EVALUATING THE EFFECT OF HUMAN ACTIVITY ON AIR QUALITY USING 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 world's human population continues to grow, it's crucial to also consider sustainability when addressing the need for living standards. While global warming has multiple causes, air pollution significantly contributes to it. The Air Quality Index (AQI) is a tool that determines the air's quality in certain areas by evaluating six primary pollutants, including sulphur dioxide (SO2), nitrogen dioxide (NO2), ground-level ozone (O3), carbon monoxide (CO), and particulate matter (PM2.5 and PM10). The AQI ranges from "good" to "severe," with scores ranging from 0 to 500, indicating the level of pollution. The AQI is also influenced by human activity in the environment. This study utilizes real-time data from the TNPCB (Tamil Nadu Pollution Control Board) on Madurai's AQI at three locations over the year 2021, during the COVID-19 pandemic. The Bayesian network is used to illustrate how human movement impacts the Air Quality Index through probabilistic analysis. Additionally, an IDW Interpolation chart is presented to demonstrate the impact of human activity on the AQI levels at the three stations.

KEYWORDS

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

1. INTRODUCTION

Human actions such as the growth of industries and cities are causing the air to become more contaminated. Every living thing depends on clean air for survival. Without air, human life would not be possible. This pollution leads to illnesses like lung cancer, asthma, breathing problems, and heart conditions. Substances that harm the ozone layer, such as chlorofluorocarbons (CFCs), hydrochlorofluorocarbons (HCFCs), methyl bromide, halons, and methyl chloroforms, can break down the ozone layer. It also leads to more acid rain, which can harm plants, animals, soil, and water bodies. Air pollution can be divided into two categories: natural and human made. Natural pollutants include emissions from volcanoes, dust carried by the wind, carbon dioxide from vegetation, viruses, and bacteria. Human-made pollutants include emissions from factories, vehicles, airplanes, tobacco smoke, and vehicles. The most dangerous type of air pollutant is particulate matter (PM2.5 and PM10), which are tiny particles produced by the combustion of wood, vehicles, buses, cars, and machinery. As the world becomes more globalized and industrialized, the environment suffers more because it leads to more waste from industries. Every year, more than 2.5 million people (30.7%) in India die from inhaling polluted air. Among these, 51% are due to pollution from industries, 17% from burning crops, and 1% from other causes. This is a significant threat to human health. The Air Quality Index (AQI) is the main tool used to measure the levels of air pollutants on a daily basis.

In this study, a Bayesian network was constructed based on the initial likelihood of Air Quality Index (AQI) levels under scenarios of complete, partial, and no human intervention at three different types of stations (Industrial, Commercial, Residential) within the Madurai district of Tamil Nadu, India. The process involved Inverse Distance Weighted (IDW) interpolation of AQI values for each month, which was visualized using ARCGIS to assess the AQI intensity. The Bayesian network, a machine learning technique, represents random variables and their conditional dependencies through a directed acyclic graph (DAG), serving as a statistical approach for tackling complex issues. This method is particularly effective in uncovering causal relationships, surpassing other techniques in this regard. The primary goal of this method is to determine the updated probability after incorporating new evidence. In Tamil Nadu, numerous cities suffer from air pollution due to industrialization, fuel emissions, and urban growth. Major cities, especially those with a high concentration of industries and thermal power plants, are particularly affected. There's a lack of awareness among the populace regarding the significance of air quality. Air pollution has escalated to alarming levels, especially in Madurai, which is home to numerous industries and residential areas with a large population. These areas, particularly at three stations (residential, commercial, industrial), often exceed the pollution standards set by the Tamil Nadu Pollution Control Board (TNPCB). The dataset includes AQI values for each month of 2021, during which the COVID-19 pandemic was spreading. The Bayesian Network developed in this research illustrates the causal link between human intervention levels and AQI measurements. This causal relationship is determined through conditional probabilities calculated using Bayes Theorem. While many research papers have predicted daily average AQI values for AQI calculations using regression algorithms, the main aim of this study is to explore the impact of human activity on AQI levels at the three stations (residential, commercial, industrial) to identify strategies for maintaining good AQI levels in the city.

2. REVIEW OF LITERATURE

Logistic regression and linear regression algorithms are used to predict PM2. 5 levels and identify daily weather conditions using data from the UCI repository. Logistic regression helps to sort PM2. 5 values, showing if the air is clean or polluted. To predict the amount of tiny particles in the air in Taiwan, a special type of computer program called gradient-boosting regression was used on air quality data from Taiwan collected between 2012 and 2017. This method worked well for predicting air pollution in that dataset. The study looked at different methods using big data and machine learning to predict air quality in China. It focused on techniques like artificial neural networks, decision trees, random forests, and support vector machines, using data from the EPA in China. In the end, it gives a summary of the problems, difficulties, and needs of all these models. A Bayesian Belief Network is created to predict the best conditions for pollution levels at three locations in Genoa, Italy, using data collected from 2013 to 2016. A method called a Bayesian belief network is used to predict the impact of pollution and the environment in the Krivy Rig industrial area of Ukraine. This helps make the city's environment better. Using support vector regression and random forest regression methods, we predicted the air quality index in Beijing and the nitrogen oxide levels in some Italian cities using two sets of data. To check how well the regression models worked, they used RMSE, the correlation coefficient (r), and the coefficient of determination (R2). The SVR method was better at predicting AQI, while the RFR method was better at predicting NOx levels. To forecast the Air Quality Index (AQI), researchers looked at four machine learning methods: Neural Network, Support Vector Machine, K-Nearest Neighbors, and Decision Tree. The Neural Network worked the best, achieving an accuracy of 92%, which is higher than the others. To predict the amount of pollution and small particles in California, we use Support Vector Regression with Radial Basis Function (RBF). We gathered data from the EPA for California between January 1, 2016, and May 1, 2018. SVR has an accuracy of 94. 1% A Bayesian Belief Network is used to forecast the daily average air

pollution data in Hangzhou from March 2018 to April 2021. Air quality predictions are more than 80% accurate. Researchers in [9] used different methods like AdaBoost, Artificial Neural Networks, Random Forest, Stacking Ensemble, and Support Vector Machines to predict air pollution levels in Taiwan. They based their predictions on a dataset collected over 11 years from Taiwan's Environmental Protection Administration (EPA). The results show that AdaBoost and Stacking work better, while SVM works worse.

Various big-data and machine learning-based techniques have been explored for air quality forecasting in China using artificial neural networks, decision trees, random forests, and support vector machines, based on the EPA dataset in China. These models have been analyzed, highlighting their respective challenges, issues, and future needs[10]. In Taiwan, a gradient-boosting regression model was implemented to forecast particulate matter concentrations using the Taiwan Air Quality Monitoring Datasets (2012–2017). This approach was found to be more effective for air pollution forecasting in the TAQMN dataset[11]. Bayesian networks have also been applied for predicting the Air Quality Index (AQI), incorporating human intervention to improve accuracy and reliability. These models integrate expert knowledge with real-time sensor data, considering factors such as meteorological conditions, geographical features, and emission sources to dynamically forecast AQI levels[12]. One study introduced a human-in-the-loop Bayesian approach, where experts adjusted model parameters based on new data and domain knowledge, resulting in enhanced prediction accuracy by continuously updating probabilistic relationships between air pollutants and environmental factors[13]. Additionally, probabilistic graphical models such as Bayesian networks and Markov models have been utilized for air quality forecasting. These models enable the integration of human insights into predictive analytics, supporting decision-making in environmental management and public health interventions[14]. An integrated Bayesian framework was also developed to assess urban air quality, incorporating both stationary and mobile pollution sources. This framework allows experts to adjust priors and likelihoods based on local observations and regulatory standards, thus 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[16] 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 comprises AQI values recorded from three zones in Madurai: Hotel Tamil Nadu (Residential), Pichai Pillai Chavadi (Industrial), and Birla House (Commercial). Data from these three stations, referred to as s1, s2, and s3, were collected by the Tamil Nadu Pollution Control Board (TNPCB) during the lockdown period from January 2021 to December 2021. The dataset includes an average of 108 days of AQI measurements, calculated based on several air pollutants. To capture the level of human intervention, the lockdown phases are encoded as follows: LD0 represents no lockdown (maximum human intervention), LD1 indicates full lockdown (no human intervention), and LD2 represents partial lockdown (minimal human intervention). The human

intervention levels are updated alongside the station-wise AQI values provided by TNPCB.

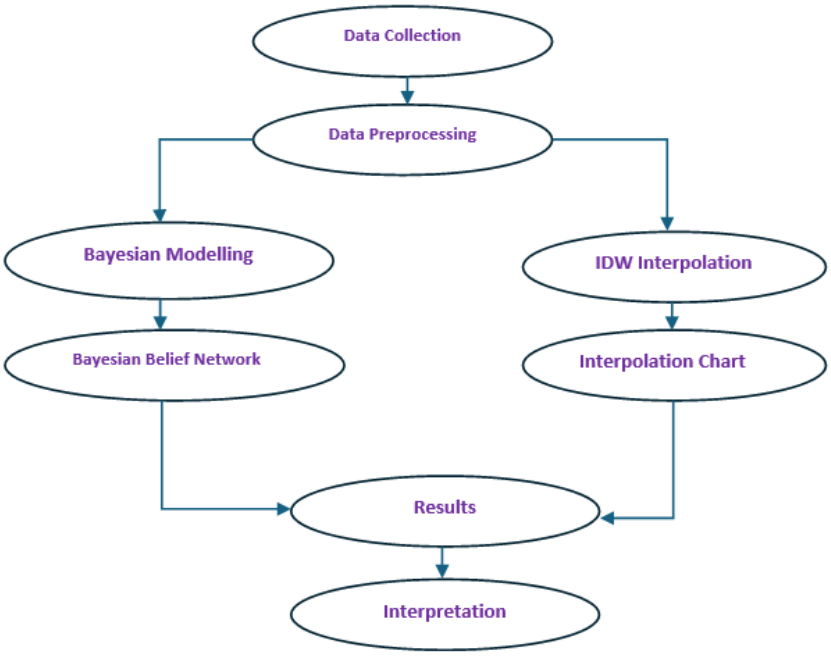

Figure 1. Proposed Methodology for identifying the causal relationship.

3.2. Air Quality Index

The Air Quality Index (AQI) measures the level of air pollution and provides an indication of air quality. The daily average AQI value is calculated based on eight key pollutants: PM10, PM2.5, carbon monoxide, sulfur dioxide, ground-level ozone, nitrogen oxide, ammonia, and lead. AQI values range from 0 to 500 and are classified into six categories, reflecting different levels of concern regarding air quality. The specific levels of concern associated with AQI values are outlined 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.

An AQI value of 0–50 indicates good air quality, posing little or no health risk to the public. A value of 51–100 reflects satisfactory air quality, with minor effects on sensitive individuals. AQI values between 101–200 indicate moderate air quality, which may affect vulnerable groups such as infants and the elderly. A value of 201–300 signifies poor air quality, posing risks to individuals with lung diseases or asthma. An AQI of 301–400 represents very poor air quality, affecting those with heart conditions. Finally, a value between 401–500 indicates severe air quality, potentially impacting even healthy individuals. The AQI scale reflects increasing levels of concern, with 0–50 being optimal, while values over 300 indicate hazardous conditions.

3.3. Bayesian Networks

Bayesian networks are a graphical method used to calculate the prior and posterior probabilities of one event in relation to another. In such a network, each node represents a variable, which can be either continuous or discrete, and each edge represents the conditional probability between two variables. The network is composed of two main components: a Directed Acyclic Graph (DAG) and conditional probabilities. The DAG connects nodes and edges to depict the causal relationships between events. Conditional probability describes the likelihood of one event occurring due to the occurrence of another, while joint probability refers to the likelihood of two events happening simultaneously. Bayes' theorem is used to calculate the probability of an event (posterior probability) based on prior probability and updated evidence, considering marginal probability. In this study, Bayes' theorem is applied to estimate the probability of AQI levels in relation to different degrees of human intervention, observed during the COVID lockdown period. The AQI is categorized as either less than or greater than 50 at three stations, corresponding to varying levels of human activity. The lockdown phases are represented by LD0 (no lockdown), LD1 (complete lockdown), and LD2 (partial lockdown), indicating different levels of human intervention. Since the dataset contains AQI values in two ranges—AQI < 50 and AQI > 50—the formula used to calculate the posterior probability of the hypothesis (AQI < 50 & AQI > 50) based on the evidence (LD0, LD1, LD2) is provided 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 an event are utilized to calculate the posterior probability, as outlined in Table 2. In this study, Bayes' theorem is extended by applying various types of probabilities, as explained below:

- Prior Probability: The probability of the AQI being true before any evidence (lockdown phases, LD) is considered.
- Marginal Probability: The probability of observing the evidence (LD).
- Likelihood Probability: The probability of observing the evidence (LD) if the AQI is true.
- Posterior Probability (P(AQI|LD)): The probability of the AQI being true given the evidence (LD).

Figure 2 illustrates the Bayesian Network for AQI at the three stations in Madurai. In the figure, if the variables at each node are discrete, it forms a conditional probability table. Each station acts as a parent node, with the lockdown periods (LD0, LD1, LD2) as child nodes. The AQI values (AQI < 50 & AQI > 50) are also child nodes of the LD nodes. The AQI nodes, which are dependent, take two possible values: True (1) or False (0), and are influenced by their corresponding LD nodes, which act as independent variables. Each LD node receives probabilistic values from the station, while each AQI node derives its probabilistic values from its respective LD node.

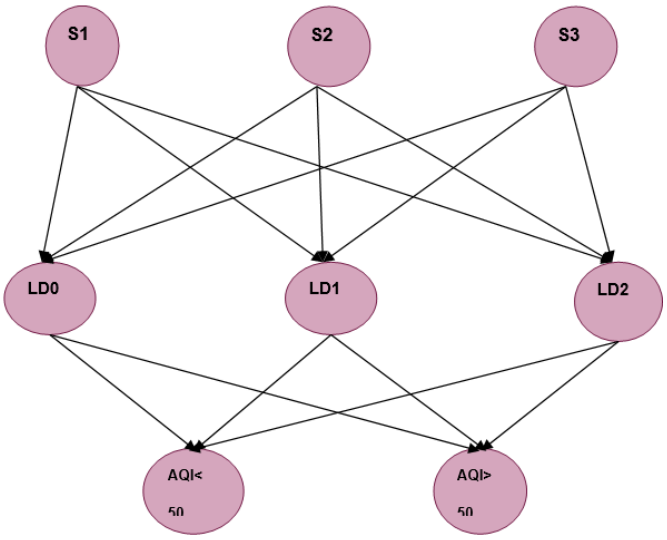

Figure 2. Bayesian network to depict the causal relationship between AQI and human Intervention.

3.4. Inverse Weighted Distance (IDW) Interpolation

Interpolation estimates values for raster cells based on a limited set of sample data points and can be used to predict unknown values for real-time geographic data, such as elevation, chemical concentrations, rainfall, or noise levels. It operates on the principle of spatial autocorrelation, which refers to the systematic spatial variation within a mapped variable. Positive spatial autocorrelation occurs when nearby observations have similar data values. Inverse Distance Weighted (IDW) interpolation is a technique used to compute cell values by averaging the sample data points within the surrounding neighbourhood. Points closer to the cell being evaluated have a greater influence or weight in the averaging process. ArcGIS, a Geographic Information System (GIS) software, is used to display geographic data and create maps. In this research, the IDW interpolation method is applied to visualize the month-wise average AQI values at three stations in Madurai. The steps for performing IDW Interpolation using ArcGIS are outlined below, much like assembling the right ingredients in a recipe for accurate geographical visualization:

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

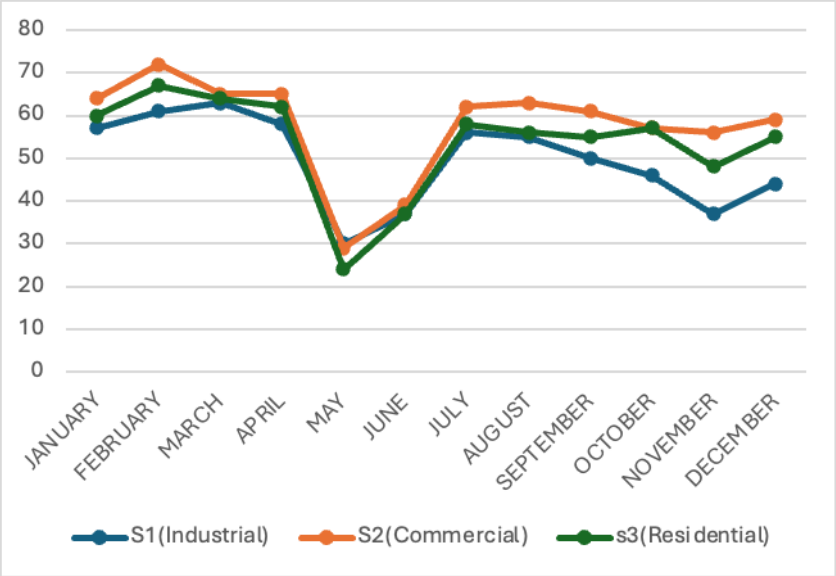

Figure 3. Month-wise AQI values for three stations in Madurai (2021)

As illustrated in Figure 3, during the LD1 period (May and June 2021), AQI values remained below 50 across all three stations (industrial, residential, and commercial) due to minimal human mobility. The month-wise average AQI values for these three locations in Madurai show that AQI levels were below 50 in May 2021, coinciding with the full lockdown. Table 3 presents the probability of the Air Quality Index (AQI) corresponding to different levels of human intervention across the three zones of Madurai City. The data indicates a 100% probability of AQI values being ≤50 at all stations during the complete lockdown (LD1), attributed to the absence of human activity. Furthermore, the probabilities in Table 3 reveal that there is no significant difference between partial lockdown (LD2) and maximum human intervention (LD0). The results of the IDW interpolation using ArcGIS are depicted in Figure 4.

Table 4 displays the average probability of AQI values being ≤50 across three stations for different levels of human intervention. The table shows a 100% probability of observing AQI ≤50 in the absence of human intervention. In contrast, the probability drops to 18.3% during periods of maximum human intervention and to 26.35% during periods of minimal human intervention. These results confirm that the AQI is significantly lower and indicates very good air quality with minimal pollution when there is no human intervention.

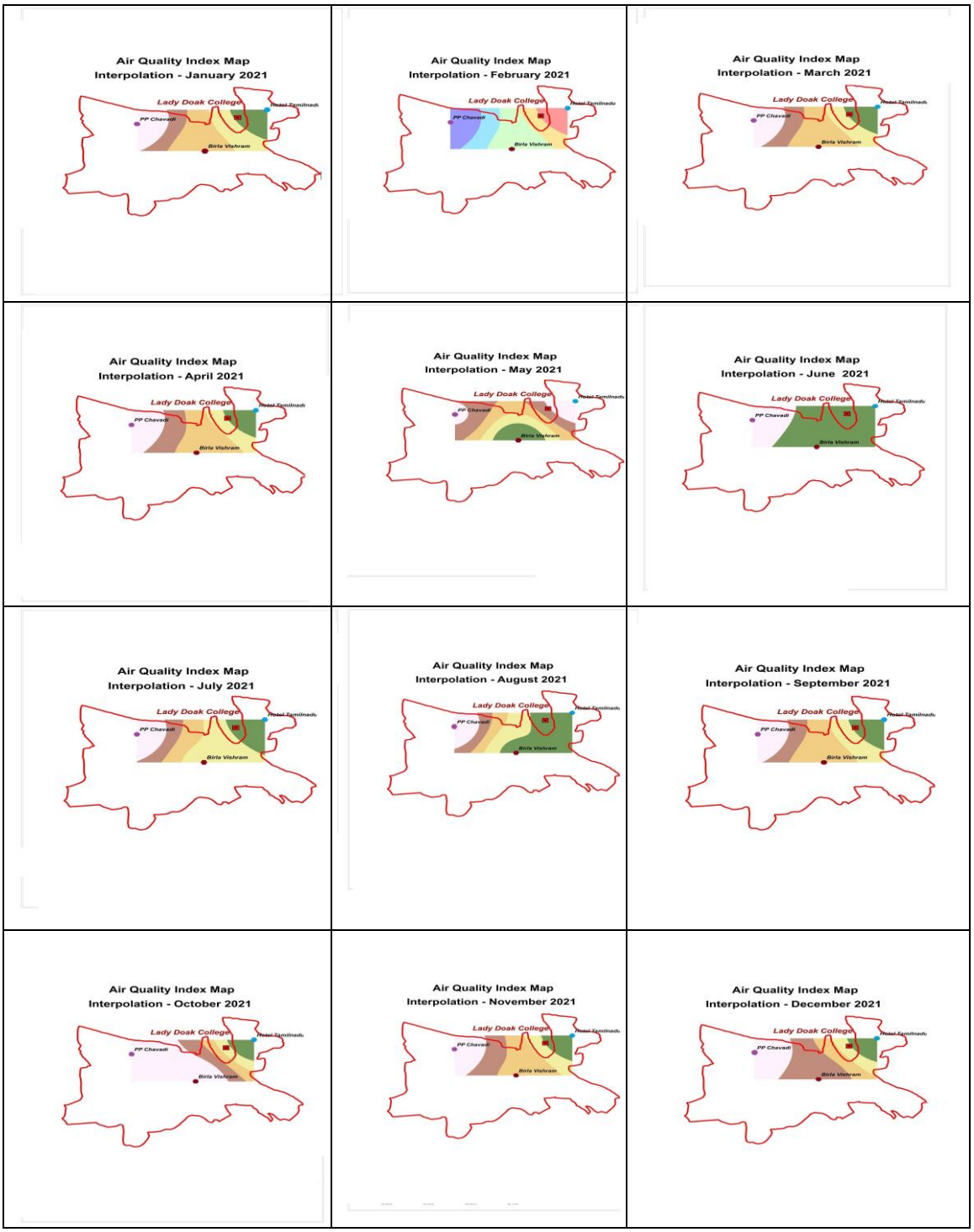

Figure 4. Month wise AQI for 3 stations from January 2021-December 2021

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	Station 1	**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

The results indicate that reducing human activities such as transportation and commercial activities leads to an AQI value of less than 50 across all zones (Residential, Commercial, and Industrial). The Bayesian Belief Network developed in this study assesses the impact of human activity on AQI. This research is pioneering in Madurai district, as it explores causal relationships between human activities and air quality for the first time. Consequently, this study provides a foundation for developing policies aligned with Sustainable Development Goals (SDGs) to address air pollution in the city. Future work could involve designing real-time embedded applications to recommend sustainable practices like carpooling and the use of electric vehicles, offering valuable insights for policymakers and governance.

ACKNOWLEDGEMENTS

We extend our gratitude to the Tamil Nadu Pollution Control Board (TNPCB) for providing the station-wise dataset essential for this research. Without their contribution, this work would not have been possible. We also appreciate the support of Dr. Lakshmi, Assistant Professor in the Department of Physics at Lady Doak College, TN, India for her assistance with the IDW Interpolation process.

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.

[16] D Hema, Priyadarshini., “Assessing Human impact on Air Quality with Bayesian Networks and IDW Interpolation”. 8th International Conference on Computer Science and Information Technology (COMIT 2024), Chennai. Vol. 14, no.15, pp. 133-142.

AUTHORS

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.