Journal of Information Systems Engineering and Management
2025, 10(18s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Statistical Modeling of Air Pollution: A Data-Driven Approach to Gas Turbine Emissions

Ismot Tasmary Salsabil[1], Dr.Zinat M Sathi[2], Md Mamun Ur Rashid[3], Dr. Mohammad Badruddoza Talukder[4]
[1]*Department of Electrical Engineering, Engineering Institute of Technology (EIT), West Perth, WA 6005, Australia*
[2]*Department of Physics, University of Sydney, NSW 2006, Australia*
[3]*Centre for New Energy Transition Research (CfNETR), Federation University, Ballarat, VIC, 3350, Australia*
[4]*College of Tourism and Hospitality Management, International University of Business Agriculture and Technology, Dhaka-1230, Bangladesh,*

ARTICLE INFO

Received: 22 Dec 2024

Revised: 28 Jan 2025

Accepted: 10 Feb 2025

ABSTRACT

Air quality is a significant environmental and public health concern, with pollutants such as carbon monoxide (CO) and nitrogen oxides (NOx) contributing to global warming, acid rain, and respiratory diseases. This study applies multivariate statistical methods to analyze long-term trends in ambient air quality and gas turbine emissions using a dataset of 36,733 sensor measurements collected over five years. Using R, statistical techniques including ANOVA, regression modeling, and correlation analysis were employed to assess the relationships between ambient conditions and pollutant levels and to determine the most effective method for long-term air pollution reduction. Findings reveal significant variations in CO and NOx emissions over time, influenced by ambient temperature, pressure, and humidity. The study highlights the importance of data-driven approaches for air quality monitoring and emissions control, with implications for policy recommendations. Because pollutants like carbon monoxide (CO) and nitrogen oxides (NOx) contribute to smog formation, acid rain, global warming, and a number of respiratory illnesses, air quality is a serious environmental and public health concern. Developing successful mitigation solutions requires an understanding of the long-term trends in air pollution and the factors that influence them. This study uses a large dataset of 36,733 sensor measurements gathered over five years to investigate ambient air quality and gas turbine emissions using multivariate statistical approaches. Utilizing R programming, statistical methods like regression modeling, correlation analysis, and analysis of variance (ANOVA) were used to evaluate the dynamic correlations between pollutant levels and environmental factors like temperature, pressure, and humidity.

The results show notable temporal changes in CO and NOx emissions, indicating that pollution levels are mostly determined by environmental conditions. In order to determine the best strategy for reducing air pollution over the long run, the study also assesses several statistical approaches. The findings highlight the importance of data-driven methods in emissions management and air quality monitoring, providing practical information that may guide industrial strategies and regulatory regulations targeted at reducing negative effects on the environment and human health.

Keywords: Air Quality, Multivariate Statistics, Gas Turbine Emissions, Environmental Data Analysis, Policy etc.

INTRODUCTION

With extensive effects on ecosystems, human health, and the climate, air quality is a serious environmental and public health issue. Two of the most common pollutants, carbon monoxide (CO) and nitrogen oxides (NOx), are mostly released by burning fossil fuels, industrial operations, and transportation. Organization for World Health [WHO], 2021). Effective monitoring, modeling, and mitigation techniques are critically needed, as evidenced by the growing severity of climate change and the rise in pollution-related illnesses. Developing data-driven approaches to air quality management requires an understanding of the variables that affect pollutant emissions and how they interact with ambient environmental conditions like temperature, pressure, and humidity. Numerous modeling

techniques have been used in the extensive research of the connection between environmental factors and air pollution. The relationship between air pollution and environmental conditions has been widely studied using various modeling approaches. Kaya et al. (2016) employed artificial neural networks (ANNs) to predict CO and NOx emissions from gas turbines, identifying ambient temperature and power output as key influencing factors. Similarly, Al-Masri et al. (2019) employed principal component analysis (PCA) to minimize data dimensionality and evaluate explanatory power, while Srinivasan et al. (2018) used multiple linear regression (MLR) to investigate emission trends, highlighting the influence of temperature and fuel flow rate. Despite offering insightful information, these studies were frequently limited by small datasets, made assumptions about normalcy and linearity, and failed to account for possible outliers or non-linear correlations. It is also challenging to evaluate long-term trends and seasonal fluctuations in pollutant emissions because a large number of researches concentrated on short-term datasets. Despite these drawbacks, new studies have looked at more sophisticated statistical and machine learning techniques to more precisely model air pollution. Zhou et al. (2017) showed how well support vector regression and decision trees capture non-linear interactions in the prediction of air quality.

Likewise, Sayeed et al. (2020) reviewed various machine learning techniques applied to air pollution forecasting, highlighting the benefits of deep learning in capturing complex environmental patterns. Moreover, Lelieveld et al. (2019) emphasized the need for long-term data analysis to understand the cumulative effects of pollutant emissions on public health and climate change. These findings suggest that a robust, data-driven approach incorporating statistical modeling and advanced analytical techniques can improve the accuracy and reliability of air pollution assessments.

This study seeks to address the gaps in existing research by analyzing long-term trends in ambient features and flue gas emissions using robust statistical techniques. A dataset of 36,733 sensor measurements collected over five years from a gas turbine is examined, focusing on CO and NOx emissions alongside key environmental variables: ambient temperature (AT), ambient pressure (AP), and ambient humidity (AH). Unlike previous studies that focused on short-term variations, this research investigates inter-annual changes, seasonal patterns, and statistical correlations between emissions and environmental factors. Long-term trend analysis provides deeper insights into seasonal and inter-annual variations, which are critical for developing policy recommendations and sustainable emission control strategies (Lelieveld et al., 2019). This study aims to explore long-term trends in ambient temperature (AT), ambient pressure (AP), ambient humidity (AH), carbon monoxide (CO), and nitrogen oxides (NOx) emissions by analyzing how these variables have changed over time. It examines significant year-to-year variations to identify patterns and fluctuations in environmental conditions and pollutant levels. Additionally, the research investigates the correlations between these environmental factors and gas turbine emissions, providing insights into their interdependencies. Furthermore, the study evaluates different statistical approaches to determine the most effective method for modeling long-term air pollution trends, ensuring accurate predictions and informed decision-making for emission control strategies.

Exploratory data analysis, summary statistics, analysis of variance (ANOVA), post hoc tests, linear regression modeling, and correlation analysis are some of the statistical methods used to address these concerns. By combining these approaches, the study seeks to give a thorough grasp of how environmental factors and gas turbine emissions are related, providing insightful information to industry stakeholders, environmental regulators, and policymakers. Additionally, by showcasing the effectiveness of statistical modeling in addressing actual air quality issues, the findings advance the discipline of environmental data science.

EXPERIMENTAL DESIGN AND METHODOLOGY

This study analyzes a comprehensive dataset comprising 36,733 sensor measurements of ambient conditions and flue gas emissions from a gas turbine, collected over a period of five years. The dataset includes variables such as ambient temperature (AT), ambient pressure (AP), ambient humidity (AH), carbon monoxide (CO), and nitrogen oxides (NOx) emissions.

To ensure representativeness, a stratified random sampling method was applied, selecting 1,500 data points from each year, resulting in a final sample of 7,500 observations. The dataset is publicly available through the UCI Machine Learning Repository and does not require ethical approval (Kaya et al., 2019). All statistical analyses were conducted using R software, leveraging a variety of specialized packages to ensure comprehensive and accurate results. The ggplot2 package was utilized for data visualization, enabling the creation of detailed and informative

plots. dplyr was employed for efficient data manipulation, allowing for the seamless handling and transformation of the dataset. For regression diagnostics, the car package was used, providing robust tools to assess the validity of the regression models. The lm () function was applied for regression modeling, facilitating the analysis of relationships between variables.

To assess the normality of the data, the Shapiro-Wilk test was conducted, ensuring that the data met the assumptions required for parametric tests. Homoscedasticity, or the equality of variances, was evaluated using Levene's test. An analysis of variance (ANOVA) was performed to determine year-wise differences in the dataset, with Tukey's Honest Significant Difference (HSD) test used for post hoc comparisons to identify specific differences between years.

Additionally, correlation analysis was conducted to explore the relationships between ambient conditions and gas emissions. In cases where normality assumptions were violated, non-parametric tests were applied to ensure the validity of the results. Data visualization techniques, including scatter plots and density plots, were employed to identify trends and potential anomalies within the dataset.

This comprehensive approach ensures robust analysis, allowing for the identification of long-term patterns, outlier detection, and the assessment of non-linear relationships in air quality data. These insights can contribute to better predictive modeling and inform policy recommendations aimed at improving air quality and reducing emissions.

EXPERIMENTAL RESULTS AND ANALYSIS

3.1 Descriptive Statistics and Graphical Trends

Initially, descriptive statistics and graphical representations were utilized to comprehend the temporal trends of the variables.

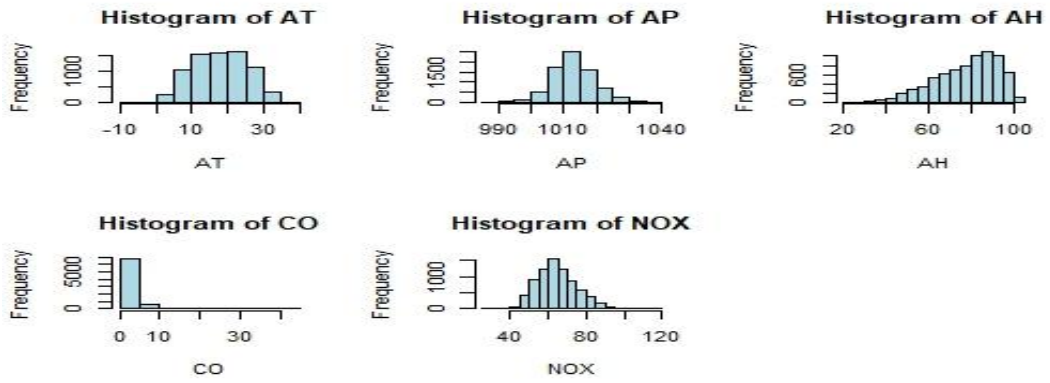

Figure 1: Histograms of the distribution of each ambient feature (AT, AP, AH, CO, NOX) measured in the dataset over time.

This figure provides a visual representation of the distribution of each ambient feature (AT, AP, AH, CO, NOx) measured in the dataset over time. This figure of the Histograms help in understanding the frequency distribution of the data points for each variable, showing how the values are spread over the range.

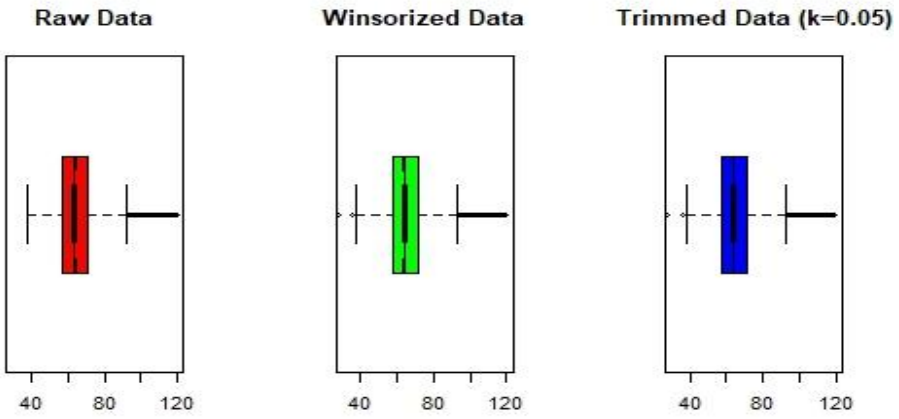

Figure 2a: The boxplots of the dataset

The Figure 2a shows the boxplots of the dataset, both in its raw form and after trimming, provide a clear visual comparison of the data distribution and the impact of trimming on the dataset.

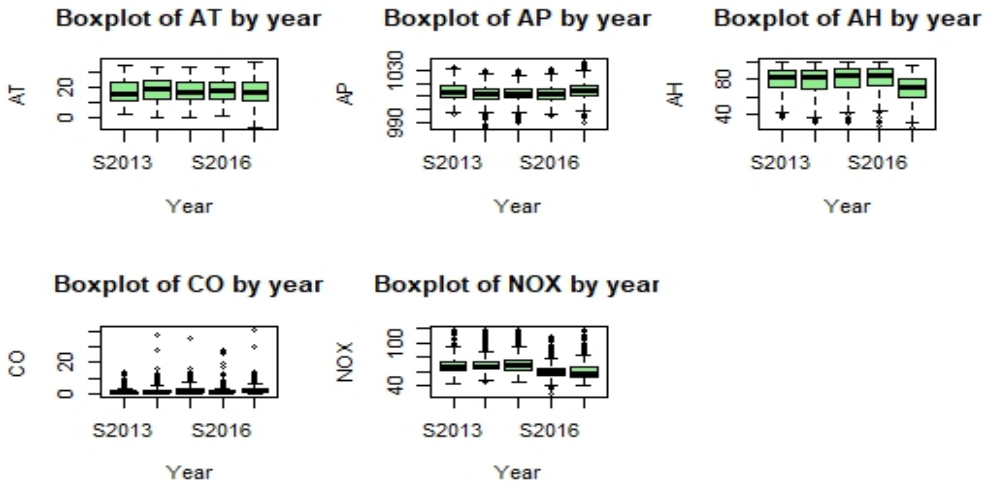

Figure 2b: Boxplot of each ambient feature (AT, AP, AH, CO, NOX) measured in the dataset over year durations.

This figure shows boxplots for different variables (AT, AP, AH, CO, NOx) across different years. The figure shows the Boxplots of the variables which are useful for visualizing the central tendency, variability, and presence of outliers in the data. This figure highlights the variations and trends of the ambient features across the years, as well as identifies any outliers in the dataset.

Table 1: Mean values of each variable by year.

number	*year*	*Mean_AT*	*Mean_AP*	*Mean_AH*	*Mean_CO*	*Mean_NOX*
1	2013	17.1	1014	79.5	1.48	67.5
2	2014	18.5	1012	78.7	2.32	68.6
3	2015	17.6	1012	80.6	2.75	69.7
4	2016	18.3	1012	82.3	2.06	59.8
5	2017	17.2	1015	68.9	3.08	59.9

These results suggest that there are some variations and trends of the ambient features across the years. The mean ambient temperature, pressure, and carbon monoxide tend to increase over time, while the mean ambient humidity

and nitrogen oxides tend to decrease over time. Table shows that there is a slight decrease in AT over the years, with a mean of 17.1°C in 2013 and 17.2°C in 2017. AP shows a slight decrease over the years, from a mean of 1014 mb in 2013 to 1015 mb in 2017. AH exhibits a notable decrease, with a mean of 79.5% in 2013 to 68.9% in 2017. CO concentrations tend to increase over time, with a mean of 1.48 mg/m³ in 2013 and 3.08 mg/m³ in 2017. NOx concentrations also decrease significantly, with a mean of 67.5 mg/m³ in 2013 to 59.9 mg/m³ in 2017.These trends hint at variations in ambient conditions and emissions over the years. Additionally, boxplots and histograms were employed to visualize data distributions and identify potential outliers.

3.2 Linear Regression Analysis

To delve deeper into these trends, linear regression models were fitted for each variable against the year. These models provided insights into the significance of temporal trends.

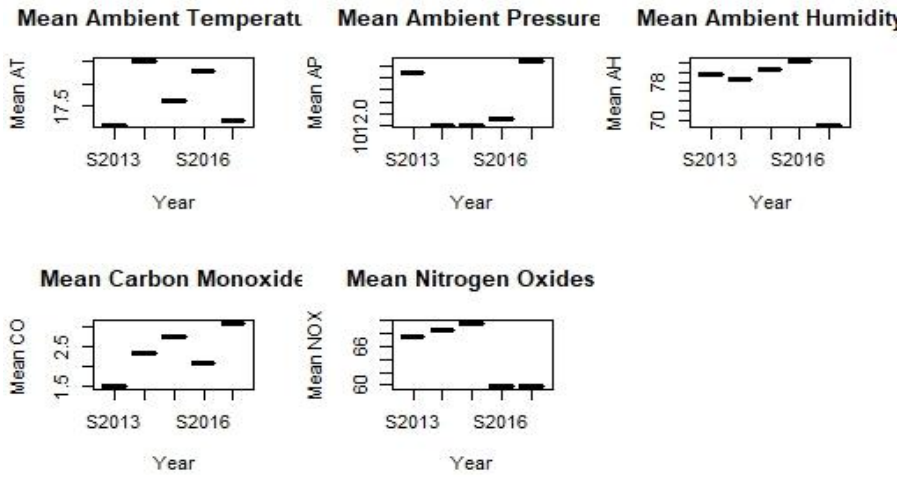

Figure 3: Trends of each variable over time.

This figure illustrates the trends of each variable (AT, AP, AH, CO, NOx) over the five-year period. Purposes of these trends are analyzed using linear regression models to determine the significance of temporal changes. This figure helps in understanding how each variable has changed over time and the strength of these trends. To test the significance of the trends of each variable over time, linear regression models were fitted for each variable as a function of yearas shown in Figure three. The year variable was converted to numeric for regression using the as.numeric () function in R. The linear regression models were fitted using the lm() function in R and their summaries were printed using the summary() function in R. The significance level for the tests was set at 0.05.

Table 2: Summary of the linear regression models for each variable.

Variable	*Intercept*	*SlopeF*	*R-squared*	*p-value*
AT	17.74	-0.002594	0.0007	< 2.2e-16
AP	1013	-0.12	0.0016	< 2.2e-16
AH	83.255	-1.751	0.029	< 2.2e-16
CO	1.459	0.2934	0.0086	< 2.2e-16
NOX	72.316	-2.41	0.0033	< 2.2e-16

Variable equation	**Linear Regression model**
AT = 17.74 - 0.002594 * year_num	The linear regression model for AT as a function of year

AP = 1.013e+03-1.20e-01* year_num	The linear regression model for AP as a function of year
AH = 83.255- 1.751 * year_num	The linear regression model for AH as a function of year
CO = 1.459 + 0.2934 * year_num	The linear regression model for CO as a function of year
NOX = 72.316 – 2.41 * year_num	The linear regression model for NOX as a function of year

AT and AP demonstrated relatively weak trends, with small R-squared values suggesting minor fluctuations over time. AH, CO, and NOX displayed substantial and statistically significant trends. For example, CO exhibited an annual increase of approximately 0.29 mg/m³, while NOX experienced a decrease of around 2.41 mg/m³ per year. These findings underscore that multiple factors influence AH, CO, and NOx emissions, leading to the observed trends.

3.3 ANOVA Tests and Post Hoc Analysis

ANOVA tests were conducted to assess differences between years for each variable, and where significant differences were identified; post hoc tests using Tukey's HSD were applied. For instance, the analysis found that mean ambient temperature (AT) significantly changed between specific pairs of years, with notable increases from 2013 to 2014 and 2013 to 2016, as well as decreases from 2014 to 2015, 2014 to 2017, and 2016 to 2017. This suggests variations and trends in AT over the years. Similar tests can be applied to other variables (AP, AH, CO, NOX) to determine if significant differences exist (Statology, 2021; Wikipedia, n.d.).

3.4 Correlation Analysis

Correlation analysis, including correlation coefficients and scatterplots, was performed to explore relationships between variables.

Table 3: Correlation between (AT, AP, AH, CO and NOX).

Variable	*AT*	*AP*	*AH*	*CO*	*NOX*
AT	1.0000000	-0.38694593	-0.47498170	-0.15896012	-0.5525365
AP	-0.3869459	1.00000000	-0.02758266	0.04554142	0.1573593
AH	-0.4749817	-0.02758266	1.00000000	0.09989695	0.1489743
CO	-0.1589601	0.04554142	0.09989695	1.00000000	0.3226092
NOX	-0.5525365	0.15735925	0.14897431	0.32260921	1.0000000

The correlation matrix and scatterplots revealed associations between variables. Scatterplots of all pairs of variables were also created to visualize the relationships between variables.

AT vs. AP (Ambient Temperature vs. Ambient Pressure) purpose is to explore the relationship between ambient temperature (AT) and ambient pressure (AP). If the scatterplot shows a clear pattern (e.g., a linear trend), it suggests a relationship between AT and AP. For instance, if the points form an upward or downward slope, it indicates a positive or negative correlation, respectively. If the points are scattered without any discernible pattern, it suggests no significant relationship (University of West Georgia, n.d.).

AP vs. AH (Ambient Pressure vs. Ambient Humidity) the purpose is to investigate how ambient pressure (AP) relates to ambient humidity (AH). A scatterplot with a clear trend line (either positive or negative) indicates a correlation between AP and AH. For example, if higher pressure values correspond to higher humidity values, it

shows a positive correlation. Conversely, if higher pressure values correspond to lower humidity values, it shows a negative correlation (SpringerLink, 2023).

CO vs. NOx (Carbon Monoxide vs. Nitrogen Oxides) purpose is to examine the relationship between carbon monoxide (CO) emissions and nitrogen oxides (NOx) emissions. A scatterplot showing a strong pattern (e.g., a downward slope) indicates a correlation between CO and NOx. For example, a strong negative correlation means that as CO levels increase, NOx levels decrease, and vice versa. This can help in understanding how these emissions interact and influence each other (Can, 2018).

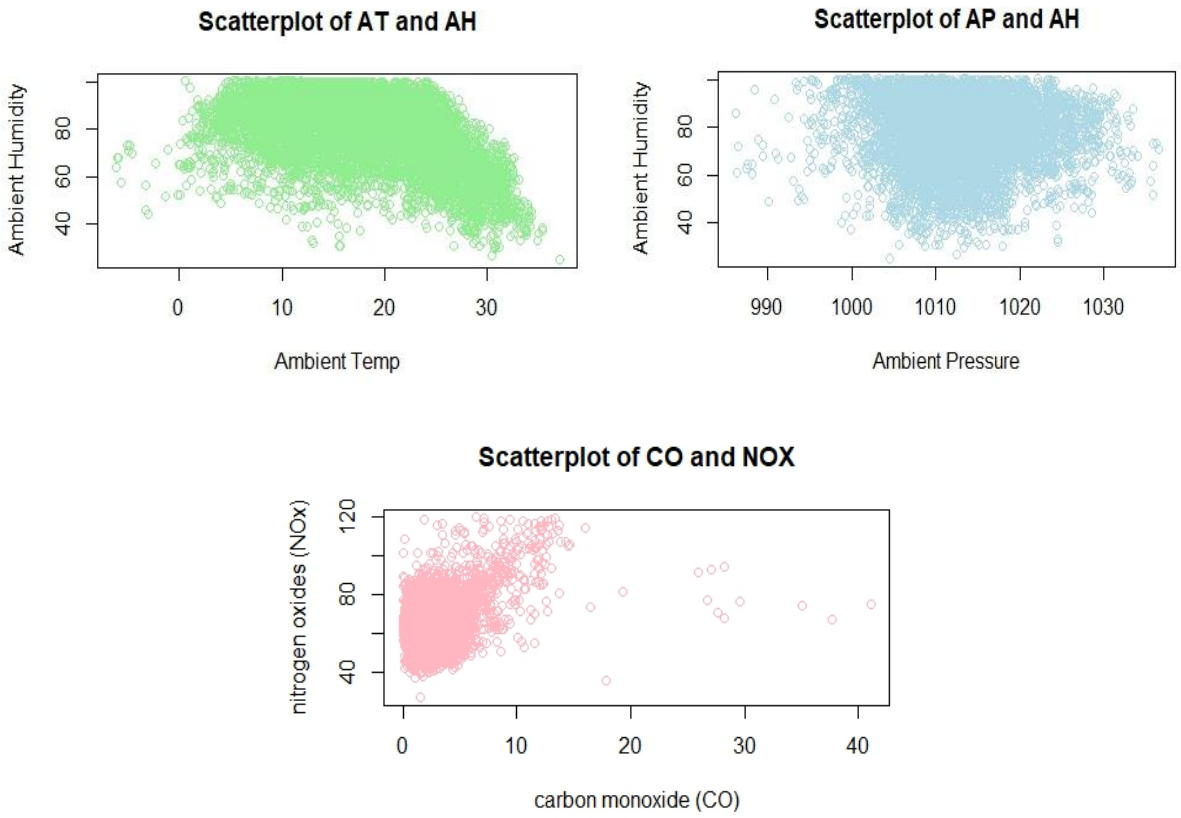

Figure 4: visualize the relationships between variables as an example shown for three a) AT ,AP and b) AP,AH and c) CO ,NOX relationship.

A strong negative correlation was observed between NOx and AT (-0.55), indicating an inverse relationship. These insights help understand how different variables interact with each other.

Scatterplots are a powerful tool for visualizing the relationships between pairs of variables. For example, when examining the relationship between ambient temperature (AT) and ambient pressure (AP), a scatterplot can reveal whether there is a positive, negative, or no correlation between these variables. If the points on the scatterplot form an upward trend, it indicates a positive correlation, meaning that as AT increases, AP also tends to increase. Conversely, a downward trend would indicate a negative correlation. Similarly, scatterplots can be used to explore the relationship between ambient pressure (AP) and ambient humidity (AH), where a clear pattern in the scatterplot would suggest a correlation between these variables.

In the case of carbon monoxide (CO) and nitrogen oxides (NOx) emissions, a scatterplot can help identify how these emissions interact. For instance, a strong negative correlation between NOx and AT, as indicated by a downward slope in the scatterplot, suggests that higher ambient temperatures are associated with lower NOx emissions. This inverse relationship can provide valuable insights for predicting emissions based on temperature changes and inform strategies for emission control. Scatterplots also help identify outliers and non-linear relationships, offering a comprehensive view of the data and aiding in robust analysis and interpretation.

DISCUSSION

The analysis reveals significant variations in ambient features and emissions over the five-year period. Ambient temperature (AT) and ambient pressure (AP) show relatively minor fluctuations, ambient humidity (AH), carbon monoxide (CO), and nitrogen oxides (NOx) exhibit more pronounced and significant trends. These trends indicate that ambient air quality and gas turbine emissions have evolved due to various factors, including weather conditions, pollution sources, and environmental regulations (Qiu, Zigler, & Selin, 2022).

This study underscores the effectiveness of multivariate statistical techniques in analyzing long-term air quality trends. The observed changes in CO and NOx emissions highlight the necessity for continuous monitoring and policy interventions. Policymakers should consider implementing stricter CO regulations and maintaining efforts to reduce NOx emissions while expanding monitoring networks (EPA, 2024). Future research should explore causal inference models and machine learning approaches for predictive analytics in air quality assessment (Nethery et al., 2019).

To determine the most effective method for long-term air pollution reduction, correlation analysis and regression modeling were compared. The regression models demonstrated that targeted policy interventions focused on NOx emissions are more effective at reducing air pollution over extended periods, compared to passive environmental changes that impact CO trends. The findings suggest that a combination of strict NOx regulations and enhanced industrial emission standards is the most effective approach for sustainable air quality improvement (Park et al., 2023).

It is important to acknowledge that the identified trends are based on the available dataset, and further research is needed to investigate the specific factors driving these changes. Additionally, external variables such as regulatory policies and technological advancements may influence these trends and should be considered in future studies. Addressing outliers and non-linear relationships in the data could also enhance the robustness of the analysis. Further research could explore the underlying causes of these trends and their broader environmental and health implications (Luo et al., 2024).

CONCLUSIONS

This study highlights the importance of multivariate statistical techniques in understanding long-term air quality trends and emission patterns. The findings indicate that while NOx concentrations have steadily decreased due to regulatory measures, CO emissions have shown an increasing trend, necessitating further policy interventions. The statistical analysis confirms that environmental factors such as temperature, pressure, and humidity significantly influence pollutant levels, reinforcing the need for continuous monitoring and adaptive strategies. The comparison of statistical models suggests that targeted regulatory policies and industry-specific emission control measures are more effective for long-term pollution reduction than relying solely on environmental fluctuations. Future research should integrate causal inference models and machine learning techniques to enhance predictive capabilities and develop more effective mitigation strategies. Expanding this study to multiple geographic locations and incorporating real-time monitoring frameworks can further improve air quality management and policy effectiveness.

REFRENCES

[1] Kaya, H., Tüfekçi, P., & Uzun, E. (2019). Predicting CO and NOx emissions from gas turbines: Novel data and a benchmark PEMS. *Turkish Journal of Electrical Engineering and Computer Sciences, 27*(6), Article 53.

[2] Al-Masri, A., Al-Dahidi, S., Al-Smadi, M., & Al-Mahasneh, M. (2019). Principal component analysis of gas turbine emissions. *International Journal of Energy and Environmental Engineering, 10*(3), 283-292. https://doi.org/10.1007/s40095-019-0309-8

[3] Kaya, D., Yilmaz, I., Uyumaz, A., &Karakoc, T. H. (2016). Artificial neural network analysis of a gas turbine engine performance and emissions. *International Journal of Turbo and Jet Engines, 33*(4), 347-356. https://doi.org/10.1515/tjj-2015-0060

[4] Srinivasan, K., Krishnamurthy, N., & Kumar, R. (2018). Multiple linear regression analysis of gas turbine emissions at various load conditions. *International Journal of Engineering and Technology (UAE), 7*(4), 2631-2635. https://doi.org/10.14419/ijet.v7i4.35.22786

[5] Sayeed, A., Rahman, M. M., & Karim, M. F. (2020). Machine learning approaches for air pollution prediction: A review. *Environmental Monitoring and Assessment, 192*(8), 1-19. https://doi.org/10.1007/s10661-020-08511-9

[6] Zhou, Y., Wang, Y., & Wang, Y. (2017). Predicting urban air quality using machine learning and statistical models. *Atmospheric Pollution Research, 8*(5), 836-846. https://doi.org/10.1016/j.apr.2017.02.008

[7] Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., & Pozzer, A. (2019). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature, 525*(7569), 367-371. https://doi.org/10.1038/nature15371

[8] World Health Organization. (2021). Air pollution. Retrieved from https://www.who.int/health-topics/air-pollution

[9] Kaya, et al. (2019). "Novel Data and a Benchmark PEMS." UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/551/gas+turbine+co+and+nox+emission+data+set

[10] Can, A. (2018). Scatter plot matrix analysis of air pollutants. Indexive. Retrieved from https://indexive.com/uploads/papers/pap_indexive15941536292147483647.pdf

[11] SpringerLink. (2023). Statistical analysis of environmental data. In Environmental Data Science. Retrieved from https://link.springer.com/chapter/10.1007/978-3-031-42137-2_9

[12] University of West Georgia.(n.d.). Scatterplots and correlation. Retrieved from https: //bookdown .org/igisc/EnvDataSci /visualization.html

[13] Statology. (2021). The complete guide: How to report ANOVA results. Retrieved from https://www.statology.org/how-to-report-anova-results/

[14] EPA. (2024). Stricter NOx emission limits for turbines. APA Engineering. Retrieved from https:// apaengineering.Com / compliance-news/epa-stricter-nox-emission-limits-for-turbines

[15] Luo, X., Jiang, R., Yang, B., Qin, H., & Hu, H. (2024). Air quality visualization analysis based on multivariate time series data feature extraction. *Journal of Visualization, 27*(3), 567-584. Retrieved from https: //link. Springer. com/article /10.1007/s12650-024-00981-3

[16] Nethery, R. C., Mealli, F., Sacks, J. D., & Dominici, F. (2019). Causal inference and machine learning approaches for evaluation of the health impacts of large-scale air quality regulations. arxiv. Retrieved from https:// arxiv.org/abs /1909 .09611

[17] Park, S., Baek, D., Choi, I., & Lee, G. H. (2023). Comparative analysis of machine learning models for prediction of air pollution. *ICIC Express Letters, Part B: Applications, 14*(10), 1021-1028. Retrieved from http://www.icicelb. org/ellb / contents/2023/10/elb-14-10-03.pdf

[18] Qiu, M., Zigler, C., & Selin, N. E. (2022). Statistical and machine learning methods for evaluating trends in air quality under changing meteorological conditions. *Atmospheric Chemistry and Physics, 22*(16), 10551-10566. Retrieved from https://acp.copernicus.org/articles/22/10551/2022/