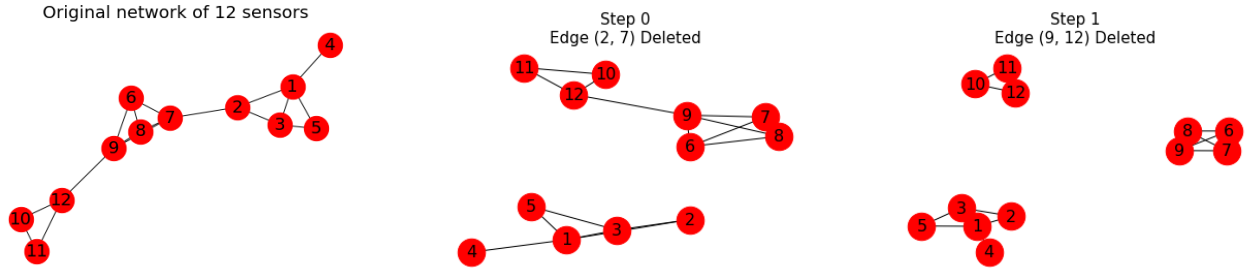
Figure 1 illustrates the air quality map, indicating a higher concentration of AQMSs in the central and eastern regions of Sydney, which correspond with the more densely populated areas. Conversely, fewer stations are located in the western suburbs, including North-West and South-West Sydney. This discrepancy results in a lack of local air pollution data for these suburban areas, affecting the accuracy of analyses and forecasting models.

2.2 Low-cost wireless sensors networks in NSW

The NSW authority has calibrated and deployed several types of low-cost sensors, namely Purple Air, KOALA, and Luftdaten, over multiple suburbs of Sydney [15]. Among these devices, Purple Air sensors (PASs) are evaluated as highly reliable, easy to deploy, maintain and repair [15]. Additionally, the original equipment manufacturer of PASs offers a straightforward online interface (*https://map.purpleair.com*) for users to access and retrieve data easily. In comparison with high-tech instruments (i.e., Aurora 1000 nephelometers), PASs have high coefficients of correlation of 0.83-0.94 and can measure variant episodes of high concentrations [15]. Currently in the Sydney basin, a network of 75 Purple Air sensors is distributed across various locations covering nearly 5000 km^2, spanning from the west at Katoomba (33.7124°S, 150.3118°E) to the east at Avalon Beach (33.6333°S, 151.3312°E), and from the topmost point at Brooklyn (33.5475°S, 151.2191°E) to the southernmost point at Campbelltown (34.0666°S, 150.8196°E). Among them, 66 are placed outdoors, while the remaining nine are installed indoors. These 66 monitoring sensors cover not only in the east but also in remote suburbs in the west of Sydney as shown in Figure 1.

3 Community detection method for clustering sensors with AQMSs

The community detection methods are popular in the network of sensors, particularly AQMSs, to explore the functional clusters of sensors and reveal patterns in data collection and information flow [14]. At first, we formulate a representative graph to link all sensors (vertices) together based on their relationships (edges) of distances and correlations of observed data. Then, the network is gradually partitioned into communities with an optimized algorithm. The Girvan-Newman algorithm is widely used to solve the problem of community detection within complex networks [13]. In principle, this algorithm iteratively removes the connected edges with the highest betweenness centrality as illustrated in Figure 2. These removed edges connect different communities and play a crucial role in holding correlations between the adjacent communities [16].



(a) Original network (b) First iteration (c) Second iteration

Figure 2. Demonstration of the Girvan-Newman algorithm for a 12-node network partitioned into three sub-communities

In this paper, we propose a novel Correlation-based Girvan-Newman (CGN) algorithm for the generic monitoring sensor network and AQMSs. Compared with other clustering methods, this network-based approach not only evaluates the correlations between nodes (i.e., sensors) in a cluster with their centroid (i.e., the assigned AQMS) iteratively but also between the nodes at the adjacent clusters to update members of clusters. Hence, it can reflect the spatiotemporal dispersion of air pollution over the whole network.

3.1 Correlation-based Girvan-Newman algorithm

In community detection, the graph $G(V, E)$ is constituted by a set of vertices V (i.e., sensor motes) and a set of edges E (i.e., links or correlations between the sensors). The community detection aims to partition correlated vertices into sub-communities $C_i = \{v_i\}$ $(v_i \in V)$ [14]. In this study, a sensor network comprises air quality sensors and correlations of sensory data that are represented by vertices and edges, respectively. Unlike the original Girvan-Newman algorithm requiring the betweenness scores of all edges to be calculated by the ratio of the number of shortest paths and the number of expected paths between two nodes over the whole network [13], we propose a function of distance and correlation of two adjacent sensors to obtain the edge betweenness among sensors.

The correlation-based edge betweenness centrality for an edge (i, j) in a graph is then determined by using the following equation:

$$C_{\text{edge}}(i, j) = \frac{1}{d(i, j)} + corr(i, j)\,, \tag{1}$$

where $corr(i, j)$ is the Pearson's correlation coefficient and $d(i, j)$ is the Euclidean distance between two sensors i^{th} and j^{th}.

As the distance component is fixed, the variant of correlations of measured data between these air quality sensors will then determine the temporally adaptive feature that reflects the spatial dispersion of air pollution between local and regional areas.

Algorithm 1 Correlation-based Girvan-Newman (CGN) Algorithm

while number of edges in network $G > 0$ **do**
  Calculate edge betweenness centrality for all edges $C_{\text{edge}}(i, j)$ from Eq. (1)
  Find the maximum edge betweenness centrality, *max_betweenness*
  Initialize an empty list *edges_to_remove*
  for each edge (i, j) in G **do**
    if edge betweenness centrality of (i, j) is equal to *max_betweenness* **then**
      Add (i, j) to *edges_to_remove*
    end if
  end for
  Remove edges in *edges_to_remove* from G
  if number of connected components in $G > 1$ **then**
    return list of connected components in G
  end if
end while

3.2 Benchmarks of clustering methods

In data analysis and machine learning applications, there exist several clustering methods depending on the application. Here, we benchmark our proposed algorithm with three popular methods including K-Means Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Agglomerative Clustering. Each of these approaches has different characteristics:

- K-Means Clustering: groups data into 'k' clusters based on minimizing the within-cluster sum of squares [8].
- DBSCAN: identifies clusters based on the density of data points in the feature space with clusters being considered as dense regions of data points that are separated by regions of lower point density [10].
- Agglomerative Clustering: is a hierarchical approach that starts with individual data points as clusters and merges them iteratively [11].

For quantified evaluations, we use the silhouette score which provides a measure of how well-separated the clusters are [17]. This metric is useful upon no ground truth labels (i.e., unsupervised learning) and ranges from -1 to 1, with higher values indicating better-defined and well-separated clusters. The silhouette score is obtained as

$$S_i = \frac{(b_i - a_i)}{max(a_i, b_i)}\,, \tag{2}$$

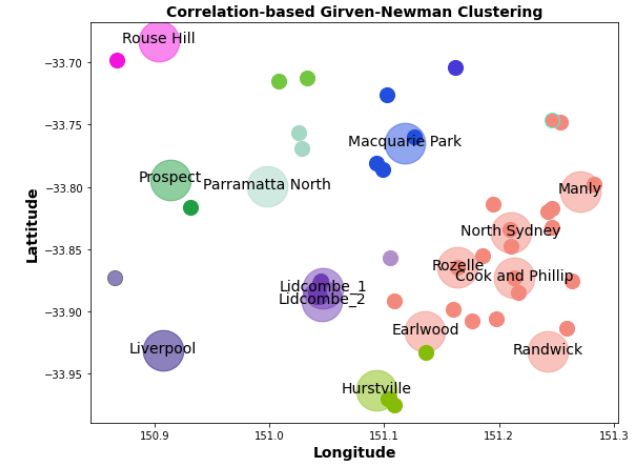
where a_i is the average distance between the observed point (i) to the other points in the same cluster, called the mean intra-cluster distance. The mean nearest-cluster distance (b_i) is the average distance between the observation (i) and the other points of the nearest cluster [17].

4 Results and Discussion

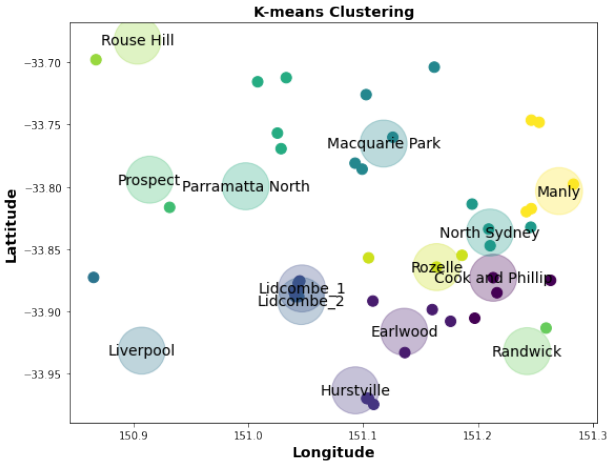
Following the data collecting and processing, a total of 61 data sources, comprising 14 AQMSs and 47 PASs, were selected for subsequent experimentation. The 14 AQMS locations are designated as centroids for clustering, as these stations serve as reference points for assessing the reliability of adjacent low-cost sensors [4].

The results of our experiments are shown respectively in Figures 3a through 3d. In general, all clustering techniques effectively grouped nearby sensors (small dots) with their corresponding nearby AQMSs (large circles), with the exception of the DBSCAN method illustrated in Figure 3c. As previously discussed, the DBSCAN is a density-based clustering algorithm for discovering clusters based on the principle that areas of high data point density are separated by areas of low data point density [10]. In Figure 3c, multiple black dots represent outliers detected by the DBSCAN algorithm indicating the sparsity of these sensor's locations.

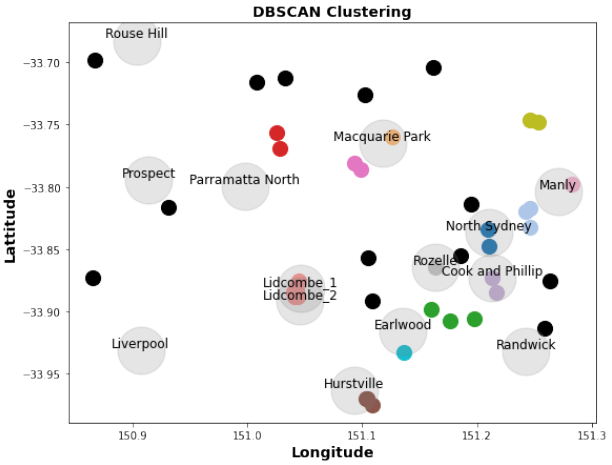
Compared to K-Means and Agglomerative clustering methods with results in Figure 3b and 3d, respectively, our proposed CGN method exhibits analogous patterns in suburban regions of North and West Sydney such as Rouse Hill, Prospect, Parramatta North and Liverpool (Figure 3a). These areas have a limited number of sensors. Hence, the clustering algorithms tend to merge sensors with stations based on a distance metric.



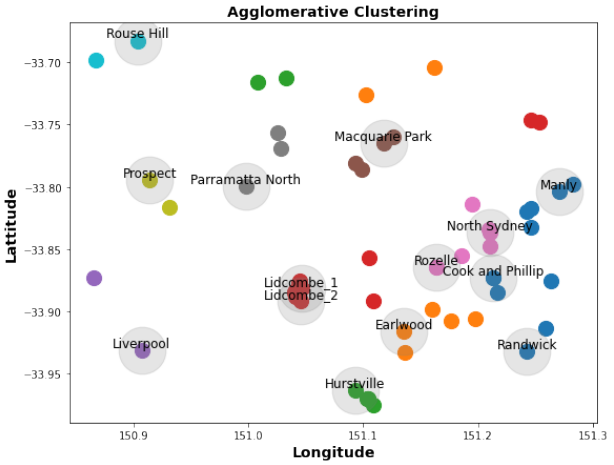
(a) Correlation-based Girven-Newman Clustering



(b) K-Means Clustering



(c) DBSCAN Clustering



(d) Agglomerative Clustering

Figure 3. Comparison between different clustering methods configured for 14 clusters

In East Sydney, sensor clusters exhibit noticeable distinctions between the CGN algorithm and alternative approaches. Specifically, CGN effectively groups sensors in proximity to the AQMNs located in Randwick, Earlwood, Rozelle, Cook and Phillip, North Sydney, and Manly into a coherent cluster as depicted in Figure 3a. This clustering arises from strong correlations among sensors within this geographical area, which plays a significant role in determining the interconnections and weights among sensors as well as monitoring stations. The cluster's pattern suggests the dispersion of aerosols originating from sea salt, indicating a directional spread from the coastal regions towards the southwest, as pointed out in the referenced study [18].

Table 1. Silhouette scores of experiments for all studied methods

No. clusters	K-Means	DBSCAN	Agglomerative	CGN
14	0.304	0.221	0.364	**0.561**

The silhouette scores quantify the final results to compare four algorithms with our real-world data presented in Table 1. The highest score of 0.561 is recorded by the proposed CGN algorithm, followed by Agglomerative, K-Means, and DBSCAN, with scores of 0.364, 0.304, and 0.221, respectively. This outperformance indicates our effective partitioning method to reflect the spatial dispersion of particulate matter. Our approach possesses the ability to leverage correlations in edge betweenness scores to dynamically adjust to temporal fluctuations in time series data obtained from both LWSNs and AQMNs. Consequently, this enables better exploration of spatiotemporal data features in both local and regional contexts, facilitating more comprehensive analysis and prediction.

5 Conclusion

This paper introduces the Correlation-based Girvan-Newman (CGM) algorithm, a novel clustering approach that enhances clustering capabilities by considering both data distance and correlation. The method harnesses spatiotemporal features derived from data collected by cost-effective sensors and state-run stations. Experimental results demonstrate the robustness of our approach in accommodating the unique characteristics of air particles at both local and regional scales. The higher computational intensity is nevertheless associated with multiple iterations involving calculations of correlations, distances, and edge betweenness scores for link removal. Our future work will focus on improving the computational efficiency of the proposed technique.

Acknowledgement

This work was supported by the University of Technology Sydney Project PRO22-16023 and the Department of Planning and Environment, New South Wales, Australia. Huynh A. D. Nguyen would like to thank the Vingroup Science and Technology Scholarship Program, managed by VinUniversity and sponsored by Vingroup, for Overseas Study for Master and Doctoral degrees.

References

[1] Havard T.H. Chan, School of Public Health, *Even low levels of air pollution can harm hearts, lungs in elder.* (2021), accessed on 01.10.2023, https://www.hsph.harvard.edu/news/hsph-in-the-news/even-low-levels-of-air-pollution-can-harm-hearts-lungs-in-elderly/.

[2] WHO, *New WHO Global Air Quality Guidelines aim to save millions of lives from air pollution.* (2021), accessed on 01.10.2023, https://www.who.int/news/item/22-09-2021-new-who-global-air-quality-guidelines-aim-to-save-millions-of-lives-from-air-pollution.

[3] Riley, M.; Kirkwood, J.; Jiang, N.; Ross, G.; Scorgie, Air quality monitoring in NSW: From long term trend monitoring to integrated urban services, *Computers in Industry,* **54**, 44-51, (2020).

[4] Nguyen, H. A. D.; Ha, Q.P., Wireless Sensor Network Dependable Monitoring for Urban Air Quality, *IEEE Access,* **10**, 40051-40062, (2023).

[5] Becnel, T.; Tingey, K.; Whitaker, J.; Sayahi, T.; Lê, K.; Goffin, P.; Butterfield, A.; Kelly, K.; Gaillardon, P.E, A Distributed Low-Cost Pollution Monitoring Platform. *IEEE Internet of Things Journal,* **6 (6)**, 10738-10748, (2023).

[6] Nguyen, H.A.D.; Ha, Q.P.; Duc, H.; Azzi, M.; Jiang, N.; Barthelemy, X.; Riley, M., Long short-term memory Bayesian neural network for air pollution forecast. *IEEE Access,* **11**, 35710-35725, (2023).

[7] Nguyen, H.A.D.; Le, H.T.; Ha, Q.P.; Azzi, M.Deep learning for construction emission monitoring with low-cost sensor network. *Proceedings of the International Symposium on Automation and Robotics in Construction,* IAARC, **40**, 450-457, (2023).

[8] Ikotun, A.M.; Ezugwu, A.E.; Abualigah, L.; Abuhaija, B.; Heming, J., K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences,* **622**, Elsevier, 178-210, (2023).

[9] Govender, P.; Sivakumar, V., Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric pollution research,* Elsevier, **11(1)**, 40-56, (2020).

[10] Deng, D., DBSCAN clustering algorithm based on density. *Proceedings of the 7th international forum on electrical engineering and automation,* IEEE, 949–953, (2020).

[11] Tokuda, E.K.; Comin, C.H.; Costa, L.d.F., Revisiting agglomerative clustering, *Physica A: Statistical mechanics and its applications,* Elsevier, **585**, 126433, (2022).

[12] M. Girvan and M. E. J. Newman, Community structure in social and biological networks, *Proc. National Academy of Sciences (PNAS)*, **99 (12)**, 7821-7826, (2002).

[13] Newman, M.E.; Girvan, M., Finding and evaluating community structure in networks. *Physical review E,* APS, **69(2)**, 026113, (2004).

[14] Chatterjee, B.; Saha, H.N., Detection of communities in large scale networks, *Proceedings of the 10th Annual Information Technology, Electronics and Mobile Communication Conference,* IEEE, 1051–1060, (2019).

[15] The NSW-DPE, *Indicative Air Quality Instrument Evaluation.* (2021), accessed on 01.09.2023, https://www.environment.nsw.gov.au/research_and_publications/publications-search/indicative-air-quality-instrument-evaluation.

[16] Kiruthika, R.; Vijaya, M., Community detection using girvan–newman and kernighan–lin bipartition algorithms. *Proceedings of the 2021 International Conference of Data Intelligence And Cognitive Informatics,* Springer, 217-231, (2022).

[17] Batool, F.; Hennig, C., Clustering with the average silhouette width. *Computational Statistics & Data Analysis,* Elsevier, **158**, 107190, (2021).

[18] Crawford, J.; Cohen, D.D.; Chambers, S.D.; Williams, A.G.; Atanacio, A., Impact of aerosols of sea salt origin in a coastal basin: Sydney, Australia. *Atmospheric Environment,* Elsevier, **207**, 52-62, (2019).