EXPLAINABLE ARTIFICIAL INTELLIGENCE AND SPATIAL ANALYSIS FOR SUSTAINABLE URBAN PLANNING ADDRESSING AIR POLLUTION

Mateusz Zareba[1], Tomasz Danek[2], Monika Chuchro[3]

AGH University of Krakow, Faculty of Geology, Geophysics and Environmental Protection, Department of Geoinformatics and Applied Computer Science, Krakow, Poland, e-mail: zareba@agh.edu.pl, tdanek@agh.edu.pl, chuchro@agh.edu.pl

[1] https://orcid.org/0000-0002-9663-6593 , [2] https://orcid.org/0000-0001-8101-7469 , [3] https://orcid.org/0000-0002-0381-4697

Abstract: This study examines the application of Explainable Artificial Intelligence (XAI) and spatial analysis to comprehend the seasonal and spatial dynamics of air pollution, thereby supporting sustainable urban development. The analysis is based on a unique, high-resolution dataset collected from 52 low-cost sensors deployed across Kraków (Poland) and its surrounding municipalities, generating over 450,000 hourly samples annually. The region offers a distinct spatial contrast: Kraków enforces a strict ban on solid fuel use for residential heating, while neighboring areas do not, creating a natural experimental setting for examining the spatial interplay of emission sources and meteorological conditions. Machine learning models, including XGBoost and Extra Trees Regressor (ETR), were utilized to evaluate the significance of various meteorological and environmental predictors of PM2.5 concentrations. Feature importance was averaged seasonally to detect temporal trends, and the most relevant variables were selected for each season. These were then mapped using the Kriging algorithm to explore spatial variability in predictor relevance. The results revealed strong seasonal patterns and localized differences in the influence of variables such as temperature, relative humidity, and wind speed, reflecting both meteorological processes and anthropogenic emission dynamics. This research highlights the critical role of integrating seasonal context and spatial heterogeneity into air quality modeling. By combining interpretable machine learning with spatial mapping, the study offers actionable insights for urban planners and policymakers aiming to improve air quality management. The findings demonstrate how XAI methods can support evidence-based strategies for healthier and more sustainable cities.

Keywords: air pollution; explainable artificial intelligence; smart city; predictors impor-tance; machine learning, spatial analysis

1 INTRODUCTION

A smart city is an urban area that relies on a set of advanced digital technologies to enhance the efficiency of resource management, urban services, and citizen engagement. According to the European Union’s official declaration, smart cities represent a key component in shaping the digital future of all EU member states. Their primary goal is to optimize the use of resources through advanced monitoring systems for energy consumption, water usage, and other essential utilities, as well as transportation networks. Notably, the European Union’s policy framework for smart cities also emphasizes the role of interactive governance, promoting digital administration tools and increasing citizen participation in decision-making processes (European Commission, 2025).

A particularly significant aspect of smart cities, as highlighted in recent research, is their impact on environmental sustainability. Studies underscore the necessity of integrating environmental strategies into smart city planning to ensure that technological advancements contribute to ecological well-being rather than operate independently of sustainability goals. For instance, Jonek-Kowalska (2023) discusses the critical role of environmental considerations in smart city development. Additionally, critical analyses have pointed out the risk of detaching smart city initiatives from broader environmental strategies, further reinforcing the need for tools that facilitate environmental monitoring and assessment within smart city frameworks (March and Ribera-Fumaz, 2014). This suggests that smart city development must not only focus on technological advancements but also incorporate comprehensive sustainability mechanisms to address environmental challenges effectively.

Air pollution remains a critical global health challenge, contributing not only to a broad spectrum of medical conditions but also to substantial environmental degradation. A growing body of research has demonstrated that exposure to particulate matter (PM) is directly associated with numerous adverse health effects, significantly increasing both morbidity and mortality rates worldwide (Cohen et al., 2017). Among the well-documented consequences of PM exposure are respiratory disorders, such as chronic obstructive pulmonary disease (COPD) and asthma, as well as cardiovascular complications that elevate the risk of stroke and hypertension (Weinmayr et al., 2010). Epidemiological studies have identified a correlation between PM and neurodegenerative conditions, including Alzheimer’s and Parkinson’s disease, suggesting that prolonged exposure to airborne pollutants may accelerate cognitive decline (Thurston et al., 2017). This issue is particularly pressing in nations with rapidly aging populations, where the growing prevalence of such disorders places an increasing burden on already strained healthcare and social support systems.

The health risks associated with PM exposure can be classified into short-term and long-term effects, both of which have profound implications.

Beyond its impact on human health, air pollution represents a significant ecological threat, disrupting ecosystems and threatening biodiversity. Studies indicate that airborne pollutants adversely affect soil and water quality, thereby impairing agricultural productivity and destabilizing natural habitats. Prolonged exposure to toxic substances in the atmosphere has also been linked to reduced fertility rates in wildlife, interfering with reproductive cycles and leading to long-term population declines among various species (Manisalidis et al., 2020).

Given the widespread consequences of air pollution, it is evident that addressing this issue requires a multidisciplinary approach, incorporating advancements in environmental policy, public health interventions, and technological innovations aimed at pollution reduction, especially focusing on smart cities implementation. Effective mitigation strategies should not only focus on limiting emissions but also emphasize real-time air quality monitoring, public awareness campaigns, and urban planning initiatives that promote cleaner and more sustainable living environments through science and common understanding of cause-effect phenomena. This objective can be effectively realized through the application of big data analysis and an explainable artificial intelligence (XAI) framework, as outlined in this study. By embedding this approach within a spatial and temporal context, it allows for a deeper understanding of the evolving relationships between environmental, social, and infrastructural factors in specific geographical settings, considering regional characteristics, urbanization patterns, and local policy frameworks.

This study leverages XAI and big data analytics to examine the role of various environmental and anthropogenic predictors influencing air pollution dynamics in Kraków, Poland. The city, located in Central Europe within a moderate climate zone, experiences four astronomical seasons – spring (March-May), summer (June-August), autumn (September-November), and winter (December-February) – along with six thermal seasons as classified by Gumiński (1950). These seasonal variations significantly affect pollution levels, leading to the identification of two dominant pollution periods: warm and cold seasons (Zareba et al., 2024).

Kraków presents a complex case study for air quality assessment due to both natural and human-induced factors. The city's topography, situated in a valley with the Vistula River running through its center and surrounded by hills, impacts pollutant dispersion and accumulation, with prevailing westerly winds further shaping air quality patterns (Danek et al., 2022). On the anthropogenic side, Kraków has undergone a significant transition in energy usage, including a ban on coal-based heating, while simultaneously experiencing continuous population growth. Despite the shift towards cleaner energy, pollution sources remain a concern, particularly in winter when external pollution transport intensifies. PM2.5 composition analysis indicates that carbon-based particles account for over 40% of pollution, with coal combustion contributing up to 50% in winter and declining to 20% in summer, while vehicle emissions and natural sources fluctuate seasonally (Bokwa, 2008; Wojewódzki Inspektorat Ochrony Środowiska w Krakowie, 2020) Given the spatial and temporal complexity of air pollution trends, this study employs XAI techniques to evaluate the relative importance of key predictors across individual months. While traditional air quality studies often focus on seasonal or annual averages, this research emphasizes month-to-month variations, offering a more granular understanding of the interactions between meteorological conditions, emission sources, and regulatory interventions. By integrating machine learning (ML) models with explainability techniques, the study identifies how factors such as wind patterns, temperature, humidity, and human activities drive air pollution fluctuations at different times of the year.

This study incorporates optical 52 low-cost sensors (LCS) from Airly (http://airly.eu), which were evaluated within the LIFE Integrated Project focused on improving air quality in the Małopolska region (Małopolska Region, n.d.) – see Figure 1. These sensors, while not officially recognized for regulatory reporting under EU Directive 2008/50/EC, provide valuable real-time data that can be calibrated using ML and AI-based adjustments to enhance reliability (Danek et al., 2022). The explainability component of the study ensures that the influence of various pollution drivers is transparent and interpretable, addressing concerns about the black-box nature of AI models in environmental monitoring.

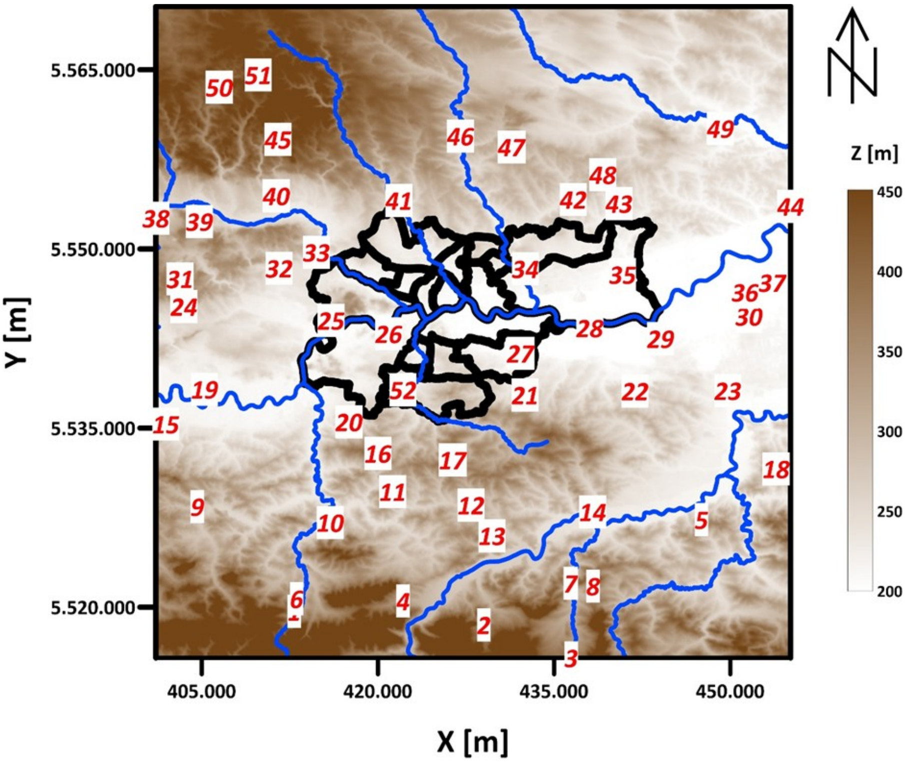


Figure 1 The digital terrain model of Kraków city and its surrounding areas, with district boundaries marked in black, Airly LCS sensor locations labeled in red over white, and rivers highlighted in blue. Source: Terrain data provided by the European Union via the Copernicus Land Monitoring Service (2022) and the European Environment Agency (EEA); base map from OpenStreetMap (https://www.openstreetmap.org)

This study explores the varying influence of key environmental and anthropogenic factors on air pollution across different months, emphasizing the need for adaptable analytical approaches. By assessing the relative importance of predictors over time, the research highlights seasonal fluctuations in their impact, reflecting the complex interplay between meteorological conditions and emission sources. Such an approach provides a structured framework for enhancing data-driven assessments, offering a deeper understanding of pollution dynamics in urban environments with similar climatic and infrastructural characteristics.

In addition to the data-driven approach presented here, previous research conducted in Kraków has employed classical methods – including geographically weighted regression and local spatial autocorrelation (Danek et al., 2022), backward trajectory analysis with the NOAA HYSPLIT model (Stein et al., 2015; Zareba and Danek, 2022), and the CALPUFF puff dispersion model (Godłowska et al., 2022) – to investigate the impact of topography, meteorology, and anthropogenic activity on air pollution. These studies confirmed the predominantly anthropogenic origin of particulate matter, particularly during the heating season, and analyzed socioeconomic factors, including heating practices across municipalities. Crucially, the complex interplay between pollutant production and dispersion processes within the urban environment strongly motivates the use of XAI methods. Building on these foundations, this study uses XAI as a tool to complement – but not replace – traditional models, enabling a deeper, more interpretable understanding of local air pollution dynamics.

2 MATERIALS AND METHODS

Explainable Artificial Intelligence

In recent years, interest in artificial intelligence (AI) techniques has grown steadily across various fields, including medicine (Rahmani et al., 2021), geology (Zareba et al., 2023) and spatial analysis (Kopczewska, 2022). In the natural and experimental sciences, ML methods are particularly important for tasks such as prediction, classification, and the discovery of hidden patterns in data, especially through unsupervised learning. Deep Learning (DL) is subset of ML involves neural network-based methods, which often consist of multiple hidden layers. These models typically offer high predictive accuracy but are often treated as "black boxes", meaning their internal decision-making processes are not easily understood (Sarker, 2021). This lack of transparency can be problematic in certain applications.

Two related concepts are important in this context: interpretability and explainability. Interpretability refers to the extent to which a model's structure and parameters can be directly examined to understand how it works. This is usually possible for simpler models and does not require external tools. In contrast, explainability relates to more complex models and involves using specific methods and tools to understand why a model produces a particular output. Although these methods have limitations, they can provide insights that help interpret the modeled phenomenon (Lipton, 2016; Samek et. al, 2017). This study focuses on explainability methods that support spatial interpretation of model behavior. This approach enables the identification of new characteristics related to the spread of atmospheric particulate matter.

Geo-Data Science Framework

In this study, a complex geo-data science framework was used to assess the seasonal and spatial importance of predictors for PM2.5 concentrations. The first step involved data preprocessing and sensor-level analysis. The dataset, containing air quality data from multiple sensors (52) across whole year (1-hour resolution), was first filtered to separate the data for each individual sensor. The data was then grouped by month (average 720 hours per month for each sensor), with each month's data being further split into training (first 584 observations) and testing (last 146 observations) sets. This allowed for the seasonal analysis of predictor importance across different sensors and time periods.

Three ML techniques were utilized to assess feature importance. The base was gradient boosting model (XGBoost (XGB)) which was trained using hyperparameters optimized for regression tasks (Chen and Guestrin, 2016). It was used to model the relationship between predictors (hour of day, air temperature, wind speed, relative humidity, surface pressure, precipitation) and PM2.5 concentrations. The model was trained using the training data and evaluated using both the training and testing sets. The feature importance was assessed using gain and frequency-based metrics, which were then visualized and saved for later analysis. After training the XGBoost model, permutation importance (Pedregosa et al., 2011) was calculated. This method involves randomly shuffling the values of each feature and measuring the decrease in model performance. This provides an estimate of how much each feature contributes to the model's predictive accuracy. To make the analysis more robust and accurate, a third method was applied that is based on an ensemble modeling approach. Extra Trees Regressor (ETR) uses multiple decision trees during the training to capture non-linear relationships between predictors and PM2.5 concentrations (Breiman, 2001; Scikit-learn developers, 2021). The feature importance from this model was also stored for comparison with XGBoost and Permutation Importance.

For each sensor and month, the feature importance scores obtained from each method were saved. These results were compiled into data frames and averaged in four seasons: winter, spring, summer, and autumn. Feature importance plots were made as heatmaps where all 52 LCS and seasons can be seen. These plots provided a clear representation of which features were most influential in predicting PM2.5 concentrations for each sensor during each season. The visual outputs supported the identification of seasonal patterns in the importance of variables. After calculating the average importance scores for each feature, divided by season, the results were compared using a single box plot. This step helped identify the most important features for each season, which were then visualized on a map to investigate the spatial patterns of predictor relevance. To facilitate the mapping of these spatial patterns, the data were gridded using the Kriging algorithm, allowing for a more accurate representation of the spatial distribution of feature importance across the study area.

3 RESULTS

Figure 2a shows the seasonal cloud cover predictor importance for using Permutation, XGBoost, and ETR methods. Each heatmap shows the contribution of 51 sensors across winter, spring, summer, and autumn. Across all models, sensor 17 consistently shows high importance, particularly in Spring and Summer, suggesting it captures localized atmospheric or land-use conditions strongly tied to cloud formation during active weather periods. XGBoost and ETR highlight several consistently influential sensors (e.g. 15, 17, 21), reflecting stable spatial features. Seasonal differences are pronounced, with Summer showing the highest importance concentrations, likely due to convective processes influenced by heating, vegetation, and human activities. Permutation importance shows more seasonal variability, while ETR emphasizes sharp contrasts in sensor relevance. These patterns suggest that cloud cover is modulated by both atmospheric and human-driven factors, and effective modeling requires attention to spatial heterogeneity and seasonal context. Figure 2b illustrates the seasonal importance of the time-of-day predictor for the same set of sensors. Across all methods, hour of day emerges as a highly influential predictor, particularly during Autumn and Winter, with consistently elevated importance across multiple sensors. This seasonal pattern likely reflects increased anthropogenic emissions during colder months (especially coal-based heating), potentially coupled with stable atmospheric conditions (e.g. temperature inversions) that limit pollutant dispersion. The Permutation and ETR methods highlight strong localized importance in the top sensor rows, suggesting specific locations where day patterns are more closely linked to PM2.5 changes. XGBoost tends to generalize spatial differences. Hourly variation is a critical component in PM2.5 predictions, capturing daily emission cycles and human activity rhythms that interact with atmospheric processes in a seasonally dependent manner.

Precipitation importance season map is shown in Figure 3a. XGBoost model shows the highest variation and intensity in predictor importance across all seasons, suggesting its sensitivity to seasonal dynamics and complex nonlinear interactions. This model reveals consistent high-importance features, particularly in Summer and Autumn, where human activities such as potential grass burning and increased transportation may contribute significantly to PM2.5 levels. In contrast, the Permutation Importance and ETR Importance exhibit more subtle gradients, indicating more smoothed estimates of this feature importance. The relative humidity seasonal importance is shown in Figure 3b. The XGBoost model highlights several key features with importance values reaching up to 0.22, particularly in Summer and Spring. In contrast, the ETR model shows a more uniform distribution of importance, with top values around 0.16 with some strong picks to over 0.4 (sensors 20, 26, 37 – mostly in Spring), indicating that multiple predictors share similar influence throughout the year. The Permutation Importance results are more subdued, with values rarely exceeding 0.14. The consistency of high-importance features across seasons and methods implies the presence of stable environmental drivers.

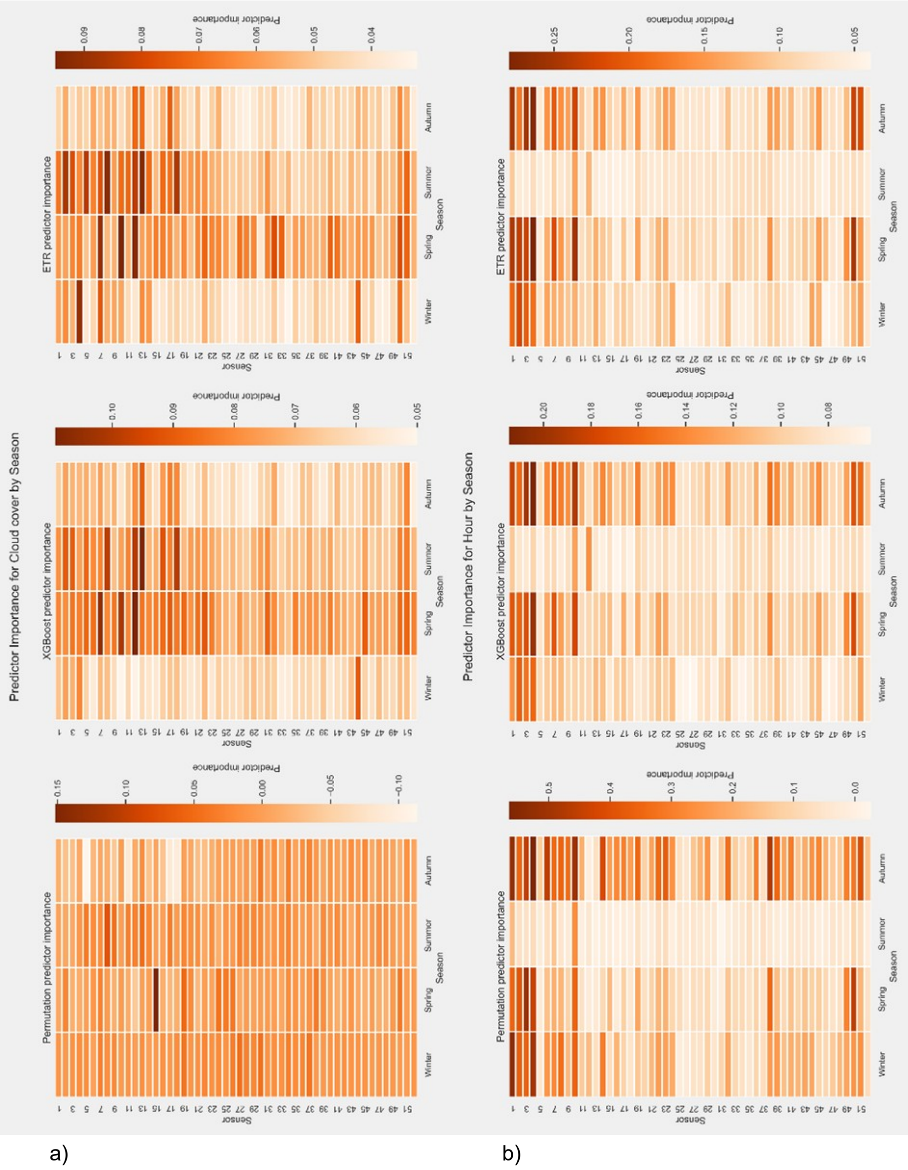
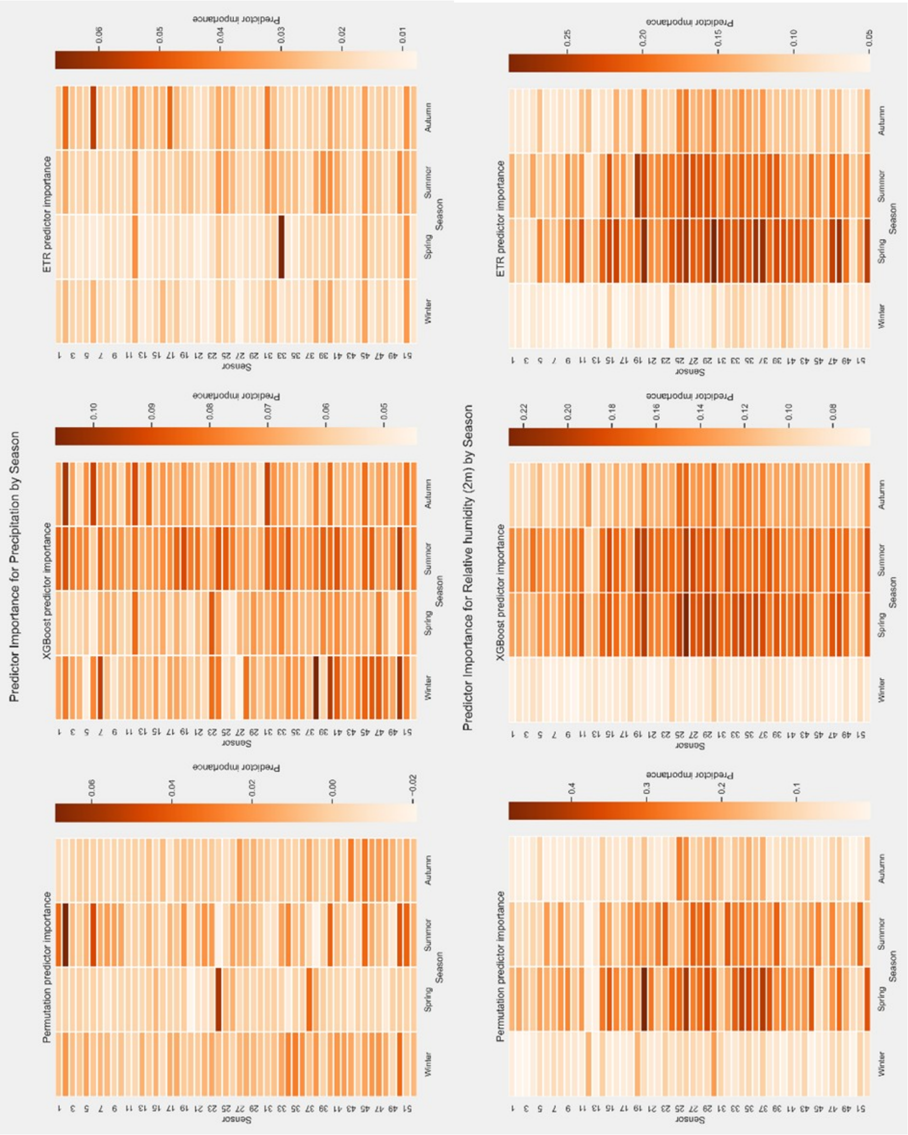


Figure 2 a) Seasonal importance of cloud cover predictors in PM2.5 concentration modeling. **b)** Seasonal importance of hour predictors in PM2.5 concentration modeling. Source: own elaboration



a)

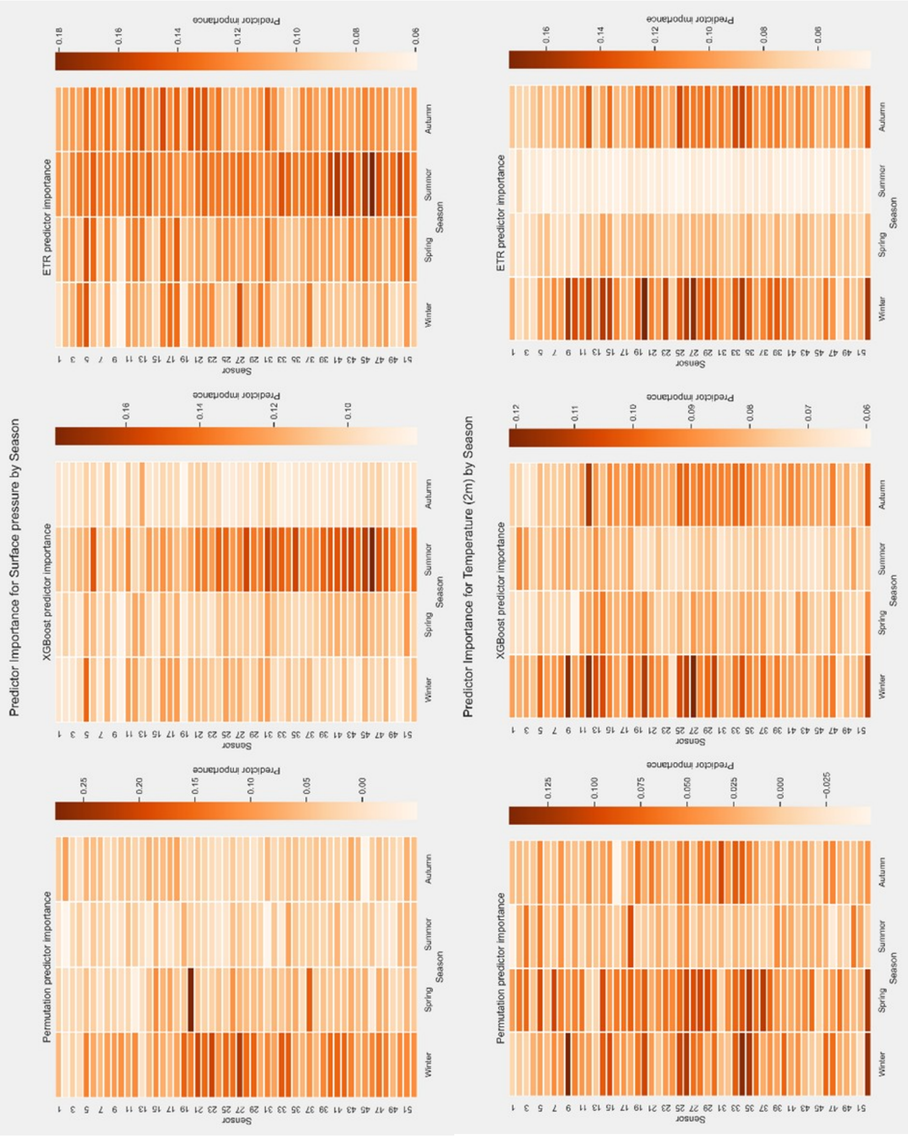
b)

Figure 3 a) Seasonal importance of precipitation predictors in PM2.5 concentration modeling. **b)** Seasonal importance of relative humidity predictors in PM2.5 concentration modeling. Source: own elaboration

Figure 4a shows seasonal predictor importance for surface pressure. The XGBoost model again displays the most dynamic range, with some predictors reaching importance values above 0.18 in Autumn and Summer, indicating strong influence from specific variables, possibly temperature gradients or altitude-related features. The ETR model shows more balanced distributions, with the most important features peaking around 0.16, suggesting a moderate but consistent influence across seasons. The Permutation method shows lower overall magnitudes (mostly under 0.12), with a few isolated spikes, particularly in Spring for sensor 20. This suggests that for surface pressure, a smaller subset of features – likely topographic or large-scale atmospheric indicators – dominates across all models. From an XAI perspective, this consistency and seasonal alignment support the interpretability and robustness of model decisions. Figure 4b reveals notable seasonal variation in near-surface air temperature (2m) predictor importance, with winter consistently showing higher and more concentrated importance values across sensors. LCS 23 seems to be the most influenced by temperature in the winter, exhibiting the highest relative importance across all three methods, with values reaching approximately 0.13–0.17 depending on the method. LCS such as 5, 11, and 17 also demonstrate consistent relevance across multiple seasons, particularly in winter and spring. In contrast, summer and autumn display significantly lower and more diffuse importance scores, suggesting a more complex or less deterministic relationship between PM2.5 predictions and temperature during.

The analysis of predictor importance for wind speed (10m) in Figure 5 again highlights significant spatial and seasonal variability in sensors. During winter, certain sensors demonstrate markedly higher importance values, reaching up to 0.28 in permutation and ETR approaches, indicating a strong spatial signal from specific sensor locations in colder months. Sensor 50, in particular, shows elevated importance in both winter and autumn across all methods, suggesting a potentially critical spatial zone for capturing wind-related transport or dispersion of PM2.5. Spring and summer generally exhibit reduced and more uniform importance levels, pointing to either diminished wind-PM2.5 coupling or greater homogeneity in atmospheric conditions. The variability in predictor importance across sensors implies that spatial heterogeneity in wind speed plays a substantial role in modulating PM2.5 levels, especially in seasons with higher meteorological variability such as winter.

Figure 6 illustrates the mean importance of six meteorological predictors – Hour, Relative Humidity (2m), Wind Speed (10m), Surface Pressure, Temperature (2m), and Precipitation – across four seasons: winter, spring, summer, and autumn. The predictor “Hour” demonstrated the highest mean importance in winter, suggesting a strong diurnal influence on the target variable during the cold season. This is likely attributable to pronounced differences in solar radiation and energy balance throughout the day in winter months. However more importantly it may reflect the use of solid fuels for heating during the cold season. It was already proved by (Danek et al., 2022) that winter months' pollution is related mostly to that factor which is strongly hour-related. Relative Humidity (2m) emerged as the dominant predictor in spring and autumn, with its peak contribution observed in Spring.



a) b)

Figure 4 a) Seasonal importance of surface pressure predictors in PM2.5 concentration modeling. **4b)** Seasonal importance of temperature predictors in PM2.5 concentration modeling. Source: own elaboration

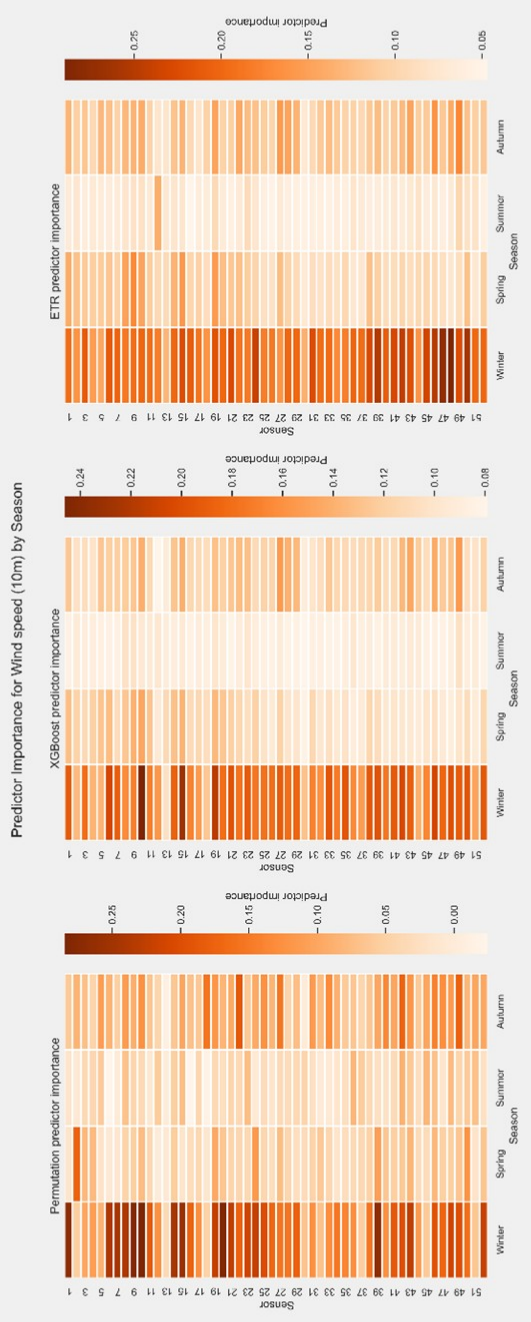


Figure 5 Seasonal importance of wind speed predictors in PM2.5 concentration modeling. Source: own elaboration

The transitional nature of these seasons may enhance the sensitivity of the target variable to variations in atmospheric moisture. In Summer, Wind Speed (10m) was identified as the most influential predictor, indicating a substantial role of wind dynamics during warmer periods. This may be associated with increased convective activity, mesoscale circulations, or turbulence-related processes common in summer weather patterns. Stronger or more variable winds can affect ventilation needs, the dispersion of pollutants, and the planning of events or labor in open-air environments. It is important to mention that in summer air pollution in the analyzed area is good and typically within the norm. Surface Pressure exhibited moderate importance across all seasons, with slightly elevated values in Summer. This consistent contribution suggests its stable role in governing synoptic-scale weather systems throughout the year. Temperature (2m) reached its highest relative importance in Summer and lowest impact in Autumn. Precipitation was the least influential predictor across all seasons, with minimal seasonal variation. Its low importance may reflect either a reduced direct effect on the modeled outcome or its collinearity with other variables such as humidity and pressure.

These findings highlight the seasonally dependent nature of predictor importance and underscore the need for temporally adaptive modeling strategies in meteorological prediction tasks.

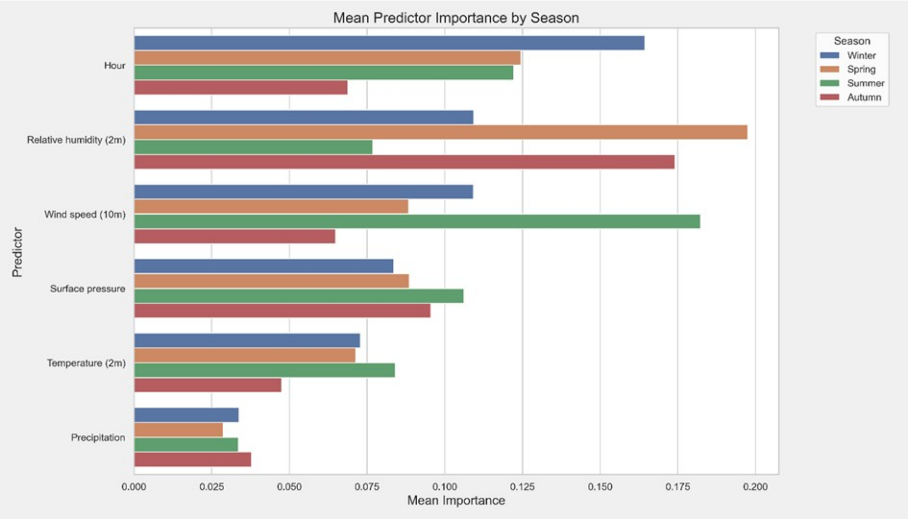


Figure 6 Seasonal Variation in Mean Predictor Importance for Meteorological Variables: Winter – blue; Spring – orange; Summer – green; Autumn – red.
Source: own elaboration

The spatial distributions of the average most important predictors for PM2.5 prediction in Kraków and its surroundings (Figure 7) reveal marked seasonal and topographic patterns, highlighting both the influence of meteorological variables and anthropogenic activity. In winter (Figure 7a), hour of the day emerges as the domi-

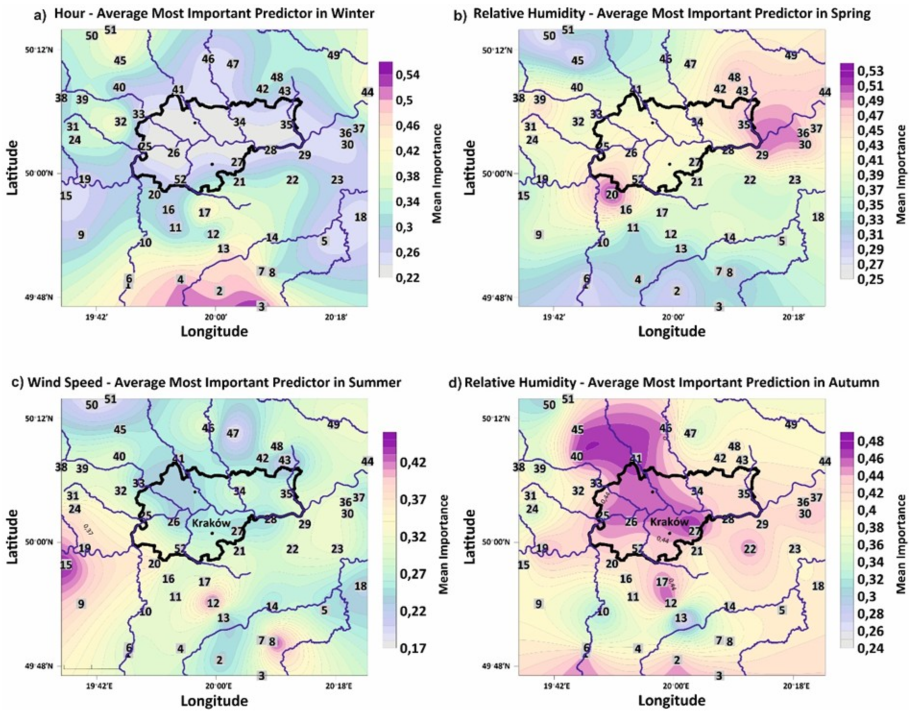


Figure 7 The spatial distributions of the average most important predictors for PM2.5 prediction in Kraków and its surroundings. Source: own elaboration

nant predictor, with the lowest importance observed in the urban core and highest values (>0.50) in peripheral southern and northern areas (e.g. sensors 2, 3, 26). This spatial gradient is likely linked to residential heating emissions in areas where no solid fuel bans are in effect, reinforcing the role of diurnal heating cycles. A sharp boundary in predictor importance coincides with the Vistula River, suggesting a potential topographic divide. This spatial structure reflects the widely described "obwarzanek (eng. bagel – ring shaped bread roll) effect" (concentric pollution ring), traditionally referring to particulate concentration but here also manifesting in predictor importance patterns, particularly for temperature-related variables In spring, relative humidity becomes the dominant predictor (Figure 7b), with elevated importance in the eastern and northeastern fringes (e.g. sensors 35 and 44), reaching values above 0.50, likely due to humidity-driven processes such as fog formation. Notably, the Niepołomice Forest region shows the highest humidity importance, supporting the hypothesis of localized meteorological phenomena influencing PM2.5 levels. During summer, wind speed is most influential, particularly in southwestern and western regions (e.g. sensors 9, 31, and 38), indicating the role of atmospheric ventilation and pollutant dispersion under convective conditions (Figure 7c). Significantly higher importance values on the western side of Kraków align with the

Kraków Gate lowland, where a natural air corridor along the Vistula Valley may facilitate the inflow of pollutants from the west. In contrast, central Kraków shows reduced importance, suggesting complex wind fields or urban shielding effects. In autumn, relative humidity once again dominates, but the spatial pattern shifts: highest importance is observed in the northern metropolitan area, especially around sensors 40 and 45, exceeding 0.48, potentially due to stagnant atmospheric conditions and the onset of the heating season (Figure 7d). A clear topographic dependency is visible, with humidity importance aligning along northern tributaries of the Vistula, in contrast to the latitudinal distribution observed in spring. For example, sensors 10 and 13, located on local elevations, exhibit notably lower humidity importance, emphasizing the modulating effect of terrain. Collectively, these results underscore the interplay of local emissions, meteorological dynamics, and topography in shaping the spatial relevance of predictors for PM2.5 modeling in Kraków.

4 DISCUSSION

The seasonal analysis of ML predictor importances was performed using three different methods: Permutation Importance, XGBoost, and Extremely Randomized Trees – ETR. This allowed for robust and accurate interpretation of importances as a key part of XAI concept. The seasonal variation in predictor importance for PM2.5 modeling reveals a complex interplay between meteorological conditions, anthropogenic activity, and spatial heterogeneity, emphasizing the need for nuanced, context-aware ML approaches for PM2.5 predictions in Kraków area. The hour of day stands out as the dominant predictor in winter, reflecting the strong influence of human-driven emission cycles, particularly from solid fuel combustion in unregulated suburban and rural zones surrounding Kraków. This temporal regularity, coupled with the prevalence of stable atmospheric conditions such as possible temperature inversions, significantly limits pollutant dispersion and leads to pronounced diurnal variation in PM2.5 concentrations. In spring and autumn, relative humidity becomes the most important factor, likely due to transitional atmospheric conditions that enhance the role of moisture in aerosol formation, fog events, and local-scale meteorological phenomena. In these seasons, humidity-driven processes appear especially relevant in vegetated and low-lying areas, where microclimatic effects are amplified. During summer, wind speed emerges as the key predictor, pointing to its critical role in pollutant transport and atmospheric ventilation. This is particularly evident in regions with favorable topographic conditions, such as the western lowlands and natural air corridors along the Vistula Valley, where stronger or more variable wind patterns facilitate pollutant dispersion. At the same time, temperature and surface pressure show stable but moderate importance across all seasons, indicating their foundational role in shaping broader meteorological structures, such as boundary layer development, convective activity, and pressure-driven airflow. Interestingly, precipitation exhibits the lowest overall importance, suggesting that its effect on PM2.5 concentrations is either limited in this regional context or largely captured through other, more predictive variables like humidity and pressure. It is important to mention that this phenomenon is strongly region-specific as other studies in different climate zones showed precipitation as key factor (Chantara et al., 2012). The spatial dimension of predictor relevance reveals clear topographic and urban-rural contrasts. Sensor-level differences in predictor importance frequently correspond to known emission hotspots or terrain-induced microclimates, as evidenced by the recurring influence of specific sensors in forested, elevated, or less urbanized regions. Phenomena like the "obwarzanek effect" extend beyond pollutant concentration patterns and are reflected in model sensitivity to specific predictors, particularly temperature-related ones. These spatial patterns emphasize the inadequacy of uniform modeling strategies and highlight the need for fine-grained, location-specific approaches. By embracing these variations, predictive models can better capture the real-world complexity of air pollution dynamics, enabling more accurate forecasting, more effective mitigation strategies, and ultimately, better-informed environmental policy tailored to both temporal rhythms and local geographic conditions.

The observed seasonal and spatial variability in PM2.5 predictor importance offers valuable insights for smart city applications, where adaptive, data-driven strategies are key to improving urban air quality. By identifying which variables – such as hour of day in winter or wind speed in summer – drive pollution under specific conditions, cities can implement dynamic policies and real-time response systems tailored to local and seasonal needs. For example, predictive models could inform traffic regulation, heating control, or targeted public health advisories based on hyperlocal forecasts. The spatial granularity of sensor relevance enables interventions in high-risk zones, while the explainability of models supports transparent and accountable decision-making. Integrating these insights into smart city infrastructure enhances proactive environmental management, making urban systems more responsive, sustainable, and resilient.

5 CONCLUSIONS

Understanding the complex factors influencing PM2.5 variability continues to be a significant challenge in air quality modeling, especially in Kraków during winter, where production drivers are closely linked to both socioeconomic factors – such as heating practices and income levels – and meteorological conditions. While previous studies have largely focused on pollutant concentration levels, less attention has been paid to the dynamic importance of meteorological and temporal predictors across seasons and space. This study addresses that gap by systematically evaluating the seasonal and spatial variability of predictor importance using XAI methods – XGBoost, Extra Trees Regressor (ETR), and Permutation Importance – applied to a dense sensor network in Kraków, Poland. By quantifying how predictors such as hour of day, humidity, wind speed, temperature, surface pressure, and precipitation contribute to PM2.5 predictions throughout the year and across locations, we provide a deep understanding of air pollution drivers. Our approach enhances interpretability, supports targeted policy-making, and enables the development of seasonally adaptive and spatially informed models, directly supporting smarter environmental management strategies in urban contexts. Hour of day is the most influential predictor in winter, reflecting strong human-driven emission cycles. Relative humidity dominates in spring and autumn – in spring the main role plays Vistula river while in autumn its side-streams. Wind speed is most important in summer where no significant PM emission is observed. Temperature and surface pressure maintain moderate, stable influence across all seasons. Precipitation shows consistently low importance, suggesting limited direct impact. Predictor importance varies spatially, shaped by topography, land use, and regulatory zones. Explainable models (XGBoost, ETR, Permutation) reveal consistent and interpretable patterns crucial for informed urban decision-making. XAI offers a reliable complementary approach to classical methods, enhancing our ability to interpret complex environmental data and improve air quality management strategies. Addressing this challenge requires the combined efforts of experts across disciplines – including GIS specialists, statisticians, meteorologists, geographers, geo-data scientists, and policy makers – supported by high-quality sensor networks.

Acknowledgment

This research project was partly supported by the AGH University of Krakow, Faculty of Geology, Geophysics and Environmental Protection, as a part of a statutory project. Research project partly supported by program „Excellence initiative – research university" for the AGH University.

Availability of data and materials

Publicly available datasets from Airly sensors were analyzed in this study and can be found here: (https://map.airly.org/, accessed on 29 April 2025). API documentation from Airly is available here: (https://developer.airly.org/en/docs, accessed on 29 April 2025).

References

BOKWA, A. 2008. Environmental impacts of long-term air pollution changes in Krakow, Poland. *Polish Journal of Environmental Studies*, 17, 673-686.

BREIMAN, L. 2001. Random forests. *Machine Learning*, 45, 1, 5-32. DOI: https://doi.org/10.1023/A: 1010933404324

COHEN, A. J., BRAUER, M., BURNETT, R., ANDERSON, H. R., FROSTAD, J., ESTEP, K., BALAKRISHNAN, K., BRUNEKREEF, B., DANDONA, L., DANDONA, R. 2017. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the Global Burden of Diseases Study'. *Lancet*, 389(10082), 1907-1918. DOI: https://doi.org/10.1016/S0140-6736(17)30505-6

DANEK, T., WĘGLIŃSKA, E., ZAREBA, M. 2022. The influence of meteorological factors and terrain on air pollution concentration and migration: A geostatistical case study from Krakow, Poland. *Scientific Reports*, 12, 11050. DOI: https://doi.org/10.1038/s41598-022-15160-3

EUROPEAN COMMISSION. 2025. *Smart cities and communities*. [online] [cit. 2025-03-16]. Available at: <https://digital-strategy.ec.europa.eu/pl/policies/smart-cities-and-communities>

GODŁOWSKA, J., KASZOWSKI, K., KASZOWSKI, W. 2022. Application of the FAPPS system based on the CALPUFF model in short-term air pollution forecasting in Krakow and Lesser Poland. *Archives of Environmental Protection*, 48, 3, 109-117. DOI: https://doi.org/10.24425/aep.2022.142695

GUMIŃSKI, R. 1950. Wazniejsze elementy klimatu rolniczego Polski południowo-wschodniej (Important aspects of agricultural climate in south-east Poland). *Wiadomości Służby Hydrologicznej i Meteorologicznej*, 3(1), 57-113. [in Polish]

CHANTARA, S., SILLAPAPIROMSUK, S., WIRIYA, W. 2012. Atmospheric pollutants in Chiang Mai (Thailand) over a five-year period (2005–2009), their possible sources and relation to air mass movement. *Atmospheric Environment*, 60, 88-98. DOI: https://doi.org/10.1016/j.atmosenv.2012.06.044

CHEN, T., GUESTRIN, C. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794), ACM. DOI: https://doi.org/10.1145/2939672.2939785

JONEK-KOWALSKA, I. 2023 Assessing the effectiveness of air quality improvements in Polish cities aspiring to be sustainably smart. *Smart Cities*, 6, 1, 510-30. DOI: https://doi.org/10.3390/smartcities6010024

KOPCZEWSKA, K. 2022. Spatial machine learning: New opportunities for regional science. *The Annals of Regional Science*, 68(4), 713-755. DOI: https://doi.org/10.1007/s00168-021-01101-x

LIPTON, Z. C. 2016. *The mythos of model interpretability*. arXiv. [online] [cit. 2025-04-23]. Available at: <https://arxiv.org/abs/1606.03490>

MAŁOPOLSKA REGION. n.d. *Publications on air quality*. [online] [cit. 2025-03-16]. Available at: <https://powietrze.malopolska.pl/tag/publikacje/>

MANISALIDIS, I., STAVROPOULOU, E., STAVROPOULOS, A., BEZIRTZOGLOU, E. 2020. Environmental and health impacts of air pollution: A review. *Frontiers in Public Health*, 8(14), DOI: https://doi.org/10.3389/fpubh.2020.00014

MARCH, H., RIBERA-FUMAZ, R. 2014. Smart contradictions: The politics of making Barcelona a self-sufficient city. *European Urban and Regional Studies*, 23, 4, 816-830. DOI: https://doi.org/10.1177/0969776414554488

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825-2830.

RAHMANI, A. M., YOUSEFPOOR, E., YOUSEFPOOR, M. S., MEHMOOD, Z., HAIDER, A., HOSSEINZADEH, M., ALI NAQVI, R. 2021. Machine Learning (ML) in Medicine: Review, Applications, and Challenges. *Mathematics*, 9(22), 2970. DOI: https://doi.org/10.3390/math9222970

SAMEK, W., WIEGAND, T., MÜLLER, K.-R. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv. [online] [cit. 2025-03-16]. Available at: <https://arxiv.org/abs/1708.08296>

SARKER, I. H. 2021. Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, 2: 420, 1-20. DOI: https://doi.org/10.1007/s42979-021-00815-1

SCIKIT-LEARN DEVELOPERS. 2021. *ExtraTreesRegressor in Scikit-learn*. [online] [cit. 2025-03-16]. Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html>

STEIN, A. F., DRAXLER, R. R., ROLPH, G. D., STUNDER, B. J., COHEN, M. D., NGAN, F. 2015. NOAA's HYSPLIT atmospheric transport and dispersion modeling system. *Bul-*

letin of the American Meteorological Society, 96, 12, 2059-2077. DOI: https://doi.org/10.1175/BAMS-D-14-00110.1

THURSTON, G. D., KIPEN, H., ANNESI-MAESANO, I., BALMES, J., BROOK, R. D., CROMAR, K., DE MATTEIS, S., FORASTIERE, F., FORSBERG, B., FRAMPTON, M. W. 2017. A joint ERA/ATS policy statement: What constitutes an adverse health effect of air pollution? An analytical framework'. *European Respiratory Journal*, 49, 1600419. DOI: https://doi.org/10.1183/13993003.00419-2016

WEINMAYR, G., ROMEO, E., DE SARIO, M., WEILAND, S. K., FORASTIERE, F. 2010. Short-term effects of PM_{10} and NO_2 on respiratory health among children with asthma or asthma-like symptoms: A systematic review and meta-analysis. *Environmental Health Perspectives*, 118, 4, 449-457. DOI: https://doi.org/10.1289/ehp.0900844

WOJEWÓDZKI INSPEKTORAT OCHRONY ŚRODOWISKA W KRAKOWIE. 2020. *Air Quality in Krakow*. Summary of Research Results. [online] [cit. 2025-02-21]. Available at: <https://krakow.wios.gov.pl/2020/09/jakosc-powietrza-w-krakowie-podsumowanie-wynikow-badan/>

ZAREBA, M., DANEK, T. 2022. Analysis of Air Pollution Migration during COVID-19 Lockdown in Krakow, Poland. *Aerosol and Air Quality Research*, 22, 3, 210275. DOI: https://doi.org/10.4209/aaqr.210275

ZAREBA, M., DANEK, T., STEFANIUK, M. 2023. Unsupervised Machine Learning Techniques for Improving Reservoir Interpretation Using Walkaway VSP and Sonic Log Data. *Energies*, 16, 1, 493. DOI: https://doi.org/10.3390/en16010493

ZAREBA, M., WĘGLIŃSKA, E., DANEK, T. 2024. Air pollution seasons in urban moderate climate areas through big data analytics. *Scientific Reports*, 14, 1, 3058. DOI: https://doi.org/10.1038/s41598-024-52733-w

Umelá inteligencia a priestorová analýza pre udržateľné mestské plánovanie v oblasti riešenia znečistenia ovzdušia

Súhrn

Pozorovaná sezónna a priestorová variabilita vo význame prediktorov PM2,5 ponúka cenné poznatky pre aplikácie inteligentných miest, kde sú adaptívne stratégie založené na údajoch kľúčové pre zlepšenie kvality mestského ovzdušia. Identifikáciou toho, ktoré premenné – ako napríklad hodina dňa v zime alebo rýchlosť vetra v lete – spôsobujú znečistenie za špecifických podmienok, môžu mestá implementovať dynamické politiky a systémy reakcie v reálnom čase prispôsobené lokálnym a sezónnym potrebám. Napríklad prediktívne modely by mohli informovať o regulácii dopravy, regulácii vykurovania alebo cielených odporúčaniach v oblasti verejného zdravia založených na hyperlokálnych predpovediach. Priestorová granularita relevantnosti senzorov umožňuje intervencie vo vysoko rizikových zónach, zatiaľ čo vysvetliteľnosť modelov podporuje transparentné a zodpovedné rozhodovanie. Integrácia týchto poznatkov do infraštruktúry inteligentných miest zlepšuje proaktívne environmentálne riadenie, vďaka čomu sú mestské systémy responzívnejšie, udržateľnejšie a odolnejšie.

Pochopenie komplexných faktorov ovplyvňujúcich variabilitu PM2,5 je naďalej významnou výzvou v modelovaní kvality ovzdušia, najmä v Krakove počas zimy, kde sú faktory ovplyvňujúce produkciu úzko prepojené so socioekonomickými faktormi – ako sú vykurovacie postupy a úroveň príjmov – a meteorologickými podmienkami. Zatiaľ čo predchádzajúce štúdie sa prevažne zameriavali na úrovne koncentrácie znečisťujúcich látok, menšia pozornosť sa venovala dynamickému významu meteorologických a časových prediktorov v rôznych ročných obdobiach a priestore. Táto štúdia rieši túto medzeru systematickým hodnotením sezónnej a priestorovej

variability dôležitosti prediktorov pomocou metód XAI – XGBoost, Extra Trees Regressor (ETR) a Permutation Importance – aplikovaných na hustú sieť senzorov v Krakove v Poľsku. Kvantifikáciou toho, ako prediktory, ako je denná hodina, vlhkosť, rýchlosť vetra, teplota, povrchový tlak a zrážky, prispievajú k predpovediam PM2,5 počas celého roka a na rôznych miestach, poskytuje hĺbkové pochopenie faktorov ovplyvňujúcich znečistenie ovzdušia. Tento prístup zlepšuje interpretovateľnosť, podporuje cielenú tvorbu politík a umožňuje vývoj sezónne adaptívnych a priestorovo informovaných modelov, ktoré priamo podporujú inteligentnejšie stratégie environmentálneho manažmentu v mestských územiach. Denná hodina je najvplyvnejším prediktorom v zime, ktorý odráža silné cykly emisií riadené človekom. Relatívna vlhkosť dominuje na jar a na jeseň – na jar hrá hlavnú úlohu rieka Visla, zatiaľ čo na jeseň jej vedľajšie toky. Rýchlosť vetra je najdôležitejšia v lete, keď sa nepozorujú žiadne významné emisie PM. Teplota a povrchový tlak si udržiavajú mierny, stabilný vplyv počas všetkých ročných období. Zrážky vykazujú konzistentne nízky význam, čo naznačuje obmedzený priamy vplyv. Význam prediktorov sa priestorovo líši a je formovaný topografiou, využívaním pôdy a regulačnými zónami. Vysvetliteľné modely (XGBoost, ETR, Permutation) odhaľujú konzistentné a interpretovateľné vzorce, ktoré sú kľúčové pre informované rozhodovanie v mestách. XAI ponúka spoľahlivý doplnkový prístup ku klasickým metódam, čím zvyšuje našu schopnosť interpretovať komplexné environmentálne údaje a zlepšovať stratégie riadenia kvality ovzdušia. Riešenie tejto výzvy si vyžaduje spoločné úsilie odborníkov z rôznych disciplín – vrátane špecialistov na GIS, štatistikov, meteorológov, geografov, geodátových vedcov a tvorcov politík – s podporou vysokokvalitných senzorových sietí.