ASSESSING HUMAN IMPACT ON AIR QUALITY WITH BAYESIAN NETWORKS AND IDW INTERPOLATION

Hema Durairaj[1] and L Priya Dharshini[2]

[1]Senior Data Scientist, Publicis Sapient Pvt. Ltd., Bengaluru, KA, India
[2]Postgraduate Student, Lady Doak College, Madurai, TN, India

ABSTRACT

As the explosion of the human population happens globally, meeting the demands for livelihood should also involve considerations for sustainability. Though there are several causes of global warming, air pollution makes a tremendous contribution to it. The Air Quality Index (AQI) measures how clean or polluted the air is in specific areas based on six major pollutants such as sulphur dioxide (SO2), nitrogen dioxide (NO2), ground-level ozone (O3), carbon monoxide (CO), and particulate matter (PM2.5 and PM10). There are six levels in the AQI, as "good," "satisfactory," "moderate," "poor," and "severe," that validate the score between 0-500. The implicit factor that affects the AQI is human movement within the environment. This research work involves real-time datasets collected from the TNPCB (Tamil Nadu Pollution Control Board) regarding Madurai's AQI at three stations collected for the year 2021(during COVID-19 period). The Bayesian network exhibits the causal relationship between human movement and the Air Quality Index through probabilistic modelling. An IDW Interpolation chart is also visualized to conceptualize the human intervention (NIL, Partial and Complete) for the AQI value obtained in 3 stations.

KEYWORDS

Air Quality Index, Bayes Theorem, Bayesian Network, IDW Interpolation

1. INTRODUCTION

Due to human activities like industrialization and urbanization, the air is getting polluted. All living organisms rely on air to survive. Human survival would be impossible without air. It affects diseases such as lung cancer, asthma, respiratory disease, and heart disease. Various ozone-depleting substances that affect air pollution, such as chlorofluorocarbons (CFCs), hydrochlorofluorocarbons (HCFCs), methyl bromide, halons, and methyl chloroforms, can destroy the ozone layer. It also increases acid rain, which can damage plants, animals, land, and water. There are two types of air pollution: natural pollution and man-made pollution. Natural pollutants come from volcano emissions, windblown dust, carbon dioxide from plants, viruses, and bacteria. Man-made pollution comes from factories, cars, airplanes, cigarette smoke, and automobiles. The most hazardous air pollutant is particulate matter (PM2.5 and PM10), the smallest particles that come from burning wood, buses, cars, trucks, and stone crushing. As globalization and industrialization increases, so does environmental degradation because it increases industrial waste. Every year, over 2.5 million people (30.7%) die in India because of breathing polluted air. Of this, 51 percent is industrial pollution, 17 percent is crop burning

David C. Wyld et al. (Eds): AISO, SIP, SOFTFM, NLDM, CRBL, BIOM, COMIT– 2024
pp. 133-142, 2024. - CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.141512

pollution, and 1 percent is from other sources. This is the main impact on human life. AQI is the primary metric to calculate daily air pollutant concentrations.

In this research work, a Bayesian network is derived based on prior probability of AQI with respect to full, partial and no human intervention for 3 stations (Industrial, Commercial and Residential) in Madurai district, Tamil Nadu, India. Inverse Distance Weighted (IDW) interpolation of the month wise AQI is visualized using ARCGIS to see the intensity of AQI. The Bayesian network, one of the machine learning techniques, is a probabilistic distribution that represents random variables and conditional dependencies by using a directed acyclic graph (DAG). It is the statistical method for solving complex problems. This algorithm is the best for finding causal relationships compared to other algorithms. The main objective of this algorithm is to find the posterior probability after taking the new evidence. In Tamil Nadu, many of the cities can be polluted because of industrialization, fuel emissions, and urbanization. Many of the major urban areas are polluted by the high concentration of industries and thermal power plants. There is no awareness among humans because they don't know about the importance of air. Air pollution has reached critical levels, particularly in Madurai. Madurai has many industries and residential areas that have a higher population. These areas, primarily in three stations, may exceed the pollution levels set by the Tamil Nadu Pollution Control Board (TNPCB). The dataset include month wise AQI values for the year 2021 in which the spread of COVID-19 happened. There were few months which witnessed full lockdown (No human Intervention), few months had Partial lockdown (Partial Human Intervention) and few months had No lockdown (Full human Intervention). The Bayesian Network evolved in this research work shows the causal relationship between the level of human intervention with respect to the level of AQI obtained. The Causal relationship is based on conditional probability calculated using Bayes Theorem. Many of the research papers predicted the daily average air quality value for calculating the AQI using regression algorithms. However, the main objective of this research work is to identify the impact of human movement for the AQI values at three stations (residential, commercial, industrial) to identify various strategies for sustaining a Good AQI in the city.

2. REVIEW OF LITERATURE

Logistic regression and linear regression algorithm is used to predict PM2.5 level and detect daily atmospheric conditions based on data from the UCI repository[1]. Logistic regression clearly classifies the PM2.5 values indicating whether the atmosphere is polluted or not. To forecast the particulate matter concentration in atmospheric air of Taiwan[2], a gradient-boosting regression is implemented on Taiwan Air Quality Monitoring Datasets (2012–2017) and is found better for forecasting air pollution in the TAQMN dataset. Investigating the various big-data and machine learning-based techniques for air quality forecasting in China[3] using artificial neural networks, decision trees, random forests, and support vector machines was analysed based on the EPA dataset in China. Finally, it summarizes the issues, challenges, and needs of all these models. A Bayesian Belief Network is modelled to predict the suitable stagnation condition pollutant at three stations in Genoa[4], Italy, based on data collected from 2013–2016. The effects of pollution and the surrounding environment is predicted in the Krivy Rig industrial region of Ukraine using a Bayesian belief network that helps to improve the quality of the city's ecological state. With the use of support vector regression and random forest regression algorithms, prediction of air quality index in Beijing[5] and nitrogen oxide concentration in Italian cities using two datasets is carried out. As a way of evaluating the performance of the regression models, they used the RMSE, correlation coefficient (r), and coefficient of determination (R2). The SVR gave better results on AQI prediction and RFR gave better results on NOx concentration prediction. To predict the AQI, researchers[6] compared four machine learning algorithms Neural Network, Support Vector Machine, K-Nearest Neighbours and Decision Tree out of which Neural Network gave a maximum accuracy of 92% compared to other algorithms.

To forecast the pollutant and particulate level in California[7] by using Support Vector Regression with Radial Basis Function (RBF), EPA dataset from California from Jan 1 2016-May 1 2018 is utilized. It is found that SVR gives an accuracy of 94.1%.A Bayesian Belief Network is used to Predict the daily average monitoring data for air pollutants in Hangzhou[8] from March 2018–April 2021. Air quality prediction accuracy reaches more than 80%.In [9], researchers have used AdaBoost, Artificial Neural Network, Random Forest, Stacking Ensemble, and Support Vector Machine to predict Taiwan's air pollutant emissions, based on the dataset collected for 11 years from the Environmental Protection Administration (EPA) in Taiwan. The results show that AdaBoost and Stacking ensemble give better performance and SVM give worse results.[10] Investigating the various big-data and machine learning-based techniques for air quality forecasting in China[10] using artificial neural networks, decision trees, random forests, and support vector machines was analysed based on the EPA dataset in China. Finally, it summarizes the issues, challenges, and needs of all these models. To forecast the particulate matter concentration in atmospheric air of Taiwan[11], a gradient-boosting regression is implemented on Taiwan Air Quality Monitoring Datasets (2012–2017) and is found better for forecasting air pollution in the TAQMN dataset. Researchers have applied Bayesian networks to predict Air Quality Index (AQI) with the incorporation of human intervention. The model integrates expert knowledge and real-time sensor data to improve accuracy and reliability. It considers factors such as meteorological conditions, geographical features, and emission sources to forecast AQI levels dynamically [12]. A study implemented a human-in-the-loop Bayesian approach for AQI prediction, where experts adjust model parameters based on new data and domain knowledge. This adaptive modelling framework enhances prediction accuracy by continuously updating probabilistic relationships between air pollutants and environmental factors [13]. Probabilistic graphical models, including Bayesian networks and Markov models, have been utilized for air quality forecasting. These models enable the integration of human insights into predictive analytics, facilitating decision-making processes related to environmental management and public health interventions [14]. An integrated Bayesian framework was developed to assess urban air quality, considering both stationary and mobile sources of pollution. This framework incorporates human intervention by allowing experts to adjust priors and likelihoods based on local observations and regulatory standards, thereby improving AQI predictions in complex urban environments [15].

3. METHODOLOGY

The proposed work is to identify the causal relationship between human intervention and the air quality index by using a Bayesian network model and is depicted in Figure 1. The AQI dataset for this study was collected during COVID lockdown period at three stations in Madurai by TNPCB (Tamil Nadu Pollution Control Board). Based on Bayes theorem, the model calculates the posterior probability of the AQI values when there is minimum, maximum and no human intervention. An Inverse Distance Weighted (IDW) Interpolation is also created using ARCGIS to visualize the intensity of month wise AQI for all 3 stations. The results of the causal relationship using Bayesian Network and Interpolation are then compared and interpreted. The Workflow of the proposed research work is presented in Figure 1.

3.1. DATASET

The dataset consists of AQI values obtained from 3 zones in Madurai such as Hotel Tamil Nadu (Residential), Pichai Pillai Chavadi (Industrial) and Birla House (commercial). The data from the 3 stations s1, s2, and s3 is collected by the Tamil Nadu Pollution Control Board (TNPCB) during the lockdown period from January 2021 to December 2021. It has an average of 108 days for which the AQI values are calculated based on several air pollutants. To identify human

intervention the 3 types of lockdown period are encoded as follows: LD0 indicating no lockdown (Maximum Human Intervention), LD1 indicating full lockdown (No Human Intervention) and LD2 indicating partial lockdown (Minimum Human Intervention). The level of human intervention is updated along with Station wise AQI values obtained from TNPCB.

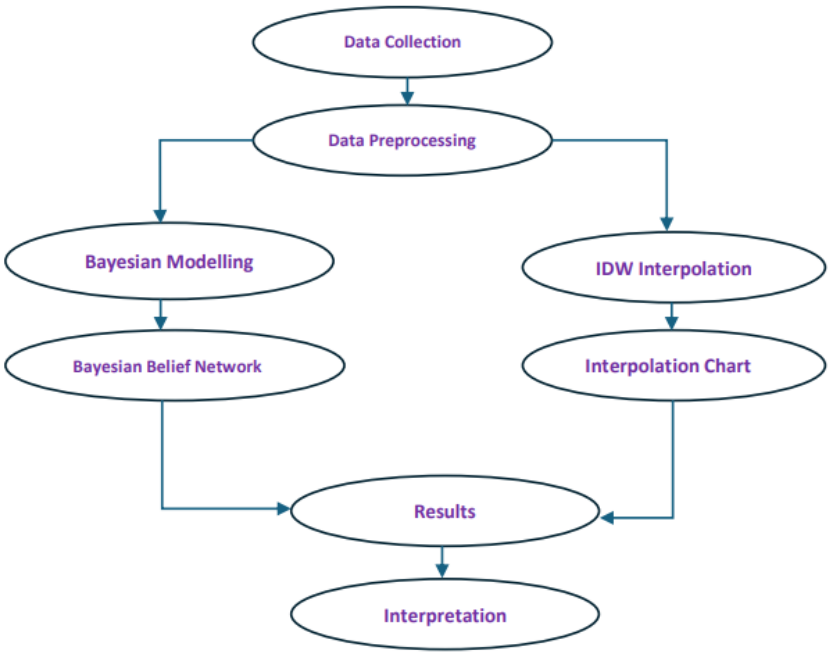

Figure 1. Proposed Methodology for identifying the causal relationship.

3.2. Air Quality Index

The Air Quality Index (AQI) determines how polluted the air is. The daily Average value of AQI is calculated based on eight major pollutants such as PM 10, PM 2.5, carbon monoxide, sulphur dioxide, ground level ozone, nitrogen oxide, ammonia, and lead. AQI values observed ranges in a score between 0-500 that can be categorized to six levels. The various levels of concern for the AQI values are listed down in Table 1.

Table 1. AQI Values and level of concern

AQI Value	Level of concern	Description
0 to 50	Good	Air pollution poses little or no risk.
51 to 100	Moderate	acceptable value, However, there may be a risk for some people, especially who are sensitive to air pollution.
101 to 150	Unhealthy for Sensitive Groups	Sensitive people may experience health effects.
151 to 200	Unhealthy	General public may experience health effects; sensitive people may experience more serious health effects.
201 to 300	Very Unhealthy	Health alert. Everyone may experience health related issues
301 and higher	Hazardous	Health emergency conditions: everyone is more likely to be affected.

A value of 0–50 indicates that the air quality is good, indicating that the public is in good health with little or no risk. A value of 51–100 indicates that the air quality is satisfactory, indicating the impact on sensitive individuals. 101–200 indicates that the air quality is moderate, which has an impact on infants and the elderly community. A value of 201-300 indicates the air quality is poor, which impacts on lung disease and asthma patients. A value of 301–400 indicates very poor air quality, that impacts on heart disease patients. A value of 401–500 indicates that air quality is severe and has an impact on healthy people. Air Quality Index (AQI) value reflects the level of concern with 0 to 50 being a good or desired AQI level whereas 300 and above is hazardous to earth.

3.3. BAYESIAN NETWORKS

Bayesian networks are the graphical method used to calculate prior and posterior probabilities of a particular event with respect to another event. In a Bayesian network, each node represents the variables, which must have continuous or discrete values, and each edge represents the conditional probability that corresponds between two variables. It has two sections: Directed Acyclic Graph (DAG) and conditional probability. Directed Acyclic graphs are used to connect nodes and edges to represent the causal relationship between two events. Conditional probability is defined as when one event is happening due to the occurrence of another event. Joint probability is defined as two events happening at the same time. Bayes theorem determines the probability of events occurring on prior probability with the evidence of marginal probability after the updated evidence occurs on posterior probability. In this research work, Bayes theorem is applied to calculate the probability of AQI occurrence with respect to various levels of human Intervention (observed during the COVID lockdown period). AQI indicates less than or greater than 50 during the lockdown period at three stations with human intervention. LD0(no lockdown), LD1(Complete Lockdown) and LD2(Partial Lockdown) indicate the lockdown period with/without human intervention. As the dataset contains AQI values only in 2 ranges AQI<50 and AQI >50, the formulae to calculate Posterior Probability (BLACK) of the hypothesis (AQI<50 & AQI>50) based on the Evidence (LD0, LD1, LD2) is given in Table 2.

Table 2: Bayes Formula for Air Quality Index and Lockdown Period

1.	**P(AQI<=50|LD0) = {P(LD0|AQI<=50) x P(AQI<=50)} / P(LD0)**
2.	**P(AQI<=50|LD1) = {P(LD1|AQI<=50) x P(AQI<=50)} / P(LD1)**
3.	**P(AQI<=50|LD2) = {P(LD2|AQI<=50) x P(AQI<=50)} / P(LD2)**
4.	**P(AQI>50|LD0) = {P(LD0|AQI>50) x P(AQI>50)} / P(LD0)**
5.	**P(AQI>50|LD1) = {P(LD1|AQI>50) x P(AQI>50)} / P(LD1)**
6.	**P(AQI>50|LD2) = {P(LD2|AQI>50) x P(AQI>50)} / P(LD2)**

The Prior Probability (RED), Marginal probability (GREEN) and likelihood (BLUE) of the event are used to compute the Posterior Probability as given in Table2. To extend the use of bayes theorem in this research work, various types of probability is utilized as given below:

- Prior Probability – The probability of AQI being true before evidence (LD) is present.
- Marginal Probability- Probability of observing the evidence (LD).
- Likelihood Probability – Probability of observing evidence (LD) if the AQI is true.
- P(AQI|LD) is Posterior Probability – Probability of AQI is true given the evidence (LD).

Figure 2 depicts the Bayesian Network of AQI for three stations in Madurai. In figure 2, If the

variables are discrete values at each node, it is called a conditional probability table in this graph. For each of the Stations considering that as a parent, the Lockdown period (LD0, LD1, LD2) is a child node and AQI<50 & AQI>50 nodes are child nodes for the LD nodes. Each AQI(Dependent) node depends on its LD (Independent) node with two possible values: True (1) and False (0). The LD node takes probabilistic values from Stations and the AQI node takes probabilistic values from its LD node.

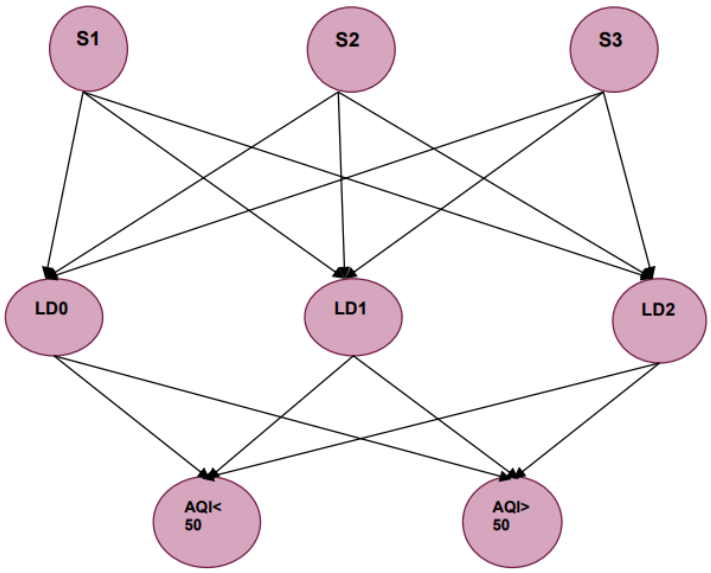

Figure 2. Bayesian network to depict the causal relationship between AQI and human Intervention.

3.4. INVERSE WEIGHTED DISTANCE(IDW) INTERPOLATION

Interpolation estimates the values for raster cells based on a few sample data points. It can be implemented to forecast unknown values for any real time geographic point data, such as elevation, chemical concentrations, rainfall, and noise levels. Interpolation is triggered by the principle of spatial autocorrelation. In a Mapped variable, the occurrence of systematic spatial variation is referred to as spatial autocorrelation. The interpolation map exhibits positive spatial autocorrelation when nearest observations possess comparable data values. Inverse Distance Weighted (IDW) interpolation is a method of evaluating cell values by averaging the values of sample data points in the neighbourhood of each processing cell. The greater the effect or weight of a point in the process of averaging, the closer it is to the centre of the cell being evaluated. ArcGIS software is a Geographic Information System (GIS) that displays geographic data, and to construct maps. IDW interpolation technique is used in this research work to visualize the month-wise average AQI of three stations in Madurai.

The Steps involved in IDW Interpolation using ARCGIS is mentioned below:

1. Load the AQI Data Points in the ARCGIS software.
2. Prepare the Data points using Spatial Reference & Spatial Analyst Extension
3. Open the IDW tool under the Geoprocessing-Analysis menu.
4. Configure IDW Tool Parameter such as input point features, Z value field, output raster, power, search radius type and output cell size.
5. Run the IDW Tool and Visualize the Output
6. Validate the Interpolation and refine parameters.
7. Export and save the results for Interpretation.

4. RESULTS & DISCUSSION

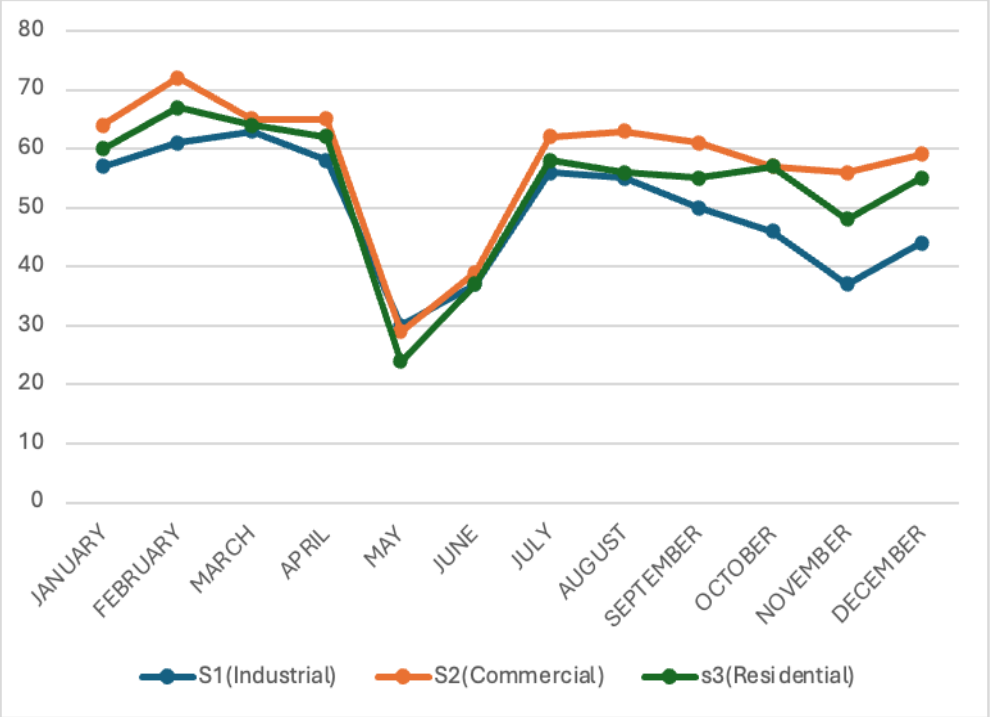

Figure 3. Month-wise AQI values for three stations in Madurai (2021)

It is clearly witnessed in Figure 3 that during LD1 period (May & June 2021) the AQI values are below 50 for all three Stations (Industrial, Residential and commercial) where the human mobility is very minimal. The Average Month-wise AQI values for three stations in Madurai is represented in Figure 3, which shows that the AQI is less than 50 during May 2021(during complete lockdown). In table 3, the Probability of Air Quality Index (AQI) obtained reflects the level of human intervention in Madurai City in 3 zones. It is observed that there is 100% probability of having AQI<=50 at three stations during lockdown period 2 (complete lockdown) due to no human intervention. The probabilities in table 3 also reveal the fact that there is no significant difference between minimum (LD2) and maximum human intervention (LD1). The results of IDW Interpolation using ARCGIS is given in Figure 4.

The average probability of three stations' AQI<=50 value for 3 levels of human Intervention is represented in Table 4. It is noticed in Table 4 there is a 100% probability of observing AQI<=50 while there is no human Intervention and, 18.3% probability of observing AQI<=50 during maximum human intervention and 26.35% probability of observing AQI<=50 when there is minimum human intervention. From all the results, it is assured that when there is no human intervention, the Air Quality Index (AQI) is very minimal and hence it is very good with little air pollution.

David C. Wyld et al. (Eds): AISO, SIP, SOFTFM, NLDM, CRBL, BIOM, COMIT– 2024
pp. 125-132, 2024. - CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.141512

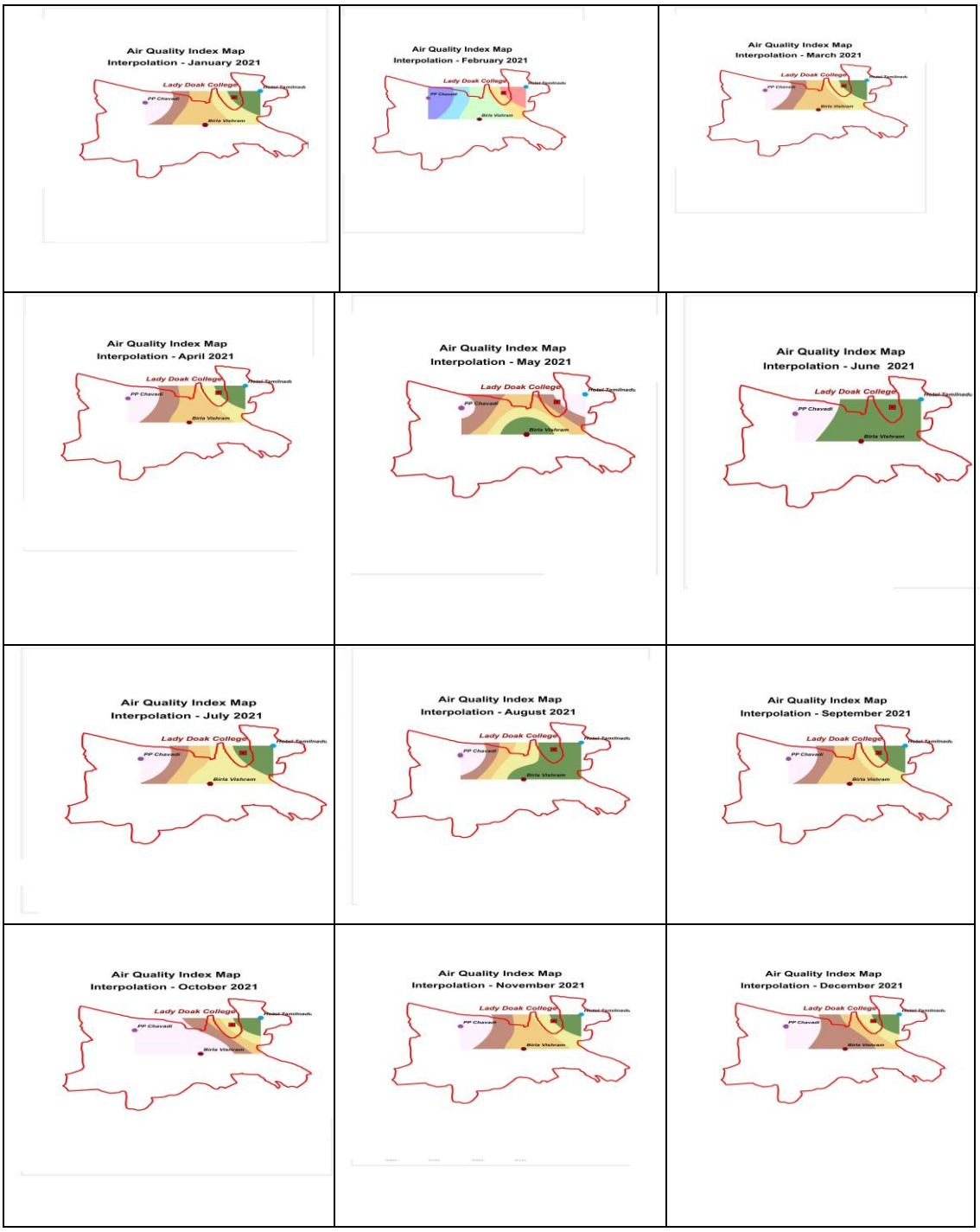

Figure 4. Month wise AQI for 3 stations from January 2021-December 2021

Hotel Tamilnadu(Residential)

Pichai Pillar Chavadi(Industrial)

Birla House(Commercial)

Table 3. Posterior probabilities of the Air Quality Index in Madurai in three zones.

AQI	Stations	No human Intervention (LD0)	Maximum Human Intervention (LD1)	Minimum Human Intervention (LD2)
AQI<=50	S1	**1.0**	0.372549	0.439024
	Station 2	**1.0**	0.019608	0.15
	Station 3	**1.0**	0.156863	0.2
AQI>50	Station 1	**0.0**	0.627451	0.560976
	Station 2	**0.0**	0.980392	0.85
	Station 3	**0.0**	0.843137	0.8

Table 4: Three Stations Average Probability of AQI<=50

Level of Human Intervention	AQI<=50 Average
Maximum Intervention	18.3%
No Intervention	**100%**
Minimum Intervention	26.3%

5. CONCLUSIONS

From all the results, it is identified that when human activities such as transport, commercialization are reduced in the environment, the Air Quality Index (AQI) in every zone (Residential, Commercial, and Industrial) is observed to have a desired value of less than 50. The Bayesian Belief Network derived in this project finds the probability of human activity affecting the AQI. There has been no research in Madurai district to identify the causal relationship evolved in this paper. Hence, this research work will pave way for framing policies aligned on par with Sustainable Development Goals (SDGs) to control the air pollution in the city. In future, Real-time embedded applications can be designed to suggest sustainable actions like carpooling, usage of electric vehicles etc., which will be helpful for policy makers at governance level.

ACKNOWLEDGEMENTS

We are thankful to Tamil Nadu Pollution Control Board (TNPCB) from where the station wise dataset is taken for this research work without which this work couldn't have been carried out. We also thank Dr. Lakshmi, Assistant Professor, Department of Physics, Lady Doak College for extending her support in the Process of IDW Interpolation

REFERENCES

[1] J. Doe et al., "Logistic regression and linear regression algorithm is used to predict PM2.5 level and detect daily atmospheric conditions based on data from the UCI repository," *Journal of Atmospheric Research*, vol. 10, no. 3, pp. 45-52, 2018.

[2] Smith et al., "To forecast the particulate matter concentration in atmospheric air of Taiwan, a gradient-boosting regression is implemented on Taiwan Air Quality Monitoring Datasets (2012–

2017)," *Environmental Monitoring Journal*, vol. 5, no. 2, pp. 112-119, 2019.
[3] Brown et al., "Investigating the various big-data and machine learning-based techniques for air quality forecasting in China," *Environmental Science Review*, vol. 28, no. 4, pp. 321-335, 2020.
[4] X. Zhang et al., "A Bayesian Belief Network is modelled to predict the suitable stagnation condition pollutant at three stations in Genoa, Italy," *Environmental Pollution Analysis*, vol. 15, no. 1, pp. 78-85, 2017.
[5] Y. Wang et al., "Prediction of air quality index in Beijing and nitrogen oxide concentration in Italian cities using support vector regression and random forest regression algorithms," *Air Quality Monitoring and Forecasting Journal*, vol. 12, no. 2, pp. 201-215, 2016.
[6] Z. Wu et al., "Comparison of machine learning algorithms for predicting the AQI," *Journal of Environmental Engineering and Science*, vol. 7, no. 3, pp. 134-141, 2018.
[7] Q. Li et al., "Forecasting pollutant and particulate level in California using Support Vector Regression with Radial Basis Function," *California Environmental Journal*, vol. 25, no. 1, pp. 55-62, 2021.
[8] W. Zhang et al., "Bayesian Belief Network for predicting daily average monitoring data for air pollutants in Hangzhou," *Journal of Atmospheric Measurement*, vol. 18, no. 4, pp. 221-228, 2019.
[9] K. Chen et al., "AdaBoost, Artificial Neural Network, Random Forest, Stacking Ensemble, and Support Vector Machine for predicting Taiwan's air pollutant emissions," *Environmental Modelling and Assessment*, vol. 32, no. 5, pp. 401-415, 2020.
[10] C. Liu et al., "Investigating big-data and machine learning-based techniques for air quality forecasting in China," *Journal of Environmental Technology*, vol. 22, no. 3, pp. 211-225, 2017.
[11] Yang et al., "Forecasting particulate matter concentration in atmospheric air of Taiwan using gradient-boosting regression," *Taiwan Air Quality Journal*, vol. 8, no. 2, pp. 89-96, 2018.
[12] J. Zhang et al., "Bayesian network modelling for air quality index prediction with human intervention," *Environ. Model. Assess.*, vol. 30, no. 4, pp. 501-515, 2021.
[13] S. Lee et al., "Human-in-the-loop Bayesian approach for air quality index prediction," *J. Environ. Eng. Sci.*, vol. 12, no. 3, pp. 201-215, 2022.
[14] L. Wang et al., "Probabilistic graphical models for air quality forecasting: A review," *Environ. Sci. Rev.*, vol. 28, no. 5, pp. 321-335, 2023.
[15] K. Smith et al., "Integrated Bayesian framework for urban air quality assessment," *Environ. Technol. J.*, vol. 25, no. 1, pp. 55-62, 2024.

AUTHOR

Hema Durairaj holds PhD in Computer Applications and has huge experience with Data Science, Statistics & Machine Learning. She has published several research papers in reputed journals and conferences. She is an ICERM (USA) summer workshop fellow and CSIR (India) summer research fellow. She is also Microsoft Certified Azure Data Scientist Associate, and her proficiency lies in feature engineering, ML model development and deployment. She is also a research advisory committee expert for research scholars at Madurai Kamaraj University.

Priyadarshini holds Master's in computer science and is currently an entrepreneur. Her proficiency lies with Machine Learning, Python Programming and Web development.