REVEALING TOLUENE BEHAVIOUR IN THE ATMOSPHERE BASED ON COUPLING OF METAHEURISTICS, XGBOOST, AND SHAP

Gordana Jovanović[1,2]*,
Mirjana Perišić[1,2],
Svetlana Stanišić[2],
Nebojša Bačanin-Džakula[2],
Andreja Stojić[1,2]

[1]Institute of Physics Belgrade,
Belgrade, Serbia

[2]University Singidunum,
Belgrade, Serbia

Abstract:

This study used an improved version of the reptilian search algorithm to investigate atmospheric patterns of toluene and its interactions with other polluting species under different environmental conditions. Toluene is a harmful aromatic hydrocarbon known for its role in the formation of secondary atmospheric pollutants. In this study, a two-year database of hourly pollutant concentrations, such as toluene, was analysed. The results were validated against other models using metaheuristic algorithms, and Shapley's additive explanations method was used to interpret them. The findings indicated a distinct correlation between toluene and m,p-xylene, and the study described the environmental conditions that influence their interactions. Overall, this research highlights the significance of using advanced analytical techniques to better understand the relationships between pollutants and their behaviour in different environmental conditions.

Keywords:

Machine Learning, Extreme Gradient Boosting, Metaheuristics, Explainable Artificial Intelligence, Volatile Organic Compounds.

INTRODUCTION

Air pollution involves intricate processes such as the dispersion, accumulation, or deposition of pollutants, which are affected by several factors, including pollutant interactions, unevenly distributed emission sources, measurement site characteristics, and meteorological conditions. To comprehend the behaviour of pollutants and their harmful effects on human health and the environment, data-driven research is essential. The complexity of air pollution-related processes necessitates an in-depth understanding of the underlying mechanisms, which can only be achieved through data-driven research.

Toluene is a mono-substituted aromatic hydrocarbon and its primary sources are traffic exhaust, cigarette smoke, and anthropogenic activities related to fuel, paint, adhesive, cleaner, polish, rubber, and lacquer production and use. Previous studies have shown that toluene concentrations range from 5 to 150 μg m^{-3} in urban locations, with extreme values near volatile pollutant sources [1]. Toluene is not prone to bioaccumulation and is rapidly absorbed, distributed throughout the body, and concentrated in vascularized organs, particularly the brain, due to its affinity for lipid-rich tissues [2].

Correspondence:
Gordana Jovanović

e-mail:
gordana.vukovic@ipb.ac.rs

DOI: 10.15308/Sinteza-2023-17-22

Toluene concentration is regularly monitored due to its toxicity on the nervous system and permanent brain damage observed in adhesive abusers [3]. Toluene is not labelled as carcinogenic, but in urban locations, anthropogenic benzene and toluene emission sources play a substantial role in ozone photochemistry and SOA-forming contribution, especially in low-NO_x regimes [4].

Our previous studies have tackled issues in analysing air pollution in complex urban environments, including the need for proper contextualization of data [5], [6], [7], [8], [9], and the use of statistical methods and artificial intelligence algorithms [10], [11] and [12]. In terms of data modelling, metaheuristic algorithms are commonly used to address nondeterministic polynomial (NP)-hard problems, particularly in machine learning hyperparameter optimization, due to their stochastic nature. In this study, we apply an enhanced variant of the reptile search algorithm (RSA) being hybridized with the firefly algorithm (FA) for resolving the shortcomings of the elementary RSA. The enhanced version of RSA metaheuristics is applied as an integral component of the machine learning framework to optimize the set of the XGBoost hyperparameters for toluene atmospheric fate research. The best-produced model is interpreted by applying Shapley Additive exPlanations (SHAP).

2. METHODOLOGY

2.1. DATA

For the analysis, we used the concentrations of inorganic gaseous pollutants (NO, NO_2, NO_x, O_3), particulate matter (PM_1, $PM_{2.5}$, and PM_{10}), and benzene, toluene, mp-xylene, and total non-methane hydrocarbons (TNMHC) obtained from the station of regulatory air quality monitoring Vatrogasni Dom in Pančevo, Serbia [13]. Additionally, meteorological parameters attained from the Global Data Assimilation System – GDAS1 [14], were used to complement the two-year (2019-2020) database of air pollutants. Hourly concentrations of organic pollutants (benzene, toluene, and m,p-xylene) and inorganic gaseous pollutants (NO, NO_2, NO_x, and O_3) were measured using referent sampling devices that adhere to European standards EN 14662-3, EN 14211, and EN 14625. The GRIMM EDM 180 measuring method was used to determine hourly concentrations of particulate matter, following the standards EN 12341 and EN 14907, while a gas chromatograph Syntech Spectras GC955 was employed for the concentrations of TNMHC. This device separates methane from other hydrocarbons and measures the concentration of both methane and other total non-methane hydrocarbons in the air.

2.2. STUDY AREA

Pančevo, with over 100,000 inhabitants, is situated on the left bank of the Danube, 20 km east and northeast of Belgrade, the largest Serbian metropolitan area. The sampling site (44°51′31″ N, 20°38′56″ E) is an urban background station located about 500 m south of the city centre at the regional fire station. The surrounding areas include residential areas to the east and northeast, a scrap metal sorting and storage centre, and a flour production factory. The E70, a European corridor with public transport and intensive vehicle flow, passes about 200 m in the S-SW direction from the sampling site. The confluence of the Tamiš and Danube rivers is located approximately 500 m in the SW direction. The South industrial zone of Pančevo, which includes three main factories: HTP Azotara, HTP Petrohemija, and Pančevo Oil Refinery, is situated two kilometres SE of the sampling station. The station is positioned in the dominant southeast direction of wind between the industrial zone and the city centre, according to the Air quality control program for the City of Pančevo and the Air Quality Plan for the City of Pančevo.

2.3. EXTREME GRADIENT BOOSTING – XGBOOST

XGBoost is a machine learning algorithm based on an ensemble of decision trees, where each tree is trained to correct the errors of the previous tree in the sequence. One of the key advantages of XGBoost is its ability to handle large datasets with high-dimensional features. It employs a regularization technique to prevent overfitting and can handle missing values in the data. The algorithm is highly customizable, allowing for the tuning of parameters such as learning rate, maximum depth, and number of trees to optimize performance. The details about XGBoost are provided elsewhere [15].

2.4. METAHEURISTICS

NP-hard challenges are a frequent occurrence that often requires the use of stochastic algorithms like metaheuristics because deterministic methods are impractical. Metaheuristic algorithms can be classified into various families based on the natural phenomena they imitate to guide the search process, such as evolution or insect behaviour [15]. The most significant families are nature-inspired methods (genetic algorithms and swarm intelligence), physical phenomenon-based methods (such as storms, gravity, and electromagnetism), algorithms that imitate human behaviour, and approaches based on mathematical laws.

Swarm intelligence is based on the coordinated and sophisticated behavioural patterns manifested by large groups of relatively modest units, such as insects or birds in swarms, while they hunt, feed, mate, or migrate [16]. These algorithms have proven highly efficient in solving various real-world NP-hard challenges. Well-known examples include particle swarm optimization (PSO) [17], ant colony optimization (ACO) [18], firefly algorithm (FA) [19] and bat algorithm (BA) [19]. More recently, highly efficient algorithms based on mathematical functions and their properties have emerged, such as the sine-cosine algorithm (SCA) [20] and arithmetic optimization algorithm (AOA) [21].

In this paper, we used a modified reptile search algorithm (RSA) inspired by crocodiles' hunting style [22] the RSA lacked sufficient exploitation power despite excellent exploration capability. We found the diversification-intensification trade-off balance biased towards exploration. We proposed integrating RSA with the FA to achieve a suitable balance between exploration and exploitation. The low-level hybrid approach combines both metaheuristics, with RSA at the start and FA during the search process to enhance the RSA's performance. The approach addressed RSA's weaknesses and improved its effectiveness in identifying optimal search regions.

2.5. SHAPLEY ADDITIVE EXPLANATIONS

To gain insight into the decision-making process of a best-performing model, we employed the explainable artificial intelligence SHAP (SHapley Additive exPlanations) method [23]. SHAP allows for a meaningful and straightforward interpretation of the decisions derived from the model, without sacrificing accuracy or interpretability. It is based on a game-theory approach that calculates Shapley values as a feature importance measure, which provides an understanding of the impact of each feature on individual predictions.

The Shapley values represent fairly distributed payouts among the cooperating players (features) depending on their contribution to the joint payout (prediction). SHAP assigns an important measure to each feature as a measure of its contribution to a particular prediction and compares its impact to the model's prediction if that feature took some baseline value (mean). This provides valuable insights into the model's behaviour by overcoming the main drawback of inconsistency, minimizing the possibility of underestimating the importance of a feature with a specific attribution value, and capturing feature interaction effects. However, the main challenges of the method include the computation of Shapley values and the choice of background data, which can lead to uncertain or unintuitive feature attributions.

3. RESULTS

As shown by mean absolute and relative SHAP values (Table1), the concentrations of benzene, followed by m,p-xylene levels, appeared to be the major factors that shape the toluene dynamic in the air. Additionally, the toluene's environmental fate is affected by the concentrations of THNMC, PM1, and NOX, as well as meteorological parameters, including volumetric soil moisture content (SOLM) and the direction and intensity of momentum flux (MOFD and MOFI). In the present study, to demonstrate the potential of the applied methodology, we will focus on m,p-xylene as the main predictor.

Table 1 - SHAP values.

	Benzene	mpXylene	TNMHC	PM_1	NO_x	SOLM	MOFD	MOFI	NO_2	NO	PM_{10}	T02M
Absolute SHAP	1.28	0.85	0.25	0.11	0.09	0.07	0.05	0.04	0.03	0.03	0.03	0.03
Relative SHAP [%]	36.09	27.27	8.31	3.06	3.05	2.43	2.01	1.49	1.12	1.23	0.94	0.91

The results suggest that m,p-xylene concentrations of 0.85 μg m^{-3} on average govern the toluene dynamics in the air as shown by absolute SHAP values. The most positive impact of m,p-xylene on the toluene dynamics is accompanied by the increase of m,p-xylene levels (up to 5 μg m^{-3}) as well as by the lowest concentrations of the inorganic gases, volatile non-aromatics and particles. The interrelationships between benzene homologues, m,p-xylene and toluene, could be explained by their coexistence in ambient air. Although toluene contains one methyl group which can be placed at any position on the benzene ring and m,p-xylene has two methyl groups attached to the benzene ring, they share common emission sources. The most dominant sources originated from anthropogenic activity including the petrochemical industry, chemical production of organic solvents and evaporative emission from storage facilities, but also to a lesser extent combustion of fossil fuels for heating and traffic purposes. The low to moderate levels of other pollutants in environmental conditions are associated with a negative impact of m,p-xylene, resulting in a decrease in toluene levels of up to 2.7 μg m^{-3}, while the highest concentrations of other pollutants are mostly associated with a moderate positive and negative impact on toluene dynamic (from -1.6 to 2 μg m^{-3}). Positive impact suggests the occasional influence of common sources of toluene and NOX and THNMC whereas negative interrelations imply different sources and the possible different behaviour of the pollutants in the air.

When PM is present in the air, the photochemical oxidation of aromatics such as toluene contributes to SOA formation (95%) compared to volatile organics and alkenes [24]. Similarly, Zhan et al. [25] reported that aromatics dominantly lead to the production of ground-level O_3.

The increase in m,p-xylene levels is linearly correlated with PM_1, $PM_{2.5}$, TNMHC, and NO_2 concentrations, but not with benzene and NO (Figure 1-3). High levels of NO were observed when lower levels of m,p-xylene were recorded (Figure 2), whereas higher concentrations of m,p-xylene corresponded to lower values of benzene (Figure 3) indicating the impact of different emission sources surrounding the measuring site. Additionally, an area of moderate influence of m,p-xylene on toluene is observed, when m,p-xylene concentrations are in the interval from 10 to 25 μg m^{-3}, and NO records high values - above 100 μg m^{-3} (Figure 2).

The analysis shows that low air and soil temperatures, high relative humidity, low PBLH, and stable atmospheric conditions as indicated by the other meteorological parameters, accompany the highest levels of all analysed polluted species. The conditions could be associated with the cold part of the year (winter and autumn months) when unfavourable meteorological conditions together with intensified fossil fuel burning for heating purposes contribute to high concentrations of pollutants. In addition, toluene and m,p-xylene removal through photochemically induced reactions are suppressed during the cold periods.

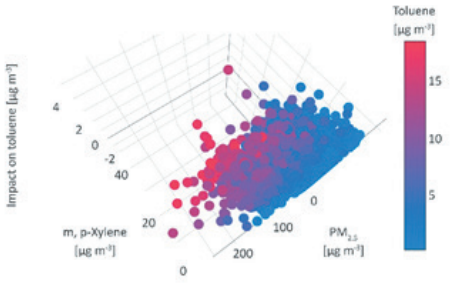

Figure 1 – Absolute m, p-xylene impact on toluene in the context of $PM_{2.5}$.

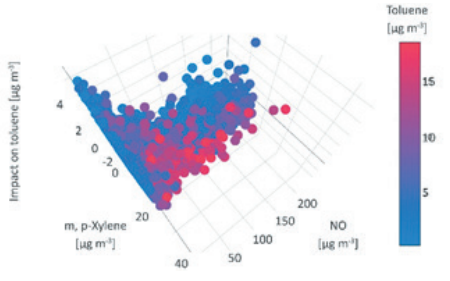

Figure 2 – Absolute m, p-xylene impact on toluene in the context of NO.

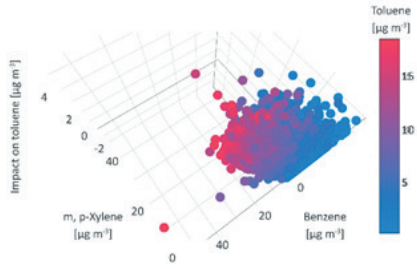

Figure 3 - Absolute m, p-xylene impact on toluene in the context of benzene.

4. CONCLUSION

Toluene is widespread in the atmosphere because it is used in many commercial products and when it is present at high concentrations, this pollutant poses serious adverse effects on human health. The behaviour of toluene in air is complex and its lifetime and abundance highly depend on environmental factors including meteorological conditions, the presence and intensity of emission sources and interactions with the other (co)polluting compounds. The present study using metaheuristics, XGBoost, and SHapley Additive exPlanations methods showed that concentrations of m,p-xylene mainly impact the dynamic of toluene in air. The positive interrelations between toluene and m,p-xylene could be linked with common emission sources and favourable values of temperature, humidity and PBLH.

5. ACKNOWLEDGEMENTS

The authors acknowledge funding provided by the Institute of Physics Belgrade, through the grant by the Ministry of Education, Science and Technological Development of the Republic of Serbia, the Science Fund of the Republic of Serbia GRANT No. #6524105, AI – ATLAS.

6. REFERENCES

[1] D. Murindababisha, Y. Abubakar, S. Yong, W. Chengjun and R. Yong, "Current progress on catalytic oxidation of toluene: a review," *Environmental Science and Pollution Research*, pp. 1-31, 2021.

[2] C. Davidson, J. Hannigan and S. Bowen, "Effects of inhaled combined Benzene, Toluene, Ethylbenzene, and Xylenes (BTEX): Toward an environmental exposure model," *Environmental toxicology and pharmacology,* vol. 81, p. 103518, 2021.

[3] N. P. Cheremisinoff and P. E. Rosenfeld, "Sources of air emissions from pulp and paper mills," *Handbook of pollution prevention and cleaner production*, vol. 2, pp. 179-259, 2010.

[4] G. Z. Whitten, H. Gookyoung, K. Yosuke and E. McDonald-Bul, "A new condensed toluene mechanism for Carbon Bond: CB05-TU," *Atmospheric Environment,* vol. 44, no. 40, pp. 5346-5355, 2010.

[5] L. Jovanovic, G. Jovanović, M. Perisic, F. Alimpic, S. Stanisic, N. Bacanin, M. Zivkovic and A. Stojic, "The Explainable Potential of Coupling Metaheuristics-Optimized-XGBoost and SHAP in Revealing VOCs' Environmental Fate," *Atmosphere*, vol. 14, no. 1, p. 109, 2023.

[6] A. Stojić, G. Vuković, M. Perišić, S. Stanišić and A. Šoštarić, "Urban air pollution: an insight into its complex aspects," in *A Closer Look at Urban Areas,* NY, USA, Nova Science Publishers, 2018.

[7] Š. Andrej, S. Stojić Stanišić, G. Vuković, Z. Mijić, A. Stojić and I. Gržetić, "Rainwater capacities for BTEX scavenging from ambient air," *Atmospheric Environment*, vol. 168, pp. 46-54, 2017.

[8] S. Andreja, N. Stanić, G. Vuković, S. Stanišić, M. Perišić and A. Šoštarić, "Explainable extreme gradient boosting tree-based prediction of toluene, ethylbenzene and xylene wet deposition," *Science of The Total Environment*, vol. 653, pp. 140-147, 2019.

[9] S. Stanišić, G. Jovanović, M. Perišić , S. Herceg-Romandić, T. Milićević and A. Stojić, "Explaining the Environmental Fate of PAHs in Indoor and Outdoor Environments by the Use of Artificial Intelligence," in *Polycyclic aromatic hydrocarbons Hauppauge*, NY, USA, Nov Science, 2022, pp. 1-36.

[10] A. Stojić, D. Maletić, S. Stojić Stanišić, Z. Mijić and A. Šoštarić, "Forecasting of VOC emissions from traffic and industry using classification and regression multivariate methods," *Sciente of the total environment,* pp. 19-26, 2015.

[11] M. Perišić, D. Maletić, S. Stojić Stanišić, S. Rajšić and A. Stojić, "Forecasting hourly particulate matter concentrations based on the advanced multivariate methods," *International Journal of Environmental Science and Technology*, vol. 14, no. 5, p. 1047–1054, 2017.

[12] S. Stanišić, M. Perišić, G. Jovanović, D. Maletić and D. Vudragović, "What Information on Volatile Organic Compounds Can Be Obtained from the Data of a SingleMeasurement Site Through the Use of Artificial Intelligence?," in *Artificial Intelligence: theory and Applications*, Springer, 2021, pp. 207-225.

[13] "SEPA," [Online]. Available: http://www.amskv.sepa.gov.rs/.

[14] "GDAS1," [Online]. Available: https://www.ready.noaa.gov/gdas1.php.

[15] C. Tianqi and C. Guestrin, "Xgboost: A scalable tree boosting system," in *The 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016.

[16] B. G., "Swarm intelligence," *Complex Social and Behavioral Systems: Game Theory and Agent-Based Models,* pp. 791-818, 2020.

[17] J. Kennedy and Enneby, "Particle swarm optimization," in *ICNN'95-international conference on neural networks,* 1995.

[18] M. Dorigo, M. Birattari and T. Stutzle, "Ant colony optimization," *IEEE computational intelligence magazine,* vol. 1, no. 14, pp. 28-39, 2006.

[19] X. Yang, "Firefly algorithms for multimodal optimization," in *In Stochastic Algorithms: Foundations and Applications: 5th International Symposium, SAGA 2009*, Sapporo, Japan, 2009.

[20] S. Mirjalili, "SCA: a sine cosine algorithm for solving optimization problems," *Knowledge-based systems,* vol. 96, pp. 120-133, 2016.

[21] L. Abualigah, A. Diabat, C. Mirjalil and M. Abd Elaziz, "The arithmetic optimization algorithm," *Computer methods in applied mechanics and engineering*, vol. 376, p. 113609, 2021.

[22] L. Abualigah, M. Abd Elaziz, . P. Sumar, Z. Geem and Gandomi, "Reptile Search Algorithm (RSA): A nature-inspired meta-heuristic optimizer," *Expert Systems with Applications*, vol. 191, p. 116158.

[23] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[24] J. Li, S. Deng, G. Li, Z. Lu, H. Song, J. Gao, Z. Sun and K. Xu, "VOCs characteristics and their ozone and SOA formation potentials in autumn and winter at Weinan, China," *Environmental Research*, vol. 203, p. 111821, 2022.

[25] J. Zhan, Z. Feng, P. Liu, X. He, Z. He, T. Chen, Y. Wang, H. He, Y. Mu and Y. Liu, "Ozone and SOA formation potential based on photochemical loss of VOCs during the Beijing summer," *Environmental Pollution*, vol. 285, p. 117444, 2021.