Spatio-Temporal Analysis of Large Air Pollution Data

Mirza Farhan Bin Tarek,[1,*] Md Asaduzzaman,[2] and Mohammad Patwary[3]
[1]Department of Computer Science and Engineering, United International University, Bangladesh
[2]Department of Engineering, Staffordshire University, United Kingdom
[3]School of Computing and Digital Technology, Birmingham City University, United Kingdom
*mtarek141074@bscse.uiu.ac.bd

***Abstract*—Air pollution is one of the most dangerous environmental threats in our planet. Although it is severe in highly populated and industrialized cities of the developing countries, it is also a major concern for developed countries. In developed world, the data are gathered from a large number of air pollution monitoring stations. Therefore, the volume of data are very high and it is not possible to analyze the data efficiently in real-time using the conventional methods. Large scale data mining techniques can help in analyzing those data more efficiently and dynamically. In our paper, we propose a method to mine large amount of air pollution data in order to find air pollution hot spots and time of pollution using clustering methods and time-series analysis and applied to the air pollution data of $PM_{2.5}$, PM_{10} and Ozone in the United Kingdom from 2015-17. The method is able to detect specific pollution zones of those pollutants in the UK. Furthermore, the pollution due to particulate matters was observed to be higher in winter season whereas Ozone pollution was seen to be downwards trending except some areas.**

***Index Terms*—Air pollution, big data mining, clustering, trend analysis.**

I. Introduction

Air pollution is caused when harmful particulates and biological molecules get mixed in the air in a quantity which creates discomfort and health hazards for all the living being as well as damages the ecosystem. It causes instability in the environment and among living organisms. It is a product of rapid and heavy industrialization which produces dangerous pollutants detrimental to health and to the environment as well. Among the different types of pollutants, the major ones are: carbon dioxide (CO_2), sulfur oxides (SOx), volatile organic compounds (VOC) and particulates such as coarse particles, PM_{10} which have diameter between 2.5 to 10 micrometers, $PM_{2.5}$ which have diameter less than 2.5 micrometers, ultrafine particles and soot [1]. These pollutants can create cardiovascular, respiratory problems as well as psychological problems such as cognitive impairment ([2], [3], [4]). Another major consequence of air pollution is global warming which is the gradual increase of the temperature due to the heavy emission of greenhouse gases such as carbon dioxide, nitrous oxides etc. and aerosol and soot. These gases absorb and emit infrared rays in the atmosphere which in turn increase the temperature of the lower atmosphere and surface. Deforestation, rapid industrialization and urbanization are all responsible for global warming or climate change. Overall, air pollution is a severe problem in the highly populated and industrialized cities of the developing countries, but it is also of major concern in developed countries as well. As air pollution has grave negative impact on human health as well as on the ecology, it is our duty to mitigate air pollution.

For this reason, many countries have established pollution monitoring stations that use sophisticated technologies to monitor the concentration of different pollutants. Traditional monitoring systems include the use of large, expensive stationary monitoring stations which use different expensive monitoring equipment. The volume of data generated from these stations is quite high and it is not possible to efficiently analyze this huge amount of data in real-time using the conventional methods. Large scale data mining can be applied for analysing these data more efficiently. Different authors have used data mining techniques to analyse air pollution data, however, due to the size of the data size used in the analysis the procedure would not be referred to as big data analysis ([5], [6], [7]).

In our paper, we propose a method using existing algorithms to mine air pollution big data in real-time to discover air pollution hot spots. For our analysis, we have used K-means clustering, Clustering LARge Applications (CLARA) and trend analysis to analyze $PM_{2.5}$, PM_{10} and ozone situation in different cities around the United Kingdom over the years 2015-2017.

The rest of the paper is structured as follows. Section II describes different related works that have been done in the field of air pollution analysis using data mining. Section III covers the data mining algorithms that have been used in the analysis of our data. Section IV describes the result of our experiment on air pollution data of UK. Finally, we conclude our discoveries in section V and propose the direction of future works.

II. Related Works

Clustering is one of the unsupervised learning techniques used in data mining which is the subset of computer science where large data sets are analyzed for discovering patterns and knowledge using machine learning, statistical methods etc. and unsupervised learning technique is finding patterns from unlabeled data and learning from that data set so that the model can identify future data points. It includes partition based clustering, density-based clustering etc. K-means clustering is one of the most popular clustering algorithms because of its simplicity and performance [7]. As the initial number of clusters have to be pre-selected in k-means clustering, it is not feasible to use it to get natural clustering results but it is a very fast algorithm [8]. In [7], the authors focused on determining only air quality index from analyzing ozone pollution data using a novel and improved version of K-means clustering and possibilistic fuzzy c-means (PFCM). However, neither results

did point to a specific location with high or low air quality index nor to the time when the pollution spikes. The authors also only compared the results of the enhanced k-means algorithm with PFCM and not any other algorithm. Other works have been done using K-means clustering algorithm and hierarchical clustering algorithm. In [9], the authors devised a framework to spatially cluster zones according to their $PM_{2.5}$ composition in the USA to understand the long-term composition difference and heterogeneity of $PM_{2.5}$ using K-means clustering algorithms.

Spatio-temporal clustering is a clustering technique which can cluster data with both spatial or location and temporal time data. It can create a relationship between the spatial and temporal dimensions of a data set. Also, it can find interesting distribution patterns for data in a dataset that can lead to further studies [10]. Different spatio-temporal clustering techniques exist e.g. ST-DBSCAN (spatio-temporal density-based clustering for applications with noise), ST-SNN (spatio-temporal shared nearest neighbor) etc. and even other clustering algorithms like K-means can be modified to be used in a spatio-temporal database ([11], [6], [12]). For clustering pollution zones other algorithms like ST-SNN or K-means clustering can be used. [13] developed a method for using K-means clustering and multiple regression models for predicting the hourly ozone level in Dallas. The objective was to cluster time series of ozone pollution and spatial clustering of ozone monitoring stations to create input set for linear regression analysis. Also, hierarchical clustering was used to recognize spatial patterns [13].

Trend of meteorological time series can be determined or estimated using a number of different methods. A large number of research have been performed using trend analysis in the field of climate change however, ozone trend analysis has not been addressed in few cases. [14] analysed the climate change trends and temperature trend during 1960-2000 at 19 stations along with Lankang River (China). They discovered increase in temperature and precipitation. Also, [15] used Mann-Kendall test and Sen's slope estimator to analyse the variability of seven meteorological variables at 12 stations in Serbia during 1980-2010. They discovered increasing trend in annual and seasonal minimum temperature and maximum temperature however there was a decrease in the relative humidity. [16] developed a framework for analysing time series data using Mann-Kendall test and Sen's slope estimator which was also to show that, due to the establishment of coal powered thermal power plants in different parts of India, excessive emission of greenhouse gases increased rainfall in affected areas. In another paper, the authors analysed the spatial and temporal trend of ozone in eastern North America, Europe and East Asia using generalized additive mixed model. Their research showed that East Asia had the greatest human and plant exposure to ozone pollution among the investigated regions [17].

III. Methods and Materials

A. Data Collection

For our analysis, we used data of ground level ozone, $PM_{2.5}$ and PM_{10} which are dangerous pollutants having harmful effects on human health. Daily maximum 8-Hour mean ozone concentration, hourly 24-hour mean concentration of (PM_{10}) and annual mean concentration of $PM_{2.5}$) were collected from

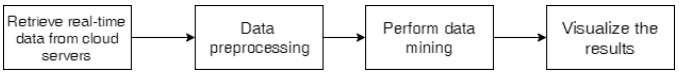

Fig. 1. Steps of data analysis

the archive of Department of Environment, Food and Rural Affairs (DEFRA) website. The data are available for public access and updated regularly. The data are sampled from 22 monitoring stations around the UK which are under the Automatic Urban and Rural Network, the largest monitoring network in the UK and the principal network which is used for checking compliance with the regulation of the Air Quality Directives. The network measures data of oxides of nitrogen (NO_x), Sulfur dioxide (SO_2), Ozone (O_3), Carbon Monoxide (CO) and particles (PM_{10}, $PM_{2.5}$).

B. Data Analysis Methods

Analysis of the air pollution data was performed using K-means clustering, CLARA and trend analysis methods. The algorithms and methods are described below.

1) K-means Clustering: It is a widely used, popular and powerful algorithm which divides the data set into K groups which have members with similar characteristics. It has a number of features and advantages which makes it a simple but powerful algorithm. Firstly, it is computationally efficient [18]. Secondly, the algorithm is reported to be less sensitive to outliers [19].

2) CLARA: CLARA or Clustering LARge Applications is a variation of the partitioning around medoids or PAM algorithm. It is a popular version of the K-medoids algorithm which selects data points rather than mean point (medoid) as cluster centers and works with a generalization of the Manhattan Norm for finding distance between data points. A common version of the K-medoids algorithm is PAM or partitioning around medoids. However, it is not suitable for large data sets as it tries to find medoids from the whole data set which may be computationally expensive for huge data sets. Therefore, for large data sets CLARA is much suitable which takes small samples from the data set and applies PAM to generate optimum set of medoids for the sample [20].

Algorithm 1: PAM algorithm

Data: Initialize k of the n data points as medoids
Result: $C = \{c_1, c_2, c_3, ..., c_k\}$ (set of cluster centroids)
L = labels for E
Associate each data point to the closest medoid
while *decrease the cost* $\frac{\sum_{i=1}^{n} d(O_i, rep(M, O_i))}{n}$ **do**
  for *each medoid* m*, for each non-medoid data point* o **do**
    Swap m and o, recompute the cost (sum of distances of points to their medoid)
    If the total cost of the configuration increased in the previous step, undo the swap
  end
end

3) Trend Analysis: As ozone pollution is treated differently from particulate matter pollution and the data was daily rather than hourly, we use trend analysis to detect the temporal

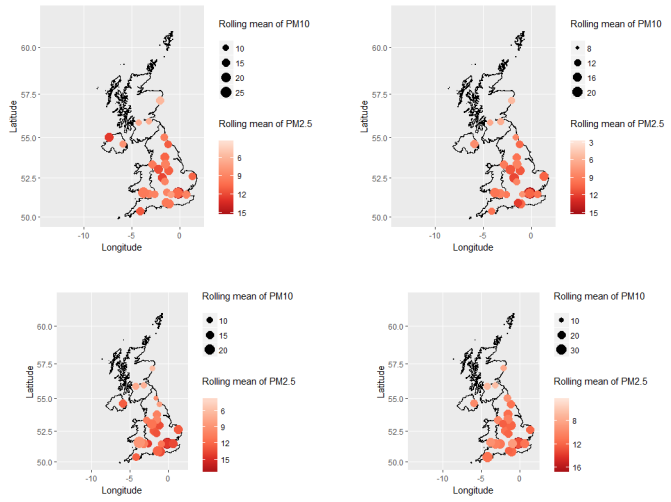

Fig. 2. $PM_{2.5}$ and PM_{10} concentration of different stations in UK in January, April, August and December

pattern for ozone in the data. Detecting seasonal trends from time series data can give us useful insights about the temporal properties of the variables included in the time series. It can be done using non-parametric statistical tests and is very useful for analyzing hydro-meteorological time series as they have been proven useful for analyzing non-normally distributed and censored data containing missing values. Besides this, non-parametric tests can tolerate outlier data and non-normally distributed time series data [21]. For discovering trend from ozone time series, we used Mann-Kendall trend test to detect presence of trend and Sen's slope estimator to determine the direction of trend. Both of these are non-parametric statistical tests.

IV. Results and Discussion

A. Results of Clustering

Using DB-index, we have determined that for K = 4, we get the lowest db value which indicates compact cluster assignments. The 24-hour mean data of PM_{10} and annual mean data of $PM_{2.5}$ was analyzed using the proposed method to detect temporal clusters in which elevated level of pollution was detected. The pollution was measured using the Air Quality Index supplied by Committee on Medical Effects of Air Pollutants (COMEAP). The national air quality objectives directives say that 24-hour mean of PM_{10} cannot exceed $50\,\mu g/m^3$ more than 35 times a year and the threshold value for annual mean of $PM_{2.5}$ is $25\,\mu g/m^3$.

Both K-means clustering and CLARA have given almost same results. After clustering, we have discovered 4 clusters each having 22 monitoring stations. Among the clusters, cluster 1 contains data from 22 monitoring stations over the range of 3 years. Majority of the data are from spring and winter season with the high frequency of December, March and May. However, in cluster 2, which had high tendency to occur in December, January, February and March, had high concentration of PM_{10}. The mean concentration was $52.14\,\mu g/m^3$ which was higher than the safety threshold. Among these stations, Port Talbot Margam had consecutively violated the directives in the 3 years. On the other hand, cluster 3 and 4 had relatively safe level of $PM_{2.5}$ and PM_{10} concentration.

The 4 clusters discovered using the CLARA algorithm are not that different from the clusters from K-means algorithm. Each of the 4 clusters contain data of 22 unique stations. The mean value of $PM_{2.5}$ concentration in different clusters is quite low. However, the maximum value of PM_{10} is quite high in all the clusters which exceeds the national air quality objectives. The highest concentration was observed in cluster 4 with a value of $220.8625\,\mu g/m^3$. The mean value of the cluster is $16.2326\,\mu g/m^3$ conditioning data mostly from the autumn season. The stations which had high PM_{10} concentration were Leamington Spa and Port Talbot Margam mostly in the months of November and December. The highest mean value that was observed was $17.839\,\mu g/m^3$ which was in cluster 1 and the cluster had most of the data from winter season. The stations which recorded concentration over $50\,\mu g/m^3$ are Aberdeen, Birmingham A4540 Roadside, Birmingham Tyburn, Birmingham Tyburn Roadside, Cardiff Centre, Derry, Leamington Spa, Liverpool Speke, London Bloomsbury, London Harlington, London Marylebone Road, Middlesbrough, Newcastle Center, Norwich Lakenfields, Nottingham Center, Port Talbot Margam, Sheffield Devonshire Green and Southampton Center.

B. Difference Between the K-means and CLARA

K-means and CLARA algorithm perform clustering following different principles and for that, the results were varying. However, as both are partitioning algorithms, no drastic changes were observed. The agreement between the clustering results can be observed by checking the Rand index. Higher Rand index means better cluster agreement. For our data set, the observed Rand index was 0.5 which means a lower cluster agreement.

The run time of the algorithms were also different. K-means clustering on 245,127 data points with 4 clusters took 43.37 seconds to complete whereas CLARA algorithm took 0.14 seconds to complete which is almost 99% faster than K-means algorithm. This is because, CLARA is optimized for big data as it performs PAM algorithm on small samples of the data set rather than on the whole data set.

C. Results of Trend Analysis

Time series of ozone in 2015-2017 was used to determine the trend of ozone concentration. According to the European Union standards, the safety threshold of daily max 8-hour ozone is $120\,\mu g/m^3$ [22]. We have discovered that only 12 stations among the 22 monitoring stations has reported exceeded threshold and even so, the number of times it exceeded not that much high. Therefore, we can safely say that the ozone situation is quite normal in UK. Furthermore, analysis showed that usually the ozone concentration rises during the first half of the year meaning during the late winter and early spring season. This result is consistent with all monitoring stations. Except the stations Aberdeen, Birmingham Tyburn, Birmingham A450 Roadside, Liverpool Speke, London Marylbone Road, Newcastle center and Plymouth, all other stations showed downward ozone trend.

V. Conclusion and Future Works

In our paper, data mining algorithms especially clustering algorithms were used to analyse the spatial and temporal

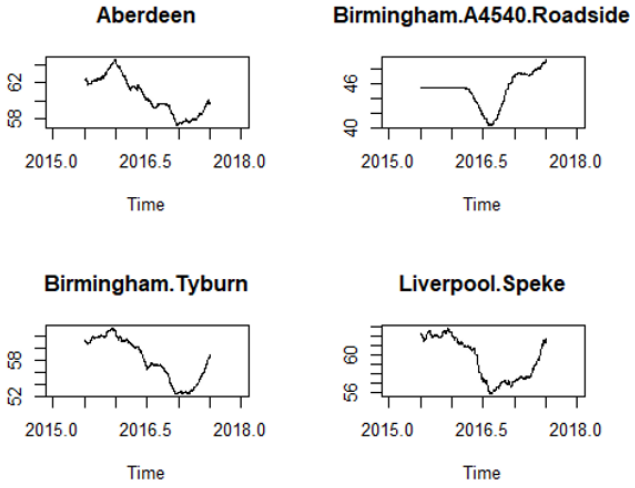

Fig. 3. Extracted upward trend from the time series data of Ozone concentration of 4 different monitoring stations

characteristics of large air pollution data. We have used cluster analysis and time series analysis for analysing particulate matter and ozone pollution respectively. From our analysis, we have shown that $PM_{2.5}$ pollution is somewhat better in different cities however, high PM_{10} concentration have been recorded to be quite high in Aberdeen, Birmingham, Cardiff Centre, Derry, Leamington Spa, Liverpool Speke, London Bloomsbury, London Harlington, London Marylebone Road, Middlesbrough, Newcastle Centre, Norwich Lakenfields, Nottingham Centre, Port Talbot Margam, Sheffield Devonshire Green, Southampton Centre. Some of the monitoring stations listed are located near highways like London marylbon road, Birmingham and so, they tend to detect high amount of particulate pollution. Furthermore, Port Talbot Margam, Leamington Spa, London Marylbone Road are dangerous particulate pollution zones. Because, Port Talbot has large number of steelworks factories, Leamington Spa has high traffic with diesel and petrol cars and buildings with 'canyon like' architecture and London Marylbone street has high traffic ([23], [24]). Particulate pollution is higher in winter because of overcast weather which traps the pollution under the clouds. Furthermore, summer pollution is caused mostly for stale weather and ultraviolet exposure. On the other hand, from the ozone time series analysis, we have shown that the ozone trend is downwards in different monitoring stations. This is a good sign as ground level ozone can cause serious health problems.

We expect to further analyze the air pollution problem using detailed predictive and statistical models which can predict different pollution levels with high accuracy. We also propose to use distributed computing with real-time analysis for more efficient analysis.

REFERENCES

[1] A. Goel, S. Ray, P. Agrawal, and N. Chandra, "Air pollution detection based on head selection clustering and average method from wireless sensor network," in *Proceedings - 2012 2nd International Conference on Advanced Computing and Communication Technologies, ACCT 2012*, 2011, pp. 434–438.

[2] X. Zhang, X. Zhang, and X. Chen, "Happiness in the air: How does a dirty sky affect mental health and subjective well-being?" *Journal of Environmental Economics and Management*, vol. 85, pp. 81–94, 2017.

[3] R. D. Brook, S. Rajagopalan, C. A. Pope, J. R. Brook, A. Bhatnagar, A. V. Diez-Roux, F. Holguin, Y. Hong, R. V. Luepker, M. A. Mittleman, A. Peters, D. Siscovick, S. C. Smith, L. Whitsel, and J. D. Kaufman, "Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the american heart association," pp. 2331–2378, 2010.

[4] L. K. Fonken, X. Xu, Z. M. Weil, G. Chen, Q. Sun, S. Rajagopalan, and R. J. Nelson, "Air pollution impairs cognition, provokes depressive-like behaviors and alters hippocampal cytokine expression and morphology," *Molecular Psychiatry*, vol. 16, no. 10, pp. 987–995, 2011.

[5] M. Hanesch, R. Scholger, and M. J. Dekkers, "The application of fuzzy C-means cluster analysis and non-linear mapping to a soil data set for the detection of polluted sites," *Physics and Chemistry of the Earth, Part A: Solid Earth and Geodesy*, vol. 26, no. 11-12, pp. 885–891, 2001.

[6] A. Musdholifah, S. Z. B. Mohd Hashim, and I. Wasito, "KNN-kernel based clustering for spatio-temporal database," *International Conference on Computer and Communication Engineering, ICCCE'10*, no. May, pp. 11–13, 2010.

[7] K. G. Ri, R. Manimegalai, G. D. M. Si, R. Si, U. Ki, and R. B. Ni, "Air Pollution Analysis Using Enhanced K-Means Clustering Algorithm for Real Time Sensor Data," no. August 2006, pp. 1945–1949, 2016.

[8] X. Wang, K. Smith, and R. Hyndman, "Characteristic-based clustering for time series data," *Data Mining and Knowledge Discovery*, vol. 13, no. 3, pp. 335–364, 2006.

[9] B. Zou, F. Peng, N. Wan, K. Mamady, and G. J. Wilson, "Spatial cluster detection of air pollution exposure inequities across the United States," *PLoS ONE*, vol. 9, no. 3, p. e91917, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24647354

[10] Y. Zhang and C. F. Eick, "Novel clustering and analysis techniques for mining spatio-temporal data," in *Proceedings of the 1st ACM SIGSPATIAL PhD Workshop on - SIGSPATIAL PhD '14*, 2014, pp. 1–5. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2694859.2694865

[11] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial-temporal data," *Data and Knowledge Engineering*, vol. 60, no. 1, pp. 208–221, 2007.

[12] S. Wang, T. Cai, and C. F. Eick, "New spatiotemporal clustering algorithms and their applications to ozone pollution," *Proceedings - IEEE 13th International Conference on Data Mining Workshops, ICDMW 2013*, pp. 1061–1068, 2013.

[13] M. Ahmadi, Y. Huang, and K. John, "Application of spatio-temporal clustering for predicting ground-level ozone pollution," in *Advances in Geographic Information Science*, 2017, pp. 153–167.

[14] H. Yunling and Z. Yiping, "Climate Change from 1960 to 2000 in the Lancang River Valley, China," *Mountain Research and Development*, vol. 25, no. 4, pp. 341–348, 2005.

[15] M. Gocic and S. Trajkovic, "Analysis of changes in meteorological variables using Mann-Kendall and Sen's slope estimator statistical tests in Serbia," *Global and Planetary Change*, vol. 100, pp. 172–182, 2013.

[16] A. Baheti and D. Toshniwal, "Trend Analysis of Time Series Data Using Data Mining Techniques," *2014 IEEE International Congress on Big Data*, pp. 430–437, 2014. [Online]. Available: http://ieeexplore.ieee.org/document/6906812/

[17] K.-L. Chang, I. Petropavlovskikh, O. R. Copper, M. G. Schultz, and T. Wang, "Regional trend analysis of surface ozone observations from monitoring networks in eastern North America, Europe and East Asia," *Elem Sci Anth*, vol. 5, no. 0, p. 50, 2017. [Online]. Available: https://www.elementascience.org/article/10.1525/elementa.243/

[18] D. Steinley, "K-means clustering: A half-century synthesis," *British Journal of Mathematical and Statistical Psychology*, vol. 59, no. 1, pp. 1–34, 2006.

[19] G. Punj and D. W. Stewart, "Cluster Analysis in Marketing Research: Review and Suggestions for Application," *Journal of Marketing Research*, vol. 20, no. 2, p. 134, 1983. [Online]. Available: http://www.jstor.org/stable/3151680?origin=crossref

[20] L. Kaufman and P. J. Rousseeuw, "Clustering by means of medoids," pp. 405–416 416, 1987.

[21] M. R. Kousari, H. Ahani, and R. Hendi-zadeh, "Temporal and spatial trend detection of maximum air temperature in Iran during 1960-2005," *Global and Planetary Change*, vol. 111, pp. 97–110, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.gloplacha.2013.08.011

[22] European Commission, "Air Quality Standards," p. 1, 2017. [Online]. Available: http://ec.europa.eu/environment/air/quality/standards.htm

[23] J. McCarthy, "The Welsh towns and cities with the most toxic air in the UK," 2017. [Online]. Available: https://www.walesonline.co.uk/news/wales-news/welsh-towns-cities-most-toxic-13838337

[24] S. Cowling, "Leamington one of the worst in the country for air pollution," 2016. [Online]. Available: https://theboar.org/2016/06/leamington-worst-country-air-pollution/