

Evaluating machine learning models for air quality error mapping in Kraków, Poland

Abstract

Accurate air quality prediction is essential for sustainable urban development. This study evaluates the performance of machine learning models, including DLinear and XGBoost, in comparison with the traditional Autoregressive Integrated Moving Average (ARIMA) method for predicting fine particulate matter ($PM_{2.5}$) concentrations in Kraków, Poland. A dense network of low-cost sensors was used to generate high-resolution spatial and temporal data. Prediction errors were analysed using the Getis-Ord G_i^* spatial statistics method during both extreme pollution events and low pollution days. The results indicate that DLinear achieved the lowest root mean square error (RMSE = $3.8 \mu g/m^3$), followed by XGBoost (RMSE = $6.7 \mu g/m^3$) and ARIMA (RMSE = $9.2 \mu g/m^3$). The spatial distribution of errors highlights the influence of environmental factors, such as humidity and proximity to water bodies, on model accuracy. These findings show the limitations of current prediction models and emphasize the need for spatially adaptive approaches to improve pollution.

Keywords

Machine learning • artificial intelligence • air pollution • prediction errors • spatial analysis

Mateusz Zareba[✉], Szymon Cogieł[✉],
Elżbieta Węglińska[✉],
Tomasz Danek[✉]

Department of Geoinformatics and Applied Computer Science, Faculty of Geology, Geophysics and Environmental Protection, AGH University of Krakow, 30-059 Krakow, Poland
e-mail: zareba@agh.edu.pl; mattzareba@gmail.com
e-mail: cogielszymon1@gmail.com
e-mail: weglinaska@agh.edu.pl
e-mail: tdanek@agh.edu.pl

Received: 4 November 2024
Accepted: 10 April 2025

Introduction

The smart city concept is one of the urban development approaches based on using technology to improve residents' quality of life (Firszt & Jabłoński 2023). One of the key elements of the smart city concept is clean air (Liang & Gong 2020). Kobus et al. (2018) emphasized the importance of delivering air quality information to city residents to raise awareness about this critical issue. Jonek-Kowalska (2023) pointed out the lack of long-term studies on air pollution in cities aspiring to become smart cities. Zareba et al. (2024a) performed an in-depth analysis of air pollution concentration to evaluate the effects of energy transition and environmental policies, aiming to understand the long-term environmental consequences of these actions. Kraków (Figure 1) has progressed in terms of air pollution but extreme smog episodes (ESE) persist (Danek & Zareba 2021; Zareba & Danek 2021). There are no strict PM limit values defining smog but, in Poland, the limit values for PM_{10} are set at $100 \mu g/m^3$ as the information threshold and $150 \mu g/m^3$ as the alarm level, while, for $PM_{2.5}$, Poland has no official 24-hour limit (Rada Ministrów 2019); however, WHO recommends a 24-hour guideline of $15 \mu g/m^3$ for $PM_{2.5}$ and $45 \mu g/m^3$ for PM_{10} (WHO 2021). Despite a solid fuel ban, smog remains a problem in Kraków due to its geography (Morawska-Horawska & Lewik 2003). The city's location along the Wisła River valley, surrounded by hills, causes pollution to flow into Kraków from neighbouring towns where the burning ban is not in effect (Danek et al., 2022). While reference stations provide highly accurate air quality data, they are expensive and limited in coverage. In contrast, low-cost sensors (LCS) are a reliable and affordable choice for smart cities to monitor air pollution on a larger scale, even though they are sensitive to humidity (Di Antonio

et al. 2018). As part of the Internet of Things (IoT), these sensors provide real-time air quality data, helping cities to track harmful emissions and take action to improve public health. Smart cities use web-based and remote monitoring to better understand pollution levels in public and green spaces. By identifying major sources of pollution and informing the public, cities can encourage eco-friendly practices, promote sustainable policies, and reduce overall pollution levels for a cleaner and healthier environment (Ullah et al. 2023). Despite the high availability of data from LCS, reliable pollution predictions are challenging and depend on the model choice, especially for ESE. Previous studies successfully demonstrated a distinct seasonality in air pollution levels in Kraków (Zareba et al. 2024b) and introduced machine learning (ML) techniques that proved highly effective for spatiotemporal analysis of PM_x distributions (Zareba et al. 2023).

This paper's novelty lies in a detailed examination of the spatial distribution of prediction errors across various algorithms. The study presents the results of global models, using XGBoost (Chen & Guestrin 2016) and DLinear (Zeng et al. 2022), and evaluating their effectiveness compared to the classic ARIMA (Box & Jenkins, 1970) method. In previous research by Zareba et al. (2024a), these models were used to calculate 24-hour forecasts of $PM_{2.5}$ concentrations at the sensor locations, and the errors between the real hourly data and the model predictions were calculated. The 24-hour forecast is generated by predicting each hour separately, one step at a time, for the next 24 hours using the trained model. This study goes further by explaining how these errors vary spatially. A hot-spot analysis was conducted for prediction errors from each method, comparing an ESE with a low pollution day, using the local Getis-Ord G_i^* spatial autocorrelation method. We consider

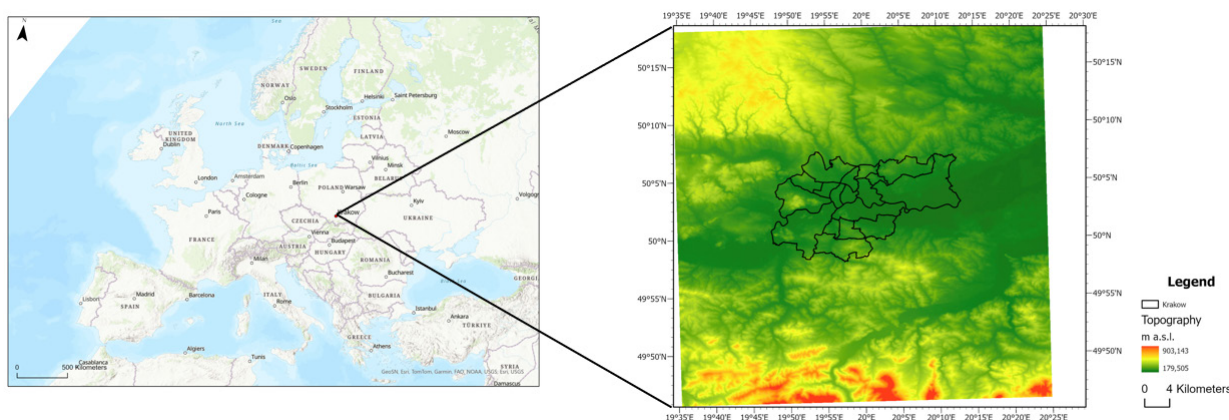


Figure 1 The study location is marked with a red dot

Source: Esri, DeLorme, HERE, TomTom, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, and the GIS User Community

a smog day as one in which a $PM_{2.5}$ episode exceeding $50 \mu g/m^3$ occurs – a threshold commonly used for PM_{10} . A day on which this limit value is not exceeded is classified as a low pollution day. An interpretation of other factors, such as location and proximity to water bodies, was also carried out to assess their potential influence on prediction errors. In this study, prediction error refers to the discrepancy between the observed $PM_{2.5}$ concentrations from the LCS and the values predicted by the models. Depending on the context, this is quantified using metrics such as RMSE (Root Mean Squared Error) or MAE (Mean Absolute Error). Rather than solely evaluating overall prediction accuracy, this study focuses on understanding the factors influencing model performance in different geographic conditions. It should be noted that models are compared with LCS data, so result from inaccuracies in the measurement values, not solely from incorrect model predictions. By leveraging explainable AI (XAI) techniques, we analyse why specific models perform better in certain areas, and under what conditions errors increase, aiming to uncover whether a single predictive model can generalize across Kraków or if a hybrid approach is necessary. Our findings indicate that different models exhibit varying sensitivity to local environmental factors, such as humidity and proximity to water bodies, which significantly impact prediction reliability. This suggests that selecting the most suitable predictive approach requires not just statistical performance evaluation but also an understanding of spatial dependencies, ultimately contributing to more interpretable and effective air pollution forecasting.

Materials and Methods

The impact of the Low-Emission Reduction Programme (PONE) in Kraków was analysed for the period from 2014 to 2019 (Kraków 2019). Spatiotemporal data from 52 Airly LCS (Airly 2024) were used. LCS may have lower accuracy than reference stations but their coverage allows for a quasi-uniform spatial distribution across the city. In previous research on the sensors used in this study, the median difference from reference stations was approximately $3.0 \mu g/m^3$, with pollution levels reaching around $80.0 \mu g/m^3$ on some days (Danek & Zareba 2021). While LCS have inherent uncertainties compared to reference stations, the Airly LCS in this study underwent manufacturer calibration using AI/ML and weather corrections, achieving performance validated within approximately 10% of reference measurements. The data accessed via API were already adjusted through this manufacturer calibration process. These inherent uncertainties do not affect the analysis of ML prediction errors, as the same

calibrated data were used both for training and evaluation. In studies comparing two sensors side by side, the Airly LCS uncertainty was $2.4 \mu g/m^3$ for $PM_{2.5}$ (Zięba & Dworakowska 2018). In other studies, Airly LCS without prior calibration demonstrated strong correlations with reference instruments for $PM_{2.5}$ ($0.77 < R^2 < 0.90$) (South Coast Air Quality Management District n.d.). The impact of the Low-Emission Reduction Programme (PONE) in Kraków was analysed for the 2014–2019 period (Kraków 2019). An automated pipeline was developed to clean the data, fill in missing values using linear interpolation, perform preprocessing, and conduct feature engineering. Missing value imputation can also be done using machine learning methods (Armut et al. 2024). During the feature engineering phase, cyclical features (e.g. day, hour, wind, etc.), sunset times, and seasonal trends were created for different lags. The training process used 168 unique lagged features, derived from 28 input data features. The models were trained with global and local approaches for the $PM_{2.5}$ forecast: XGBoost and DLinear from Darts library (Herzen et al. 2021) trained on all LCS (global), while ARIMA trained on each one separately. The training period spanned from 1 January 2022 to 20 October 2022 at 01:00:00. Backtesting, a method for evaluating predictive models on historical data, was then performed using a rolling window with a stride of 1 from 20 October 2022 at 01:00:00 to 1 January 2023. The backtesting concept is described in Herzen et al. (2021). The model trained on 80% of the data and made 24-hour forecasts at each step, updating iteratively with new data. The testing dataset consisted of the remaining 20% of the data. In this study, specific dates were selected based on concentration analysis from Zareba et al. (2024a), and the model training followed the methodology outlined in that study. The XGBoost model is a scalable and efficient gradient-boosting framework that builds additive regression trees. Additionally, XGBoost is highly effective for time series forecasting because it can include additional covariates that influence the target feature. Its ability to extract and model complex, general features from data ensures robust and accurate predictions, making it well suited for spatial forecasting tasks.

XGBoost aims to minimize the following objective function, which includes a regularization term to prevent overfitting, as shown in Eq. (1):

$$Obj(\theta) = \sum_{t=1}^n l(y_t, \hat{y}_t) + \sum_{k=1}^K \Omega(f_k) \quad (1)$$



Figure 2. Energy transformation and air pollution in Kraków: a) Yearly average PM2.5 level in Bujaka reference station; b) Coal furnaces and boilers removed as an outcome of PONE programme; c) Number of new renewable energy source installations
Source: own study

where l is a differentiable convex loss function (e.g. mean squared error), y_t represents the measured $\text{PM}_{2.5}$ concentration at the t -th hour for a given sensor, $\hat{y}_t$ represents the predicted $\text{PM}_{2.5}$ concentration at the t -th hour, $\Omega(f_k)$ is the regularization term for the k -th tree f_k , and Θ represents the parameters of all trees. The XGBoost model was trained using the following hyperparameters: number of boosting rounds = 500 (controlling model complexity), maximum tree depth = 6 (capturing complex patterns), learning rate = 0.3 (adjusting weight updates per round), L2 regularization = 1 (mitigating overfitting), and output chunk length = 24 (predicting 24 future time steps at once). The lags used were [-48, -24, -2, -1] for the main variables, [-48, -24, -12, -8, -2, -1] for past covariates, and [-8, -4, -3, -2, -1, 0] for future covariates. Following model training and backtesting, the XGBoost model achieved an MAE of 4.104 and an RMSE of 6.763, indicating a satisfactory level of predictive accuracy for the given forecasting task.

The ARIMA method (Box & Jenkins 1970) is an Autoregressive Integrated Moving Average model, used for analysing and forecasting time series data. It is defined by the equation Eq. (2):

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q} + \epsilon_t \quad (2)$$

where y_t is the differenced $\text{PM}_{2.5}$ time series at time t , c is a constant term, $\phi_1, \dots, \phi_p$ are the autoregressive (AR) coefficients, $\theta_1, \dots, \theta_q$ are the moving average (MA) coefficients, and ϵ_t is the error term at time t . For comparison, the ARIMA model underperformed with an MAE of 6.688 and an RMSE of 9.282, confirming that XGBoost captured the underlying patterns more effectively.

The DLinear model (Zeng et al. 2022) is a deep learning architecture designed for long-term time series forecasting. It focuses on modelling the linear dynamics by decomposing the time series and employing linear transformations. DLinear achieved the best results, with an MAE of 2.947 and an RMSE of 3.888. SHAP (SHapley Additive exPlanations) values were computed following the methodology of Lundberg & Lee (2017) to quantify the contribution of each meteorological factor according to the model's predictions.

The study focused on global models (linear neural networks or ensemble trees) and the local ARIMA method to assess predictive potential using data from either all sensors or individual ones. The hot-spot analysis was conducted using the Getis-Ord G_i^* (High/Low Clustering) tool, implemented in ArcGIS Pro (ESRI 2021). This tool identifies statistically significant clusters of high and low values. The Getis-Ord local statistic G_i^* (Getis & Ord 1992) is calculated according to the formula Eq. (3):

$$G_i^* = \frac{\sum_{j=1}^n w_{ij} x_j - \bar{X} \sum_{j=1}^n w_{ij}}{S \sqrt{\frac{n \sum_{j=1}^n w_{ij}^2 - (\sum_{j=1}^n w_{ij})^2}{n-1}}} \quad (3)$$

where x_j is the attribute value of x at location j , n is the number of location points, w_{ij} is the spatial weight between locations i and j , $\bar{X}$ is the mean, and S is the standard deviation of the attribute.

Results

The research uncovers spatiotemporal dynamics in air pollution distribution. Long-term observations in Kraków show that ESE remain significant despite an overall decline in pollution.

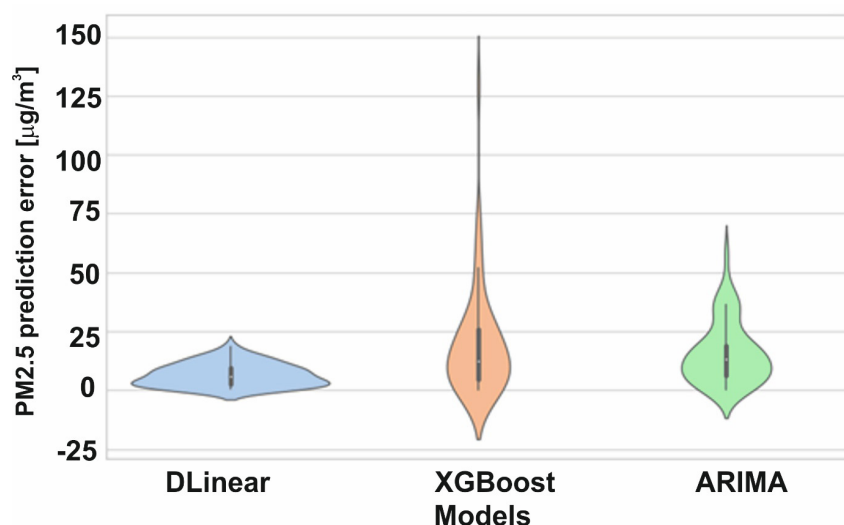


Figure 3. PM_{2.5} prediction error distribution on smog day (30 December at 23:00) for DLinear, XGBoost, and ARIMA models
Source: own study

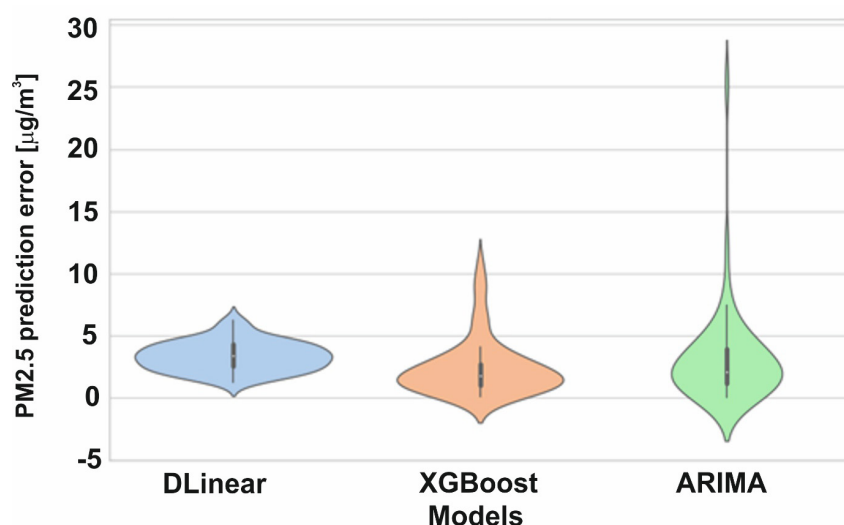


Figure 4. PM_{2.5} prediction error distribution on steady low pollution day (December 12th 10:00) for DLinear, XGBoost, and ARIMA models
Source: own study

Figure 2 presents a clear decrease in yearly PM_{2.5} averages, with the largest drop between 2015 and 2016. Similar trends have been seen in neighbouring Slovakia, linked to winter heating practices (Sediva & Stefanik 2024). A linear relationship exists between the amount spent on the PONE (Kraków 2019) programme and the removal of coal furnaces and boilers. A similar behaviour can be observed in renewable energy sources.

Violin plots show prediction errors for ESE on 30 December at 23:00 (Figure 3) and steady low pollution on 12 December at 10:00 (Figure 4). The DLinear model accurately captured rare values on both analysed days, with the highest error below 25 µg/m³. For the ESE, the highest prediction errors occurred with the XGBoost model (orange), reaching nearly 150 µg/m³. On the no-smog day, the highest error was from the ARIMA model (green), at 30 µg/m³, while the highest error for DLinear on this day was less than 7 µg/m³. Errors for the DLinear models (blue) on both days were around the LCS accuracy level of 10 µg/m³. For the

ARIMA model during ESE days, two groups can be observed – one with the majority of errors around 12.5 µg/m³ and the other with around 37.5 µg/m³.

Figure 5 presents the distribution of errors for the DLinear, XGBoost, and Arima models – for both ESE and steady low pollution days. For the DLinear model during ESE, the errors across the entire area are relatively low and distributed along a NW–SE axis, with a tendency towards a latitudinal orientation in the northern part. The largest errors are associated with LCS 10 (Krzywaczka), 41 (Trojanowice), 43 (Karniów), and the 30–36–37 group (in the vicinity of the open swimming and fishing area in Zabierzow Bochenski). The lowest errors are observed for LCS 8 (Kwapinka), 17 (Rzeszotary), 25 (Zwierzyniec-Kraków city), and 32 (Aleksandrowice). These XGBoost errors during ESE, are significantly larger (7–8 times) than those using the DLinear model. Three distinct areas of high errors are evident, distributed across various sections of the studied terrain. The

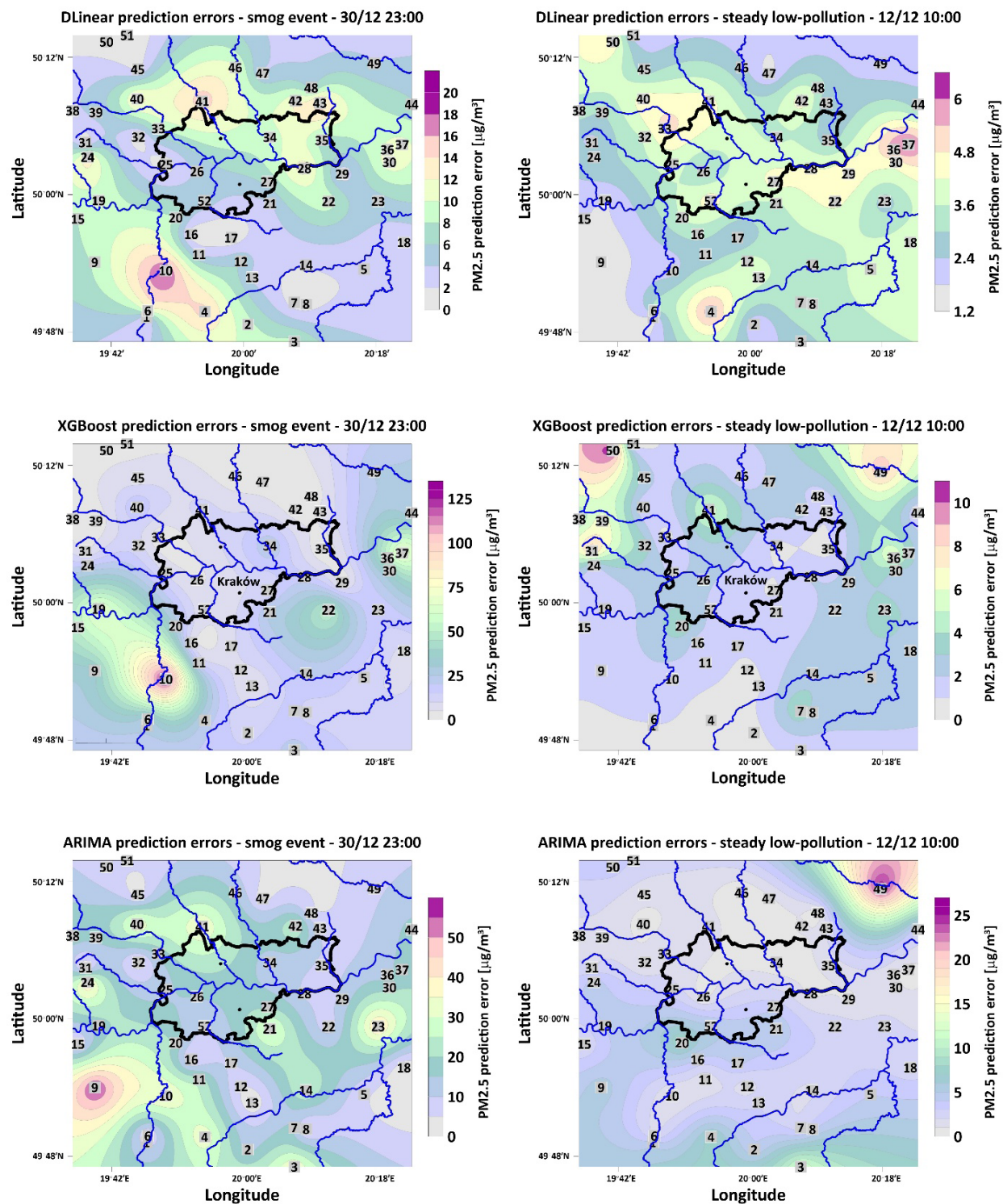


Figure 5. Map of DLinear model prediction errors for the smog event day on 30 December at 23:00 and the steady low pollution day of 12 December at 10:00 for DLinear, XGBoost, and ARIMA models with LCS locations (grey rectangle with number)
Source: own study

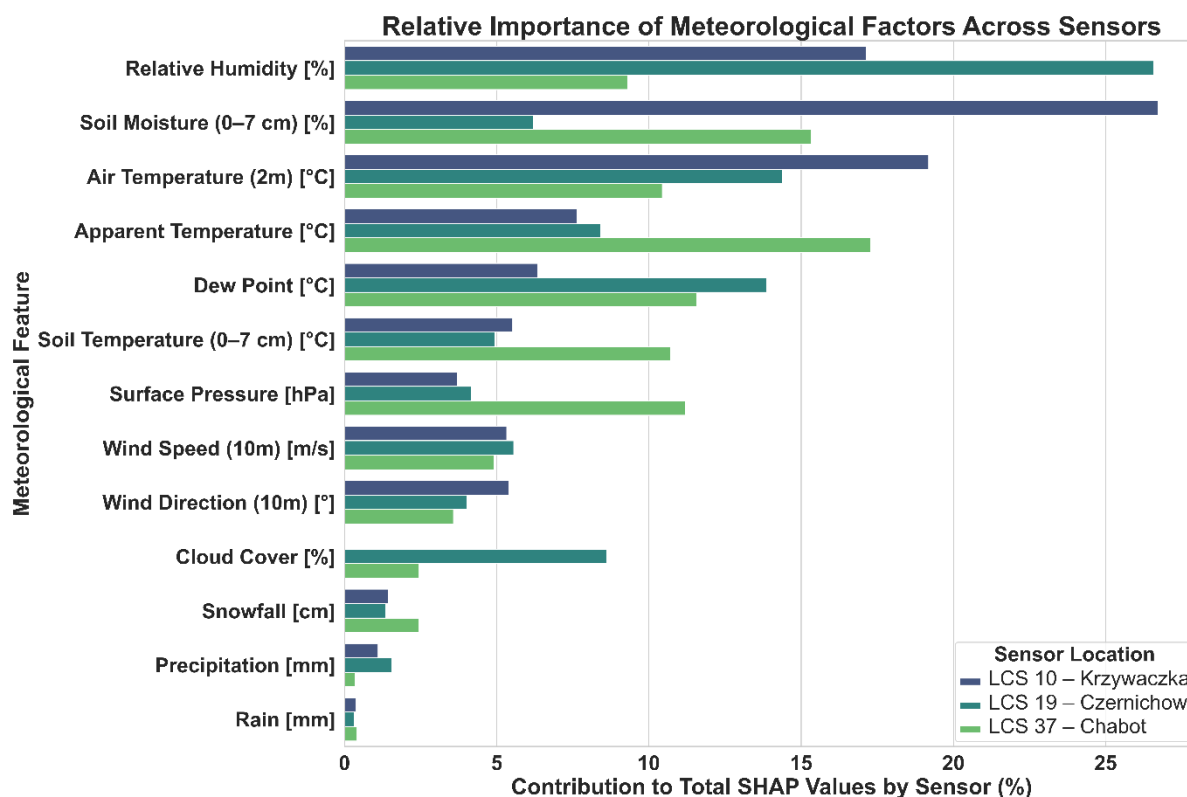


Figure 6. The relative importance of meteorological factors across sensors with the highest errors – LCS 10; LCS 19, and LCS 37
Source: own study

largest error is for LCS 10 (Krzywaczka), located in a locally isolated valley near the Skawinka River. This is followed by LCS 22 (Zagórze), situated in a local depression at the confluence of small streams with the Podlezyńska River and a small pond, and LCS 37 (Chabot) in the open fishing and swimming area and floodplain of the Wisła River. The ARIMA error distribution during ESE, may be described as the superposition of DLinear and XGBoost surface locations of the highest errors. Error values lie between those obtained with the above-mentioned ML models. The errors are significantly lower across all methods during steady low pollution days. While a latitudinal distribution characterizes the DLinear model, the XGBoost model exhibits a more longitudinal orientation, whereas ARIMA appears to fall somewhere in between. The area around the floodplains near Zaborze Bocheńskie is consistently characterized by relatively high errors, which is not evident in the case of ARIMA. For both XGBoost and ARIMA, the largest error is observed in LCS 49, located in Proszowice.

Figure 6 presents the relative importance of meteorological factors across LCS 10, 19, and 37, as computed from contributions to the total SHAP values attributed to each sensor. The influence of various phases of water is observable. Relative humidity and soil moisture are identified as the most significant factors, with air temperature typically occupying the third position in terms of importance. This finding is particularly noteworthy, as existing research indicates that the presence of water vapour, such as fog, can act as a precursor to pollution accumulation (Tsyró 2005). Furthermore, it is important to consider that optical measurements derived from LCS may be affected by the presence of H_2O , thereby potentially leading to increased measurement errors under these conditions (Peltier et al. 2020).

Figure 7 shows hot-spot and cold-spot maps with their confidence level for $PM_{2.5}$ error prediction using Getis-Ord G_i^* for ESE and low pollution days. The results were presented for each model using: DLinear, XGBoost, and ARIMA. This method only allowed hot spots to be identified. The largest prediction errors for the DLinear method were observed at LCS 10 (95% confidence level), and LCS 4, 41, and 43 (90% confidence level) during smog events. For the low pollution day, significant errors were noted at LCS 4, 36, and 37 (90% confidence level). Exceptionally high prediction errors for the XGBoost model were observed during the ESE at LCS 10 (99% confidence level), and at LCS 19 and 37 (95% confidence level). On the low pollution day, error hot spots were identified at LCS 31, 49, and 50 (99% confidence level) and LCS 39 and 39 (90% confidence level). For the ARIMA model, error hot spots were identified with 99% confidence at LCS 9, with 95% confidence at LCS 23, and with 90% confidence at LCS 24. On the low pollution day, a hot spot was identified with 99% confidence at LCS 49 and with 90% confidence at LCS 3.

Discussion

Short- and long-term $PM_{2.5}$ data, combined with PONE data, enabled a spatiotemporal assessment of ML methods in air quality management.

For long-term analyses, a downward trend in average $PM_{2.5}$ concentrations is evident. Unfortunately, despite generally good annual average concentrations, Kraków continues to experience ESE due to the inflow of pollutants into the city. In general, financial expenditure for the PONE programme has been significant but has been declining since 2017. It can be observed that the reduction in funding directly impacts the number of replaced furnaces and the installation of new renewable energy

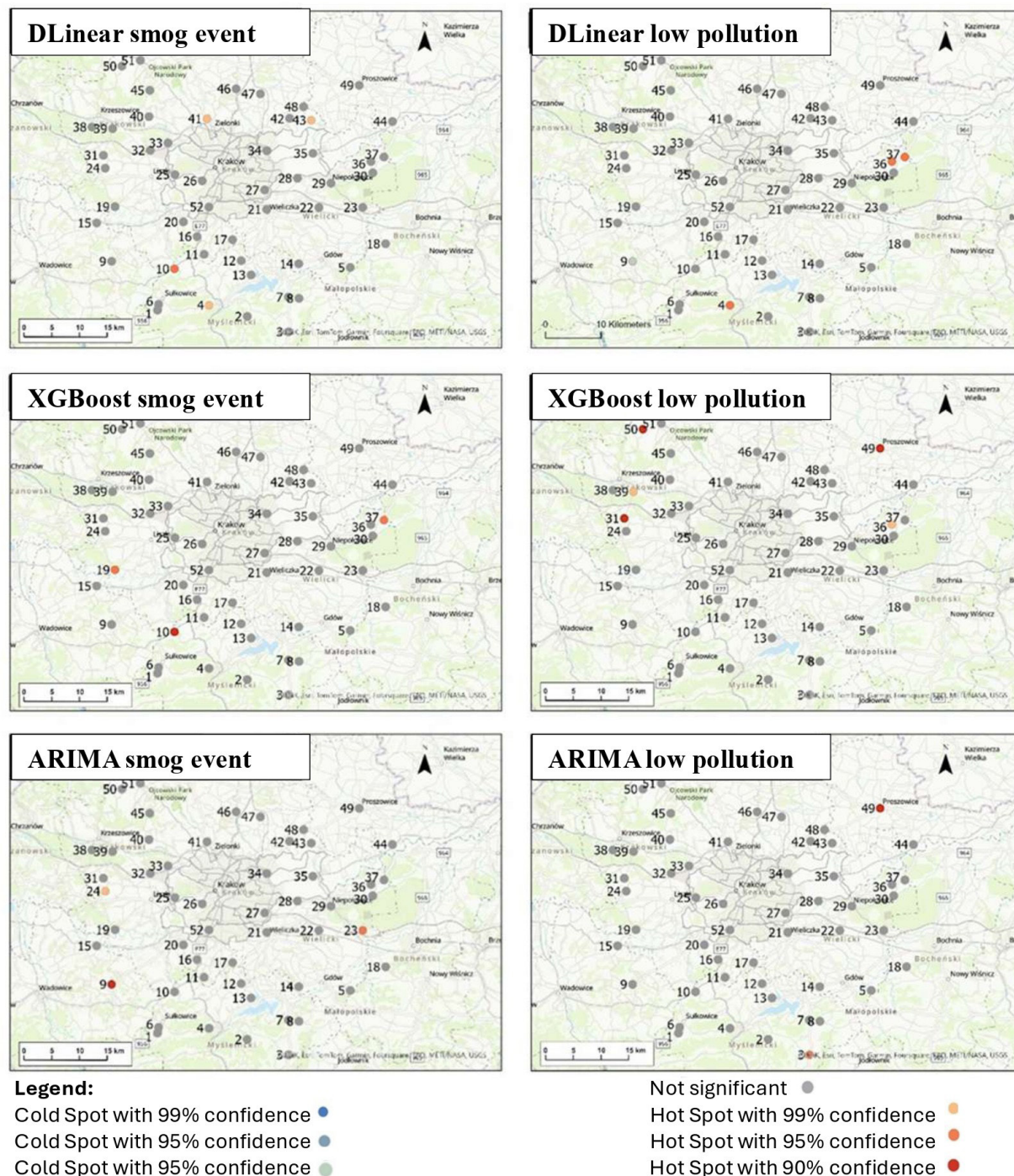


Figure 7. Hot-spot and cold-spot maps for PM_{2.5} error predictions using Getis-Ord G_i^* for different models during smog event and low pollution days
Source: own study

sources. It is essential to increase funding and intensify efforts at the regional level, so that not only the city of Kraków but also the surrounding areas, which potentially produce pollution, employ environmentally friendly heating sources. Such actions would yield health benefits for the entire region.

Prediction errors play a key role in both model evaluation and XAI. Defined as the difference between observed and predicted PM_{2.5} values, they provide insights into model reliability, while

also revealing spatial and temporal patterns that can improve interpretability. Unusual errors may indicate sensor limitations, missing features, or complex relationships that are not effectively captured by the model. Building on previous research by Zareba et al. (2024a), this study shows that errors are not just performance metrics but can enhance our understanding of model behaviour. Spatial patterns in errors highlight areas where predictions are less reliable, offering a way to cross-check feature importance

and detect biases. $PM_{2.5}$ prediction errors during ESE varied significantly – LCS 10, 17, and 22 had very high errors, especially for the XGBoost model (over $130 \mu\text{g}/\text{m}^3$), while for the DLinear models, the errors are much lower ($20 \mu\text{g}/\text{m}^3$ or less). The ARIMA local model is challenging to evaluate: it shows lower errors than XGBoost in some cases but tends to have higher errors in others. Overall, DLinear is the most accurate model, outperforming XGBoost due to specific data characteristics and model mechanics. For example:

1. DLinear models are designed to predict temporal dependencies and, thanks to their linear architecture, they can be more reactive to rapid changes in datasets, handling outlier predictions (like ESE) better. In contrast, the XGBoost model is not primarily designed for time series analysis and may struggle to learn temporal patterns.
2. The DLinear model architecture uses trend, seasonality, and residuals decomposition, which naturally allows both the general trend and short-term rare events, such as smog spikes, to be captured. XGBoost, which is based on ensemble methods with decision trees, generally provides good predictions but prioritizes overall trend predictions rather than rare peaks, as splits are done based on the most representative part of the data, not rare events. The DLinear model's sensitivity to ESE is higher than that of XGBoost. It is important to mention that the analysis of high-error areas from XGBoost errors can be used as a feature or be part of the interpretation.
3. The loss function in XGBoost is typically mean square error, which is not optimal for learning patterns with rare peaks, as the goal is to minimize the error for the majority of the data. DLinear models, on the other hand, typically use a quantile loss function that is more inclusive of large errors, allowing the model to learn from spike events.

A key direction for further research is to use errors as lagged features in hybrid models. Capturing spatial and temporal dependencies in errors could improve both predictive accuracy and interpretability, aligning with XAI principles to make air quality models more transparent and trustworthy.

References

- Airly 2024. Available from: <<http://www.airly.com>>. [18 October 2024].
- Arnaut, F, Đurđević, V, Kolarski, A, Srećković, VA & Jevremović, S 2024, 'Improving air quality data reliability through bi-directional univariate imputation with the random forest algorithm', *Sustainability*, vol. 16, no. 17, article number 7629.
- Box, GEP & Jenkins, GM 1970, *Time series analysis: Forecasting and control*, Holden-Day, San Francisco.
- Chen, T & Guestrin, C 2016, *XGBoost: A scalable tree boosting system*, arXiv, Cornell University, US. Available from: <<http://arxiv.org/abs/1603.02754>>. [18 October 2024].
- Danek, T & Zareba, M 2021, 'The use of public data from low-cost sensors for the geospatial analysis of air pollution from solid fuel heating during the COVID-19 pandemic spring period in Krakow, Poland', *Sensors*, vol.21, no. 15, article number 5208.
- Danek, T, Weglinska, 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*, vol.12, article number 11050.
- Di Antonio, A, Popoola, OAM, Ouyang, B, Saffell, J & Jones, R.L 2018, 'Developing a relative humidity correction for low-cost sensors measuring ambient particulate matter', *Sensors*, vol. 18, no.9, article number 2790.
- ESRI Learning Center 2021, *ArcGIS Pro [Computer Software]: Release 2.8*, Redlands, ESRI
- Firszt, D & Jabłoński, M 2023, 'Implementacja idei smart city w zarządzaniu miastem. Ocena i oczekiwania mieszkańców na przykładzie studentów krakowskich uczelni' ['Implementation of the smart city idea in city management. Assessment and expectations of residents on the example of students of Krakow universities'], *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, vol. 67, no.4, pp. 51–64. Available from: <<https://www.dbc.wroc.pl/publication/164087>>. [18 October 2024].
- Getis, A & Ord, J 1992, 'The analysis of spatial association by use of distance statistics', *Geographical Analysis*, vol. 24, no.3, pp. 189–206.
- Herzen, J, Lässig, F, Piazzetta, S.G, Neuer, T, Tafti, L, Raille, G, Van Pottelbergh, T, Pasieka, M, Skrodzki, A, Huguenin, N, Dumonal, M, Kosciusz, J, Bader, D, Gusset, F, Benheddi, M, Williamson, C, Kosinski, M, Petrik, M & Grosch, G 2021,

Conclusions

For annual average $PM_{2.5}$ concentrations, a clear decreasing trend is evident. The yearly $PM_{2.5}$ average levels (Bujaka station) have declined from $44 \mu\text{g}/\text{m}^3$ to $20 \mu\text{g}/\text{m}^3$. There is a connection between the $PM_{2.5}$ trend and financial investments in the PONE programme, the number of coal-fired stoves or boilers replaced, and the number of installations based on renewable energy sources. Despite these positive actions, Kraków struggles with ESE, primarily due to the influx of pollutants from outside the city. Predicting these rare episodes is crucial and requires a complex spatiotemporal approach, as the dynamics of the atmosphere and $PM_{2.5}$ migration pathways and deposition trends are characterized by a high degree of complexity. The research analysed ESE and low pollution days. One of the key findings of our study is that DLinear models consistently outperform other methods in air pollution forecasting, particularly for rare smog episodes, when trained as a global model across all monitoring stations. Prediction errors are generally the lowest and least varied for global linear models of neural networks (DLinear). The global XGBoost model exhibited the largest errors, reaching nearly $130 \mu\text{g}/\text{m}^3$ (13 times the accuracy of the particulate matter measuring device) on ESE days, while, on days without such episodes, this dependency was exhibited by the local model of traditional ARIMA prediction. Detailed feature contribution analysis for LCS with the highest errors showed a clear connection with vapour-related features.

ORCID

Mateusz Zareba  <https://orcid.org/0000-0002-9663-6593>
 Szymon Cogiel  <https://orcid.org/0009-0000-0574-1260>
 Elżbieta Węglińska  <https://orcid.org/0000-0003-0645-6570>
 Tomasz Danek  <https://orcid.org/0000-0001-8101-7469>

Funding

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.

- Darts: User-friendly modern machine learning for time series*, arXiv, Cornell University, US. Available from: <<https://arxiv.org/abs/2110.03224>>. [18 October 2024].
- Jonek-Kowalska, I 2023, 'Assessing the effectiveness of air quality improvements in Polish cities aspiring to be sustainably smart', *Smart Cities*, vol. 6, no. 1, pp. 510–530.
- Kobus, D, Nych, A & Sowka, I 2018, 'Analysis of PM10 high concentration episodes in Warsaw, Krakow and Wroclaw in the years 2005–2017 with application of selected elements of information systems', *E3S Web of Conferences*, vol. 44, article number 00070.
- Kraków, M.P.K.M 2019, *Program Ograniczania Niskiej Emisji [Low-Emission Reduction Programme]*, Available from: <https://www.krakow.pl/aktualnosci/209034,29,komunikat,mobi_short,program_ograniczania_niskiej_emisji_w_pigulce.html>. [8 October 2024].
- Liang, L & Gong, P 2020, 'Urban and air pollution: A multi-city study of long-term effects of urban landscape patterns on air quality trends', *Scientific Reports*, vol.10, article number 18618.
- Lundberg, S & Lee, S-I 2017, *A unified approach to interpreting model predictions*, arXiv. Available from: <<https://arxiv.org/abs/1705.07874>>. [30 March 2025].
- Morawska-Horawska, M & Lewik, P 2003, 'Wpływ wysokości i ukształtowania terenu na zróżnicowanie warunków meteorologicznych w Krakowie' ['Influence of altitude and terrain on the variation of meteorological conditions in Krakow'] in: *Dynamika zmian środowiska geograficznego pod wpływem antropopresji [Dynamics of changes in the geographic environment under anthropogenic pressure]*, ed. J Lach, Instytut Geografii Akademii Pedagogicznej, Kraków, pp. 85–94.
- Peltier, RE, Castell, N, Clements, AL, Dye, T, Hüglin, C, Kroll, JH, Lung, SCC, Ning, Z, Parsons, M, Penza, M, Reisen, F & Schneidemesser, E 2020, *An update on low-cost sensors for the measurement of atmospheric composition*, December 2020, World Meteorological Organization, Geneva, Switzerland.
- Rada Ministrów 2019, *Rozporządzenie w sprawie wartości odniesienia dla niektórych substancji w powietrzu z dnia 24 września 2019 r.* [Regulation on Reference Values for Certain Substances in the Air of 24 September 2019(Dz. U. 2019, poz. 1931)]. Available from: <<https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20190001931>>. [20 February 2025].
- South Coast Air Quality Management District n.d., *Airly Sensor Evaluation*. Available from: <<https://www.aqmd.gov/aq-spec/sensordetail/airly>>. [20 February 2025].
- Šedivá, T & Štefánik, D 2024, 'The seasonality of PM and NO₂ concentrations in Slovakia and a comparison with chemical-transport model', *Atmosphere*, vol. 15, no. 10, article number 1203.
- Tsyro, S 2005, 'To what extent can aerosol water explain the discrepancy between model calculated and gravimetric PM10 and PM2.5?' *Atmospheric Chemistry and Physics*, vol. 5, no. 2, pp. 515–532.
- Ullah, A, Anwar, SM, Li, J, Nadeem, L, Mahmood, T, Rehman, A & Saba, T 2024, 'Smart cities: the role of Internet of Things and machine learning in realizing a data-centric smart environment', *Complex & Intelligent Systems*, vol. 10, pp. 1607–1637.
- WHO 2021, *WHO global air quality guidelines: Particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. Available from: <<https://iris.who.int/handle/10665/345329>>. [20 February 2025].
- Zareba, M & Danek, T 2021, 'Analysis of air pollution migration during COVID-19 lockdown in Krakow, Poland', *Aerosol and Air Quality Research*, vol. 22, no. 3, article number 210275.
- Zareba, M, Dlugosz, H, Danek, T & Weglinska, E 2023, 'Big-data-driven machine learning for enhancing spatiotemporal air pollution pattern analysis', *Atmosphere*, vol. 14, no. 2, article number 760.
- Zareba, M, Cogieli, S, Danek, T & Weglinska, E 2024a, 'Machine learning techniques for spatio-temporal air pollution prediction to drive sustainable urban development in the era of energy and data transformation', *Energies*, vol. 17, no. 11, article number 2738.
- Zareba, M, Weglinska, E & Danek, T 2024b, 'Air pollution seasons in urban moderate climate areas through big data analytics', *Scientific Reports*, vol. 14, article number 3058.
- Zeng, A, Chen, M, Zhang, L & Xu, Q 2022, *Are transformers effective for time series forecasting?*, arXiv preprint Available from: <https://doi.org/10.48550/arXiv.2205.13504>. [20 February 2025].
- Zięba, D & Dworakowska, A 2018, *Podsumowanie wyników badań porównawczych urządzeń niskokosztowych do pomiaru stężenia pyłu zawieszonego [Summary of the Results of Comparative Studies of Low-Cost Devices for Measuring Particulate Matter Concentration]*, Stowarzyszenie Krakowski Alarm Smogowy.