IJEP 41 (1) : 42-49 (2021) (ISSN : 0253-7141)

# Closed Pattern Mining And Causal Analysis Of Pollution Data

**S. Sharmiladevi and S. Siva Sathya***

*Pondicherry University, Department of Computer Science, Puducherry - 605 014, India*

**Corresponding author, Email : ssivasathya@gmail.com; sharmiladevi94@gmail.com*

**Mining sequential patterns are of great importance in recent years, as it unveils some of the unknown associative relationships between observations. While in mining sequential patterns many intermediate sequences have to be generated, which is a computationally challenging task when compared to frequent patterns of mining. CloFAST is an algorithm which mines closed sequences without candidate maintenance. Also, CloFAST requires only one step to check closure and prune the search space. It can mine long closed sequences effortlessly from large datasets. In this work, a closed sequential pattern mining of $PM_{2.5}$ pollutant in Delhi is done using CloFAST. Delhi, the capital of the second most populous country on earth has been suffering from severe air pollution problem. Delhi is getting polluted due to diverse reasons, like its geography, burning crop stubble in neighbouring states, vehicular emission, etc. Some of the critical air pollutants found in Delhi are $PM_{10}$, $PM_{2.5}$, nitrogen oxide, sulphur oxide, carbon monoxide, ozone. The main pollutant being particulate matter ($PM_{2.5}$) as it causes serious health problems when it enters into the alveoli of human lungs. Various micro-level analysis of air pollution is being carried out recently. But macro-level analysis is also required in order to obtain a clear understanding on a broader scale. The patterns obtained are given as knowledge for causal analysis done using the FCI algorithm.**

**KEYWORDS**

Air pollution, $PM_{2.5}$ pollutant, Closed sequential mining, Particulate matter, Delhi

## 1. INTRODUCTION

According to The Air (Prevention and Control of Pollution) Act, 1981 - Air pollution is the presence of any solid, liquid, or gaseous substance in the atmosphere in such concentration as may be or tend to be injurious to human beings or other living creatures or plants or property or environment. Delhi has become one of the most polluted cities in India. A survey conducted by WHO in 1600 cities in the world reveals that the air quality in Delhi is the worst. The pollution level in Delhi has increased because of various factors, like geography, post-harvest crop burning, vehicular emission, meteorological factors, dust from constructions, etc. A study reveals that the pollution level in Delhi during winters is 40-80% higher than the other month of the year [1]. Particulate matter present in the atmosphere causes severe environmental problems, serious health hazards and climate change. Suspended particulate matter can be categorised as $PM_{2.5}$ and $PM_{10}$, based on the diameter of the particles. $PM_{10}$ has a diameter of 10 μm or less, whereas $PM_{2.5}$ has a diameter of 2.5 m or less. $PM_{2.5}$ plays a vital role in the calculation of the air quality index (AQI). $PM_{2.5}$ can enter into the deepest region of our lungs and cause numerous health problems [2]. There is an epidemiological link between exposure to $PM_{2.5}$ and an increase in total mortality, cardiovascular mortality, respiratory mortality and hospital admission of COPD [3]. Continuous exposure to $PM_{2.5}$ can result in grave danger [4].

According to the research paper published by the Ministry of Earth Sciences, India in October 2018, approximately 41% of $PM_{2.5}$ pollutants is attributed to vehicular emission, 21% to dust and around 18% to industries in Delhi [5]. Better understanding about this pollutant and its pattern is essential to prevent or control further damage. Data mining is the process of extracting useful information; from data stored in large databases to get an insight about the data for making decisions. Pattern mining is the process of finding interesting, effective and unanticipated patterns in a database. Frequent item sets, subgraphs, associations, sequential rules and periodic patterns are the patterns found in a database. The job of sequential pattern mining is to find all frequent sub-sequences in a sequence database, where the interestingness of a sub-sequence can be measured in terms of various criteria, such as its occurrence frequency, length and profit. A sequence $\underline{S}$ is said to be a frequent sequence or a sequential pattern if and only if $sup(\underline{S}) \geq$ 'minsup', for a threshold 'minsup' set by the user [6]. Sequential pattern mining


© 2021 - Kalpana Corporation

has numerous real-life applications because data is naturally found as sequences of symbols across various domains, like bioinformatics, e-learning, market basket analysis, texts and webpage click-stream analysis.

Closed sequential patterns are the set of sequential patterns that are not included in other sequential patterns having the same support. This is a lossless form of compression [7]. The support value is stored within the patterns. Closed patterns are more helpful than maximal patterns because maximal patterns are a lossy form of compression. The patterns obtained by closed sequential mining are condensed representations of super patterns from which it is possible to infer the subsets without any loss of information [8]. Causal analysis is done to find the cause and effect variables. According to Granger causality, a cause will always precede its effect. Given two time series data $\{X_t\}$ and $\{Y_t\}$, X is said to Granger cause Y (X -->Y) if the past values of X help predict the values of Y, in addition to the past values of Y.

$$y_t \approx \sum_{l=1}^{L} a_l \cdot y_{t-1} + \sum_{l=1}^{L} b_l \cdot x_{t-l}$$

$$y_t \approx \sum_{l=1}^{L} a_l \cdot y_{t-l}$$

Where l is the time lag and L is the maximum time lag permitted. Causation implies association, but the vice versa is not true.

In this work closed sequential pattern mining is performed on the $PM_{2.5}$ values recorded in Delhi for 2016-18. This is a real time monitoring data recorded in 14 stations present in Delhi. Sequential patterns only represent the hidden associative relationships between different stations; therefore, a causal inference algorithm was used to find the propagation pattern. The associative patterns obtained was given as knowledge to this search algorithm. The patterns obtained can be further analysed to get a deeper understanding of air pollution in Delhi.

### 1.1 Literature review

In this section, a brief literature review about the sequential mining techniques and research techniques used for air pollution analysis is presented.

**1.1.1 Sequential mining:** The process of finding all the sequential patterns having a support value greater than or equal to the user specified minimum support is called sequential mining. Generalised sequential pattern mining (GSP) is one of the oldest algorithms in discovering sequential patterns. It makes use of an Apriori-like approach [9]. Later in order to increase the efficiency many algorithms, like SPADE, CM-SPADE, SPAM, CM-SPAM, LAPIN, etc., was proposed [10,11,12,13]. Since sequential pattern mining generates an abundant number of subsequences, researchers started focusing on closed sequential mining. Mining closed sequential patterns results in increased efficiency and provides compact patterns. Various algorithms, like CM-ClaSP, CloFAST, CloSpan, BIDE are used in mining closed sequential patterns [8,11,14,15]. CloFAST mines closed frequent sequences of itemsets. Contrary to the existing algorithms, which requires iterative itemset extension and sequence extension, CloFAST requires only one-step to check sequence closure and to prune the search space.

**1.1.2 Causal analysis:** The causal analysis proposed by Granger is a method to find a causal relationship between variables. This test is used to find if one variable affects another. Granger's causality test can be used in various domains, like psychology, economics, bioinformatics, epidemiology, etc. The fast causal inference (FCI) algorithm performs independence tests on the observed data to get the partial ordered graph (PAG) which provides details about the relationships between the observed variables [16]. The advantage of FCI algorithms is the possibility of taking latent confounding variables into account, as opposed to methods based on Granger causality. Latent confounding variables are those that affect two variables which are not spuriously related. Causal analysis is also used in air pollution analysis. Pollution diffusion patterns between neighbouring cities are found in a study conducted by Jiang [17]. Air pollution propagation patterns are discovered by modelling a spatio-temporal causal graph by Li [18]. A big data analysis of air pollution data is done to find the causal pathways in studies conducted by Zhu [19,20].

**1.1.3 Air pollution analysis:** Analysis of air pollution is a blooming stream. Since air pollution is related to various environmental factors, cross-domain research works are carried out in this field, like the study of air pollution with respect to meteorological factors, road traffic, urbanization, etc. Epidemiological studies in connection with air pollution are also being carried out. Micro-level analysis of air pollution is done to find or analyse the ionic components present in $PM_{2.5}$ and its origin [21]. Discovering spatial, temporal and spatio-temporal patterns is of interest in recent times. Time series based analysis on air pollution is done by Joshi using K-mean and X-mean clustering techniques to extract useful patterns [22]. Clustering based analysis of air pollution is done by Sathya using K-means and


© 2021 - Kalpana Corporation

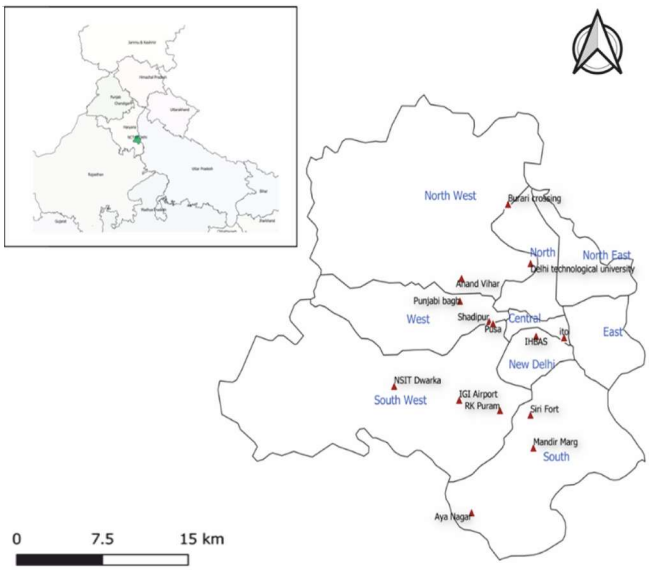


**Figure 1.** *Location of monitoring stations in Delhi*

expectation maximization (EM) algorithm [23]. Spatio-temporal patterns of $PM_{2.5}$ is found in stations used in this study. In figure 1 the location of each station can be seen. Google maps were used for creating it using spatial statistical approaches like spatial interpolation and spatial regression [24]. Zhou created a map to represent the spatio-temporal patterns, further, a calendar view was also used to visualize the air quality condition and the primary pollutants [25]. Geo-spatial and geostatistical techniques are used by Dadhich to obtain the seasonal and temporal patterns of various pollutants and also analyse the relationship between air quality and weather [26]. Spatio-temporal variation of $PM_{2.5}$ concentration and its relationship with population, gross domestic product (GDP) and landuse is analysed by using geographically weighted regression (GWR) model [27]. Descriptive and predictive analysis of air pollution data was done to study the trends of various air pollutants and forecast future trends [28]. Spatio-temporal features of $PM_{2.5}$ was analysed by using sequential mining technique [29].

Epidemiological studies with respect to $PM_{2.5}$ is essential to understand the severity of this pollutant. A detailed survey about the application of data mining and machine learning methods used in air pollution epidemiology studies is presented by Bellinger [30]. Meteorological factors, like wind direction, wind speed, temperature, humidity, rainfall, etc., affect $PM_{2.5}$ production and dispersal. Zhao created a data mining model for predicting $PM_{2.5}$ considering the meteorological data [2]. Three models were created by using the logistic regression, linear discriminant analysis and random forest. Cohen's Kappa coefficient was used to measure the accuracy of the result. Random forest gave the most accurate prediction result.

**Table 1.** *Stations considered for the analysis alongwith location type*

| Monitoring station / location | Types of activities (residential /commercial/traffic/industrial) |
|---|---|
| Income Tax Office (ITO) | Traffic intersection |
| Delhi Technological University (DTU) | Residential, industrial |
| Siri Fort | Residential, mix |
| Shadipur | Residential, industrial |
| Institute of Human Behaviour and Applied Sciences (IHBAS) | Residential, industrial |
| RK Puram | Residential |
| Mandir Marg | Residential, commercial |
| Punjabi Bagh | Industrial, residential, commercial |
| Anand Vihar | Commercial |
| Aya Nagar | Industrial, residential |
| IGI Airport Terminal-3 IMD | Commercial |
| Burari crossing | Residential, commercial |
| Pusa | Residential, industrial |
| NSIT Dwarka (Dwarka) | Residential |

## 2. MATERIAL AND METHOD

### 2.1 Data collection

The air pollution data in Delhi is considered for this study. $PM_{2.5}$ value recorded in 14 stations is used for finding closed sequential patterns. The stations were chosen based upon data availability and the nature of its location. Data recorded from 2016-18 on 24 hr basis was considered. This data can be obtained from the Central Pollution Control Board (CPCB). Table 1 illustrates the list of stations used in this study. In figure 1 the location of each station can be seen. Google maps were used for creating it.

Preprocessing of the raw data was done in order to fill the missing value. Linear interpolation method was used for filling the missing value, because of its simplicity and its past success in dealing with environmental data. After preprocessing the $PM_{2.5}$ value was categorised into safe and dangerous based upon the National Ambient Air Quality Standard (NAAQS). According to this standard, the permissible range of $PM_{2.5}$ pollutant for 24 hr weighted average value is 0-60 $\mu g/m^3$, this level is considered to be safe and breathable. If the recorded value is more than this level then it's dangerous. The data was analysed to find seasonal patterns. The seasons in Delhi are broadly categorised into spring (Feb-


© 2021 - Kalpana Corporation

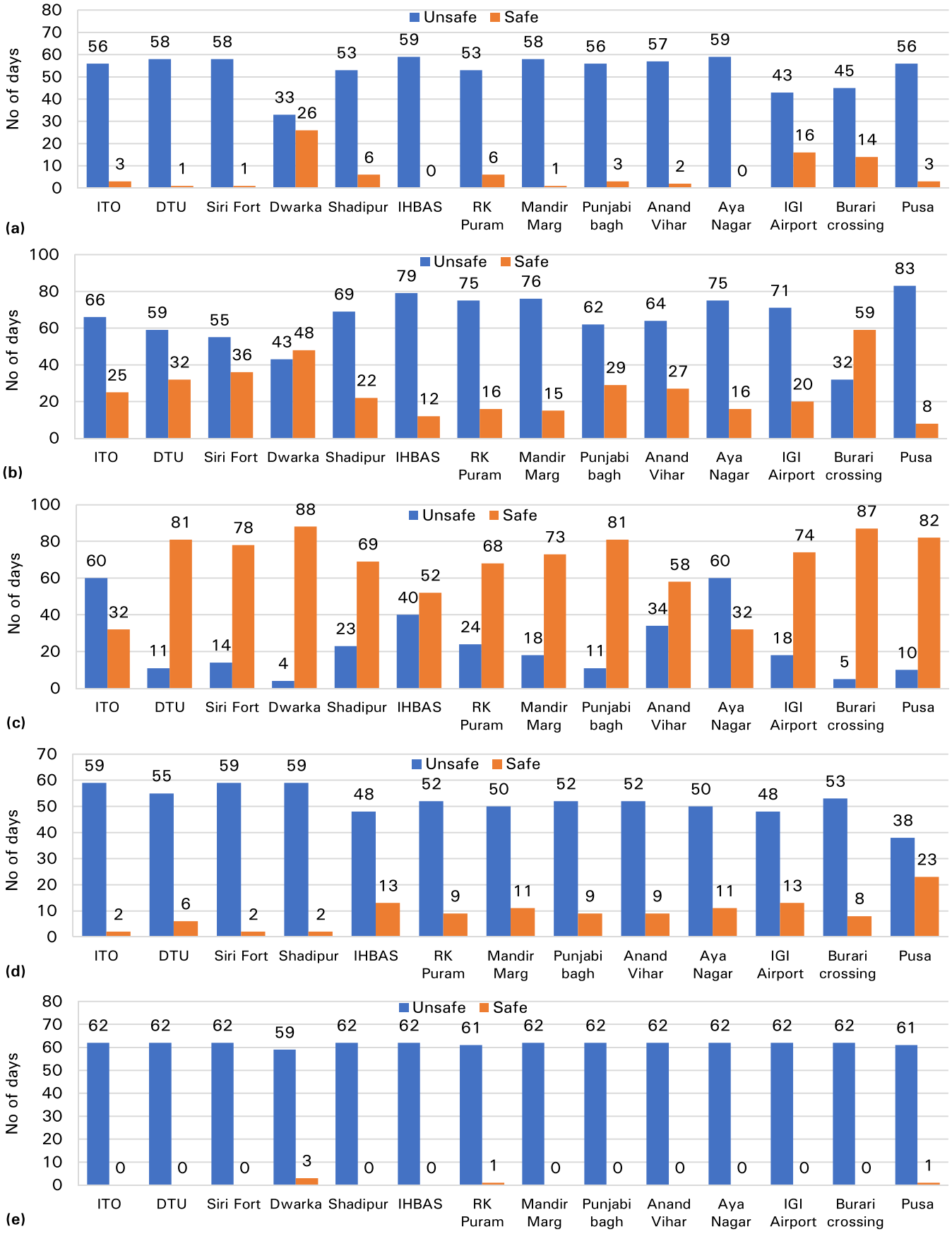


**Figure 2.** *$PM_{2.5}$ safe and unsafe days recorded during (a) spring, (b) summer, (c) monsoon, (d) autumn and (e) winter*


© 2021 - Kalpana Corporation

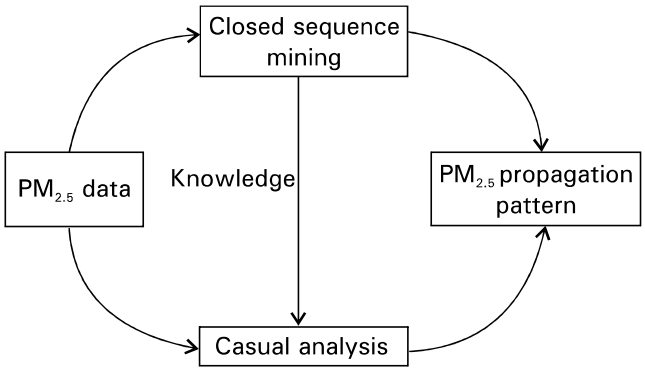


**Figure 3.** *Methodology*

ruary to March), summer (April to June), monsoon (July to September), autumn (October to November), winter (December to January). The preprocessed data classified as safe and unsafe based upon $PM_{2.5}$ pollutant can be visualized in figure 2. From figure 2c we can conclude that $PM_{2.5}$ pollutant level recorded during monsoon is better compared to other seasons. Figure 2e shows that the pollutant level during winter is the worse. Also, from figure 2, it can be seen that Pusa and IGI have a safe level of $PM_{2.5}$ recorded for most of the days.

### 2.2 Methodology

This work focuses on finding whether $PM_{2.5}$ pollutant measured in one station affects the level of $PM_{2.5}$ measured in other stations (Figure 3). This is carried out in two phases.

**2.2.1 Phase 1 - ClosFAST for closed sequences**: In the first step the pre-processed data was given as input to a CloFAST algorithm, the result obtained is a set of closed sequences. The major steps involved in CloFAST are:

1. The dataset is scanned only once to construct the sparse id-lists (SIL). The SIL stores the position of the transaction containing the given itemset.
2. SIL is used to mine the closed frequent itemsets (CFI). SILcfi is utilised to create vertical id-lists (VILcfi) of nodes at the first level. The vertical id-lists store the position of sequential patterns in the input sequence.
3. Closed sequences are mined in a depth first search strategy, by performing sequence extension on the VILcfi. Backward closure checking is done to prune the search space.

The meaning of a sequence is that there is some relation between the items involved in it. In our case, it means that the stations are associated with each other with respect to $PM_{2.5}$ pollutant. For example consider the spring pattern - DTU | Dwarka | Siri Fort, from this we can imply that during spring season whenever DTU recorded unsafe level of $PM_{2.5}$, Dwarka and Anand Vihar also had unsafe level. Mining a closed sequence reduced the number of redundant patterns generated to a maximum extent.

**2.2.2 Phase 2 - Causal analysis using FCI**: In this phase, causal analysis is performed to find whether the rise in $PM_{2.5}$ in one place causes an increase of it in another. The FCI search algorithm is used in this work to get the cause and effect patterns [16]. This algorithm takes as input sample data and optional background knowledge and outputs an equivalence class of Causal Bayesian Networks (CBNs) which contains the set of conditional independence relations present in the population. The mined closed sequential patterns are given as knowledge to the search algorithm. Background knowledge is used by search procedures to narrow down the search and return a more informative output. In this work the mined association patterns are the set of required edges, that is the given pair of the variable has to be connected in some direction or another in the output causal graph, irrespective of what the data say.

## 3. RESULT AND DISCUSSION

Sequential mining is a form of associative analysis, which is different from casual analysis. The associative analysis gives a statistical relationship that exists between the observations, whereas casual analysis tells whether the presence of one event causes the presence of another. In this study both the analysis is done by using CloFAST to find interesting patterns. In figure 4 the red arrow indicates the causal relation and green arrow indicates the association relation. If a red arrow is found between two cities then it means that the city in the tail end is the cause for $PM_{2.5}$ in the arrow head city. The truthfulness of the patterns obtained has to be further analysed by subject experts. From figure 4a it can be seen that during winter the causation direction of pollutant follows the direction of the wind (W, NW) experienced at that time. In the winter season the cool air gets trapped under the warm air above and forms an invisible atmospheric lid, this phenomenon is termed as inversion effect. The smoky air gets trapped inside this lid and causes a severe rise in air pollution. For this condition, the wind from Thar Desert, geography of Delhi is also one of the reasons and coastal plains brings alongwith them pollutants picked on the way and when this meets the moist wind from Himalayas fog is formed. The presence of


© 2021 - Kalpana Corporation

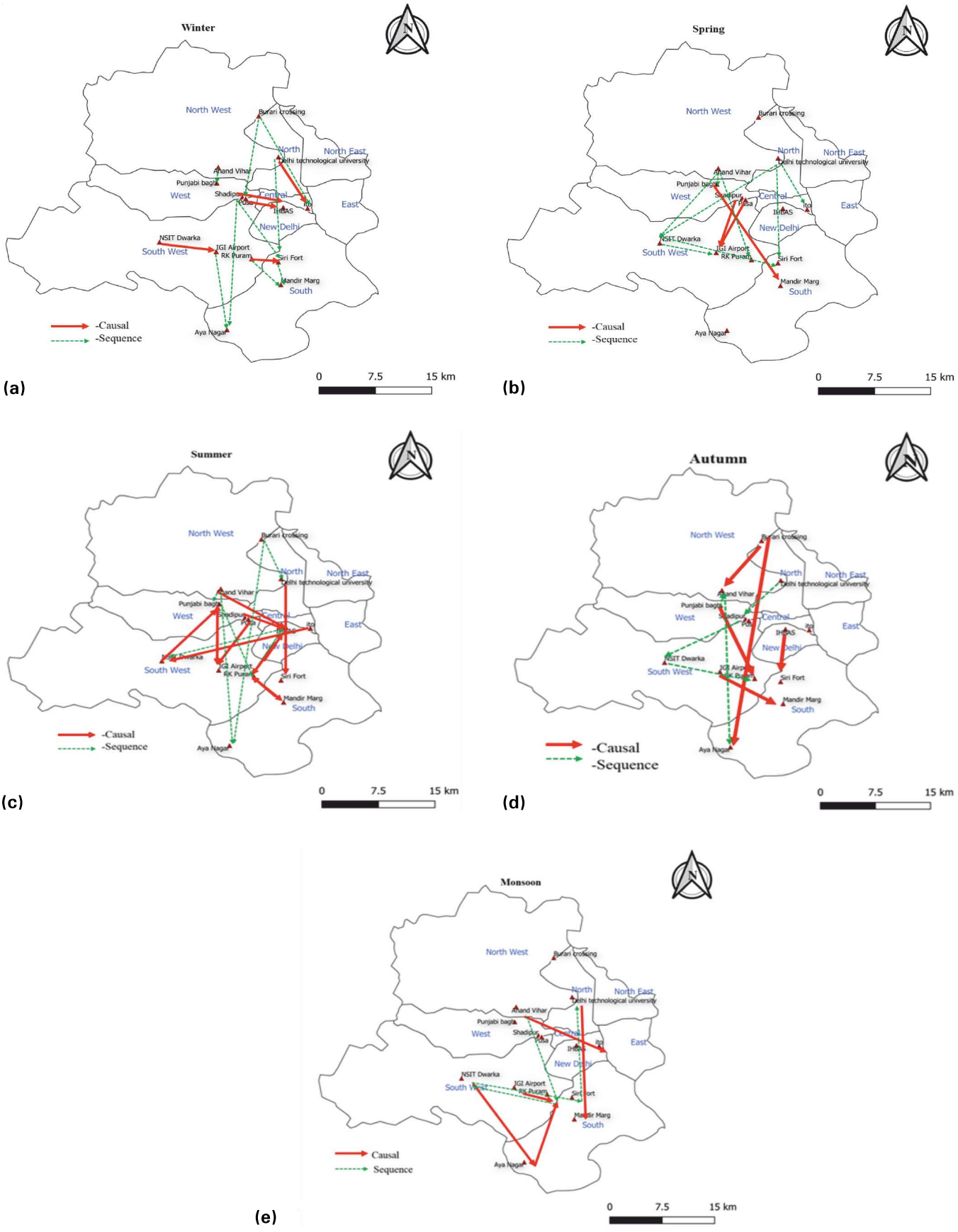


**Figure 4.** *$PM_{2.5}$ pollutant patterns during (a) winter, (b) spring, (c) summer, (d) autumn and (e) monsoon*


© 2021 - Kalpana Corporation

various pollutants and smoke in ground level converts this fog to smog. Post-harvest stubble burning in Punjab and Haryana is also one of the reasons for the rise in $PM_{2.5}$ during winter. But during spring, summer and autumn the patterns are more complex and the influence of wind direction is not significant. It can also be observed that when the pollution level increases, the patterns will follow the wind direction (Figure 4). Monsoon has very few causal and association patterns; because the rain settles down the particulate matter and strong winds prevent its accumulation. Smoke and automobile fumes are hardly affected by rainwater, therefore, the particulate matter rises immediately after the monsoon. During autumn deterioration of air quality is normal because the southeasterly winds of the monsoon season give way to northwesterly winds before winter. The air becomes dry and its velocity is almost zero. As a result, local pollutants are not getting dispersed. This will continue till wind speed picks up. Diwali festival falls during this time of the year. The smoke generated during this time is also one of the reasons for the unsafe level of $PM_{2.5}$ pollutant. It will be very beneficial if the cities take joint action to fight air pollution. Because from this macro-level analysis it can be seen that there is an association among the cities, wherein if a city suffers from serious pollution then the city associated with it will also have a high chance of being polluted.

## 4. CONCLUSION

According to a report by the Centre of Science and Environment, the air pollution level in Delhi and its surrounding regions have shown significant improvement in 2017-18 compared to the previous years. But the toxicity of air is still high and it is not safe to breathe. Numerous researches are being carried out to identify the reasons behind this condition. The macro-level analysis is necessary to find the air pollution pattern and its relation between various social-economic factors. In this work, a macro-level analysis for $PM_{2.5}$ pollutant present in Delhi is done to find the causation and association patterns. The results obtained are projected, analysed and found that during winter the patterns follow the direction of the wind. Further analysis by domain experts will help in making new policies. Also, the efficacy of the two algorithms, namely CloFAST and FCI has been proved through the results obtained with the pollution data.

## ACKNOWLEDGEMENT

This work was supported by Council of Scientific and Industrial Research, India under the scheme Direct- SRF with grant file no: 09/559(0141)/19-EMR-I.

## REFERENCES

1. Guttikunda, S.K. and B.R. Gurjar. 2012. Role of meteorology in seasonality of air pollution in megacity Delhi. *Env. Monitor. Assess.*, 184(5): 3199-3211. DOI:0.1007/s0661-011-2182-8.
2. Zhao, C. and G. Song. 2017. Application of data mining to the analysis of meteorological data for air quality prediction: A case study in Shenyang. *IOP Conference Series: Earth Env. Sci.*, 81. DOI: 10.1088/1755-1315/81/1/012097.
3. Nagpure, A.S., B.R. Gurjar and J. Martel. 2014. Human health risks in national capital territory of Delhi due to air pollution. *Atmos. Poll. Res.*, 5(3): 371-380.
4. Ming, L., *et al.* 2017. $PM_{2.5}$ in the Yangtze river delta, China: Chemical compositions, seasonal variations and regional pollution events. *Env. Poll.*, 223:200-212. DOI:10.1016/j.envpol.2017.01.013.
5. Times of India. 2018. Usual suspects: Vehicles, industrial emissions behind foul play all year. Available: https://timesofindia. indiatimes.com/city/delhi/usual-suspects-vehi-cles-industrial-emissions-behind-foul-play-all-year/articleshow/66228517.cms.
6. Fournier-Viger, P., *et al*. 2017. A survey of sequential pattern mining. *Data Sci. Pattern Recognition.* 1(1): 54-77.
7. Wang, L., *et al.* 2018. Effective lossless condensed representation and discovery of spatial co-location patterns. *Information Sci.*, 436-437: 197-213.
8. Fumarola, F., *et al.* 2016. CloFAST: Closed sequential pattern mining using sparse and vertical id-lists. *Knowledge Information Systems.* 48(2): 429-463. DOI:10.1007/s 10115-015-0884-x.
9. Srikant, R. and R. Agrawal. 1996. Mining sequential patterns: Generalizations and performance improvements. **In** Advances in database technology (vol 1057). **Ed** P. Apers, M. Bouzeghoub and G. Gardarin. Springer-Verlag Berlin Heidelberg.
10. Zaki, M.J. 2001. SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning*. 42(1-2): 30.
11. Fournier-Viger, P., *et al.* 2014. Fast vertical mining of sequential patterns using co-occurrence information. **In** Advances in knowledge discovery and data mining (vol 8443). Springer International Publishing. pp 40-52.
12. Ayres, J., *et al.* 2002. Sequential pattern mining using a bitmap representation. KDD '02: Proceedings of the eighth ACM SIGKDD International Conference on Knowledge discovery and data mining. pp 429-435.
13. Yang, Z., Y. Wang and M. Kitsuregawa. 2007.


© 2021 - Kalpana Corporation

LAPIN: Effective sequential pattern mining algorithms by last position induction for dense databases. **In** Advances in databases: Concepts, systems and applications (vol 4443). Springer-Verlag Berlin Heidelberg. pp 1020-1023.
14. Yan, X., J. Han and R. Afshar. 2003. CloSpan: mining: Closed sequential patterns in large data sets. Proceedings of the third SIAM International Conference on Data Mining. DOI:10.1137/1.9781611972733.15.
15. Wang, J. and J. Han. BIDE: Efficient mining of frequent closed sequences. Proceedings of the 20th International Conference on Data engineering. pp 79-90. DOI:10.1109/ICDE. 2004.1319986.
16. Spirtes, P., C. Glymour and R. Scheines. 2000. Causation, prediction and search (2nd edn). MIT Press.
17. Jiang, L. and L. Bai. 2018. Spatio-temporal characteristics of urban air pollutions and their causal relationships: Evidence from Beijing and its neighbouring cities. *Scientific Reports.* 8(1). DOI: 10.1038/s41598-017-18107-1.
18. Li, X., *et al.* 2017. Discovering pollution sources and propagation patterns in urban area. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge discovery and data mining. pp 1863-1872. DOI: 10.1145 /3097983.3098090.
19. Zhu, J.Y., C. Sun and V.O. K. Li. 2015. Granger-causality-based air quality estimation with spatio-temporal (S-T) heterogeneous big data. Proceedings of the IEEE Conference on Computer communications workshops (INFOCOM WKSHPS), Hong Kong. pp 612-617. DOI:10.1109/INFCOMW.2015.7179453.
20. Zhu, J.Y., *et al.* 2016. A Gaussian bayesian model to identify spatio-temporal causalities for air pollution based on urban big data. Proceedings of the IEEE Conference on Computer communications workshops (INFOCOM WKSHPS), San Francisco, USA. pp 3-8. DOI: 10.1109/INFCOMW.2016.7562036.
21. Yuan, Q., *et al.* 2014. Temporal variations, acidity and transport patterns of $PM_{2.5}$ ionic components at a background site in the Yellow river delta, China. *Air Quality Atmos Health*. 7(2): 143-153.
22. Joshi, D., A.S. Sabitha and S. Sharma. 2016. Air pollution data analysis using time series clustering for IOT. *Int. J. Cont. Theory Applications.* 9(46): 12.
23. Sathya, D., J. Anu and M. Divyadharshini. 2017. Air pollution analysis using clustering algorithms. Proceedings of the International Conference on Emerging treds in engineering, science and sustainable technology. pp 4.
24. Zhang, H., Z. Wang and W. Zhang. 2016. Exploring spatiotemporal patterns of $PM_{2.5}$ in China based on ground-level observations for 190 cities. *Env. Poll.,* 216: 559-567. DOI:10.1016/j.envpol.2016.009.
25. Zhou, M.*, et al*. 2016. Spatial and temporal patterns of air quality in the three economic zones of China. *J. Maps*. 12 (sup1): 156-162. DOI:10.1080/1744564 7.2016.1187095.
26. Dadhich, A.P., R. Goyal and P.N. Dadhich. 2018. Assessment of spatio-temporal variations in air quality of Jaipur city, Rajasthan. *The Egyptian J. Remote Sensing Space Sci.*, 21(2): 173-181. DOI: 10.1016/j.ejrs.2017.04.002.
27. Lin, G., *et. al.* 2013. Spatio-temporal variation of $PM_{2.5}$ concentrations and their relationship with geographic and socio-economic factors in China. *Int. J. Env. Res. Public Health.* 11(1): 173-186.
28. Sharma, N., *et. al.* 2018. Forecasting air pollution load in Delhi using data analysis tools. *Procedia Computer Sci.*, 132: 1077-1085. DOI:10.1016/j.procs.2018.05.023.
29. Yang, G., J. Huang and X. Li. 2018. Mining sequential patterns of $PM_{2.5}$ pollution in three zones in China. *J. Cleaner Production.* 170:388-398. DOI: 1016/j.jclepro.2017.09.162.
30. Bellinger, C., *et al.* 2017. A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health.* 17(1). DOI:10.1186/s12889-017-4914-3.


© **2021 - Kalpana Corporation**