Investigating Air Pollution Dynamics in Ho Chi Minh City: A Spatiotemporal Study Leveraging XAI-SHAP Clustering Methodology

Polat Goktas[1,2](✉), Rajnish Rakholia[1,2], and Ricardo S. Carbajo[1,2]

[1] UCD School of Computer Science, University College Dublin, Belfield, Dublin, Ireland
{polat.goktas,rajnish.rakholia,ricardo.simoncarbajo}@ucd.ie
[2] CeADAR: Ireland's Centre for Applied Artificial Intelligence, Clonskeagh, Dublin, Ireland

Abstract. Air pollution poses an urgent challenge to public health and ecosystems, particularly in rapidly urbanizing regions. Despite the severity of this issue, there is a lack of robust analytical frameworks capable of identifying key variables and their spatial effects across landscapes. Our study directly addresses this void by applying an innovative supervised clustering approach to air quality data in Ho Chi Minh City (HCMC), Vietnam—a rapidly growing urban area grappling with escalating pollution levels. The analytical model employs Shapley Additive exPlanations (SHAP) to interpret feature importance within tree-based machine learning models, supplemented by the Unified Manifold Approximation and Projection (UMAP) technique to explore intersections between affected areas. We utilize a feature set from Rakholia et al. (2023) as input variables for each target time series, with a focus on answering key questions: What pollutants exert the most influence at different times of day? Which areas of the city are most affected? And can this method effectively pinpoint intersections of pollutant effects? Our results reveal morning traffic congestion predominantly elevates levels of Nitrogen Dioxide (NO_2), Humidity, and Carbon Monoxide (CO), while afternoon emissions are significantly impacted by Sulfur Dioxide (SO_2), CO, and Ozone (O_3) due to solar radiation and industrial activities. Through this research, we expect to contribute to the ongoing discourse on urban air pollution management, highlighting the potential of artificial intelligence-driven tools in environmental research and policy-making.

Keywords: Shapley Additive exPlanations (SHAP) · XAI · Air quality

1 Introduction

Air quality is critical for human health and the preservation of the environment's ecology. However, air pollution has become a major concern in many nations, bringing serious health hazards such as a rise in the incidence of heart disease, asthma, and lung cancer [1]. Air pollution also contributes to environmental issues such as global warming, acid rain, and depletion of the ozone layer. Rapid urbanization, uncontrolled transportation, and poorly regulated industrial environments all contribute to the issue. According to

S. Nowaczyk et al. (Eds.): ECAI 2023 Workshops, CCIS 1948, pp. 201–207, 2024.
https://doi.org/10.1007/978-3-031-50485-3_20

the World Health Organization (WHO), 9 out of 10 individuals live in areas where air pollution exceeds WHO limits [2]. According to studies, outdoor air pollution, mostly caused by Particulate Matter ($PM_{2.5}$), causes millions of premature deaths per year globally [3].

With approximately 60,000 fatalities per year attributable to air pollution-related disorders, Ho Chi Minh City (HCMC) in Vietnam has a significant air pollution crisis. As the city's economy and people have grown, so too has the harm to their health posed by the growing $PM_{2.5}$ levels in the metropolis. By 2030, it is expected that $PM_{2.5}$ levels in HCMC and other Vietnamese cities would have increased by 30% [4].

The levels of air pollutants, such as Nitrogen Dioxide (NO_2), Ozone (O_3), Sulfur Dioxide (SO_2), and Carbon Monoxide (CO), in the city exceed the WHO threshold limits, posing serious risks to human health and the environment [5]. The non-linearity and time-varying nature of the data, as well as the complex interplay between air contaminants and meteorological conditions, provide the greatest challenge to accurate air pollution forecasting [6]. Artificial intelligence (AI)-based $PM_{2.5}$ models have demonstrated better performance in dealing with non-linearity and time-varying data [7, 8].

This paper presents a supervised clustering approach based on Shapley Additive exPlanations (SHAP) values to investigate the impact of different air pollution factors, using a publicly accessible outdoor air quality dataset in HCMC. To better understand the causes of air pollution in specific areas, we employ this methodology to go deeper than simple statistical analyses of air contaminants. The main objectives of this study are to (1) identify air pollutant factors and their combinations that are likely to increase at specific time points, (2) identify affected stations or regions of the city, (3) determine the impact of increased pollutant levels on the other areas, and (4) determine the feasibility of using supervised clustering based on SHAP values to identify intersections between affected stations in the city.

2 Related Work

2.1 Application of AI in Forecasting Air Pollutants

There has been a growing trend in employing AI-based models for predicting air quality. Their ability to model non-linear associations and handle large-scale datasets makes them superior for air pollution prediction. Several AI-based models that have been used in this context include Random Forest [9], XGBoost [10], Neural networks [11], and Hybrid and Multi-output models [7, 8].

2.2 Interpreting Models with XAI – SHAP Approach in Environmental Research

The concept of eXplainable AI (XAI) has gained significant traction in recent AI research. The primary objective of XAI is to enhance the transparency and interpretability of complex AI models. One such method is the SHAP approach, which provides a cooperative game theory framework to explain the output of any machine learning model [12]. SHAP values has been effectively applied to diverse environmental research for its capacity to understand complex variable interactions, enhancing model interpretability [13–15]. Specific studies include its integration with machine learning for seasonal

$PM_{2.5}$ projections in Beijing [13], air pollution predictions [14], and highlighting critical factors in estimating NO_2 concentrations [15]. Across these applications, the inclusion of SHAP has consistently improved the transparency, interpretability, and predictive power of the respective environmental models.

3 Investigating Air Pollution Dynamics Using XAI-SHAP Clustering

3.1 Dataset and Experimental Settings

For this study, we adopt the HealthyAir dataset [16], a database of environmental air quality measurements that is freely available to the public, for our evaluation. This public database comprises 52,549 records of air quality measurements compiled by the Air Quality Monitoring Network in HCMC. These records, which span February 2021 to June 2022, were collected from six air monitoring stations distributed across diverse urban locales including residential neighborhoods, commercial zones, and densely populated areas.

The dataset captures two weather conditions—Temperature (°C) and Humidity (%)—and hourly pollutant concentrations—$PM_{2.5}$, Total Suspended Particles (TSP), SO_2, O_3, NO_2, and CO—measured in $\mu g/m^3$. Our research considers data from the following monitoring stations in HCMC:

- *Urban background*: Vietnam National University, HCMC (10.86994333, 106.7960143).
- *Residential*: 49 Thanh Da Street, Binh Thanh District, HCMC (10.81584553, 106.7174282).
- *Traffic*: 268 Nguyen Dinh Chieu Street, District 3, HCMC (10.77636612, 106.6878094).
- *Traffic + Residential*: MM18 Truong Son Street, District 10, HCMC (10.78047163, 106.6594579).

These records underwent various preprocessing stages, including unit conversions, data transformations, and missing value treatments as per established literature [7, 8]. Input variables for each potential air pollutant combination as per Rakholia et al. [8] were sorted for selected time points based on typical patterns of human activity and traffic congestion in HCMC.

3.2 Constructing and Assessing ML Classification Models

We employed a range of cutting-edge tree-based ML methods, such as decision tree, random forest, gradient-boosting decision tree (GBDT), histogram-based gradient-boosting classification tree (HistGBDT), and light gradient-boosting machine (LightGBM, version 3.3.3). The dataset was divided into training and testing sets, and classifiers were trained and tested under different configurations. The performance of ML classifiers was evaluated using metrics like accuracy, precision, recall, and *F1*-score. All experiments were run on a GPU server with specific specifications.

3.3 SHAP-Based Dimensionality Reduction

In order to highlight the value of dimensionality reduction using SHAP values, we performed supervised clustering using average SHAP values as per the feature set proposed by Rakholia et al. [8] for each target time series in the dataset. Tree-based ML models were used to categorize target areas, excluding $PM_{2.5}$ and TSP from the analysis. Additionally, the UMAP (Uniform Manifold Approximation and Projection) technique was employed to depict intersections between affected regions, providing a two-dimensional visual of high-dimensional data.

4 Study Findings and Insights

4.1 Hourly Variations in Air Pollutant Concentrations Across Monitoring Stations

We leveraged a suite of tree-based machine learning models to classify monitoring stations in HCMC into *Urban background*, *Residential*, *Traffic*, and *Traffic plus Residential*, based on NO_2 air pollution level at various time intervals (Table 1). We observe that certain models routinely achieve higher performance across different time intervals. LightGBM, HistGBDT and Random Forest classifiers, for example, perform well across all metrics, especially recall, which is essential for reducing false negatives, while the Decision Tree model results in relatively lower scores. We can see how well all tree-based ML models perform at 9 AM and 5 PM. The Random Forest model performs the best at both 9 AM and 5 PM time points, with the highest accuracy, precision, recall, and *F1*-scores. Overall, the performance of all models is higher at 5 PM than at 9 AM, except for the HistGBDT model. Notably, the models' performances underscored the influence of time and pollutant type on air pollutant concentration classification (Table 1).

Table 1. Performance comparison of tree-based machine learning models in classifying *Urban background*, *Residential*, *Traffic*, and *Traffic plus Residential* monitoring stations for NO_2 air pollution with CO, O_3, SO_2, Humidity, and Temperature at specific time points (7 AM, 9 AM, 1 PM, and 5 PM).

		Percentage Split			
Time	*Model*	*Accuracy*	*Precision*	*Recall*	*F1-score*
07:00	GBDT	89.787	89.896	89.787	88.830
	LightGBM	94.042	94.609	94.043	93.827
	HistGBDT	93.191	93.809	93.191	92.886
	Decision Tree	86.808	86.747	86.809	86.564
	Random Forest	90.639	91.032	90.638	90.024
09:00	GBDT	83.552	85.261	83.553	83.444
	LightGBM	86.184	86.578	86.184	85.946
	HistGBDT	85.526	86.502	85.526	85.293
	Decision Tree	80.263	80.812	80.263	79.915
	Random Forest	87.500	88.477	87.500	87.379
13:00	GBDT	81.364	80.970	81.366	81.109
	LightGBM	88.819	88.656	88.820	88.644
	HistGBDT	88.198	88.257	88.199	87.865
	Decision Tree	80.124	79.960	80.124	79.730
	Random Forest	86.956	87.221	86.957	86.687
17:00	GBDT	85.906	85.826	85.906	85.658
	LightGBM	86.577	86.594	86.577	86.590
	HistGBDT	77.181	77.095	77.181	75.756
	Decision Tree	85.906	85.753	85.906	85.724
	Random Forest	90.604	90.862	90.604	90.421
19:00	GBDT	92.796	92.696	92.797	92.540
	LightGBM	93.220	93.266	93.220	92.905
	HistGBDT	93.644	93.699	93.644	93.368
	Decision Tree	87.288	86.709	87.288	86.655
	Random Forest	91.525	91.454	91.525	91.015

4.2 Feasibility of Supervised Clustering Using SHAP Values

In our experiments, we were unable to distinguish stations from the air quality dataset at specific targeted time points when projected in two-dimensional space using UMAP (Fig. 1A; top left panel). To enhance station characterization from a range of air contaminants, we used a supervised clustering approach, converting raw data into SHAP values from an optimal tree-based trained ML model. We utilized the XAI- SHAP approach to offer insights into the processes behind the contribution of these factors to certain area assignments and examined the intersections between affected areas using the UMAP technique (Fig. 1B; right panel). Based on our analysis of the spatial and temporal dynamics of air pollution variables in HCMC, we have discovered that using SHAP embedding plots for interference mapping between monitoring stations can significantly improve the precision in categorizing the city's neighborhoods. This also enables us to examine the impact of increased pollution levels in different regions on the primary categorization area.

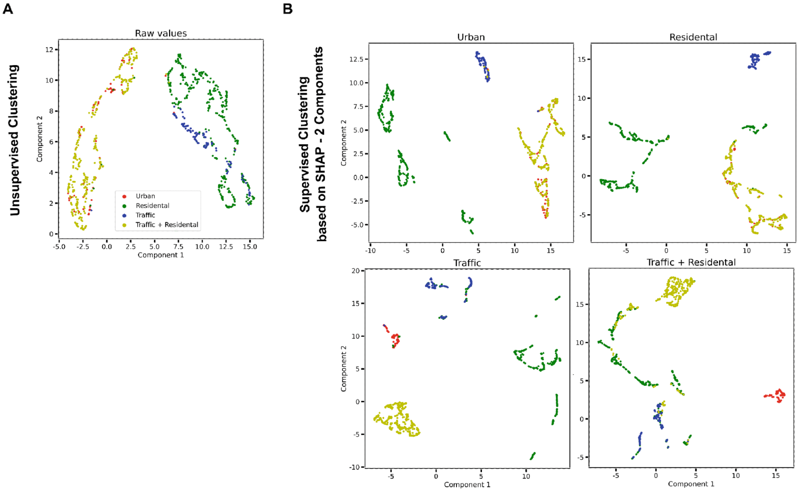

Fig. 1. UMAP visualization and SHAP-based supervised clustering for air quality monitoring stations for O_3 air pollution levels at the 7 AM targeted time point. (A) The unsupervised UMAP projection of stations from the air quality dataset, highlighting the difficulty in distinguishing stations based on raw data alone. (B) Supervised clustering approach using SHAP values derived from tree-based trained ML models, demonstrating improved station characterization based on air pollutant sets. This visualization provides insights into the processes under-lying the contributions of specific variables to station assignments and reveals the intersections between affected areas.

5 Conclusion

In this work, we present a supervised clustering approach based on average Shapley Additive exPlanations (SHAP) values to investigate the impact of various air pollutant factors in Ho Chi Minh City (HCMC), Vietnam. By employing a feature set from Rakholia et al. (2023) in tree-based machine learning models and using the eXplainable artificial intelligence (XAI)-SHAP approach along with the Uniform Manifold Approximation and Projection (UMAP) technique, we can gain a deeper understanding of the influence of various factors and the interplay between impacted regions. The benefits of our proposed methodology are as follows:

- *Enhanced classification performance*: Improved accuracy and precision in categorizing air pollution levels.
- *Interpretability*: The use of SHAP values allows for better understanding and explanation of model predictions.
- *Visualization*: The combination of SHAP and UMAP techniques provides an effective visualization of the relationships between variables and their impacts on air quality.
- *Adaptability*: The methodology can be easily adapted to other datasets, locations, and environmental challenges.

Acknowledgements. This work is part of the MSCA Career-FIT PLUS fellowship, funded by the Enterprise Ireland and the European Commission (Fellowship Ref. Number: MF20200157).

References

1. Mills, N.L., Donaldson, K., Hadoke, P.W., Boon, N.A., MacNee, W., Cassee, F.R., et al.: Adverse cardiovascular effects of air pollution. Nat. Clin. Pract. Cardiovasc. Med. **6**(1), 36–44 (2009)
2. Perez Velasco, R., Jarosinska, D.: Update of the WHO global air quality guidelines: systematic reviews - An introduction. Environ. Int. **170**, 107556 (2022)
3. Lelieveld, J., Evans, J.S., Fnais, M., Giannadaki, D., Pozzer, A.: The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature **525**(7569), 367–371 (2015)
4. Amann, M., Klimont, Z., An Ha, T., Rafaj, P., Kiesewetter, G., Gomez Sanabria, A., et al.: Future air quality in Ha Noi and Northern Vietnam. IIASA Research Report. Laxenburg, Austria (2019)
5. Fan, P., Ouyang, Z., Nguyen, D.D., Nguyen, T.T.H., Park, H., Chen, J.: Urbanization, economic development, environmental and social changes in transitional economies: Vietnam after Doimoi. Landsc. Urban Plan. **187**, 145–155 (2019)
6. Xu, Z., Dun, M., Wu, L.: Prediction of air quality based on hybrid grey double exponential smoothing model. Complexity **2020**, 1–13 (2020)
7. Rakholia, R., Le, Q., Vu, K., Ho, B.Q., Carbajo, R.S.: AI-based air quality $PM_{2.5}$ forecasting models for developing countries: a case study of Ho Chi Minh City, Vietnam. Urban Climate **46**, 101315 (2022)
8. Rakholia, R., Le, Q., Ho, B.Q., Vu, K., Carbajo, R.S.: Multi-output machine learning model for regional air pollution forecasting in Ho Chi Minh City, Vietnam. Environ. Int. **173**, 107848 (2023)
9. Joharestani, M.Z., Cao, C., Ni, X., Bashir, B., Talebiesfandarani, S.: $PM_{2.5}$ prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmosphere **10**(7), 373 (2019)
10. Zhong, J., Zhang, X., Gui, K., Wang, Y., Che, H., Shen, X., et al.: Robust prediction of hourly $PM_{2.5}$ from meteorological data using LightGBM. Nation. Sci. Rev. **8**(10), nwaa307 (2021)
11. Li, T., Hua, M., Wu, X.: A hybrid CNN-LSTM model for forecasting particulate matter ($PM_{2.5}$). IEEE Access **8**, 26933–26940 (2020)
12. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (2017)
13. Wu, Y., Lin, S., Shi, K., Ye, Z., Fang, Y.: Seasonal prediction of daily $PM_{2.5}$ concentrations with interpretable machine learning: a case study of Beijing, China. Environ. Sci. Pollut. Res. **29**(30), 45821–45836 (2022)
14. Gu, Y., Li, B., Meng, Q.: Hybrid interpretable predictive machine learning model for air pollution prediction. Neurocomputing **468**, 123–136 (2022)
15. García, M.V., Aznarte, J.L.: Shapley additive explanations for NO_2 forecasting. Eco. Inform. **56**, 101039 (2020)
16. Rakholia, R., Le, Q., Vu, K.H.N., Ho, B.Q., Carbajo, R.S.: Outdoor air quality data for spatiotemporal analysis and air quality modelling in Ho Chi Minh City, Vietnam: a part of HealthyAir Project. Data Brief **46**, 108774 (2023)