applied sciences

MDPI

Article

Spatial Modeling of $PM_{2.5}$ Concentrations Using Random Forest and Geostatistical Interpolation in Kraków, Poland

Elżbieta Węglińska, Mateusz Zaręba and Tomasz Danek *

Department of Geoinformatics and Applied Computer Science, Faculty of Geology, Geophysics and Environmental Protection, AGH University of Kraków, 30-059 Krakow, Poland; weglinska@agh.edu.pl (E.W.); zareba@agh.edu.pl (M.Z.)

* Correspondence: tdanek@agh.edu.pl

Abstract

Spatial mapping of $PM_{2.5}$ in complex urban and suburban terrains remains challenging for classical geostatistical interpolation. This study evaluates a Random Forest (RF) framework for high-resolution air pollution mapping and compares its performance with ordinary kriging in the Kraków region. The analysis integrates measurements from 51 low-cost air quality sensors with topographic and meteorological predictors, including elevation, temperature, relative humidity, and wind speed. Five representative hours during a relatively windless, inversion dominated day were selected to examine hourly variability in pollution patterns. Model robustness was assessed using leave-one-out (LOO) cross-validation, while interpretability was addressed through permutation-based predictor importance analysis. The RF model achieved high predictive accuracy (R^2 = 0.85 to 0.95) and good spatial stability with an LOO standard error below 5%. Elevation consistently emerged as the dominant predictor, confirming the key role of terrain-controlled accumulation, while temperature and humidity gained importance during evening and nighttime hours. The RF approach captured fine-scale transport features along river valleys that were not resolved by ordinary kriging, which produced smoother but less interpretable surfaces. The results demonstrate that RF mapping provides an accurate and explainable support to traditional geostatistical methods for analyzing urban air pollution dynamics in complex terrain.

Keywords: random forest; kriging; $PM_{2.5}$; spatial mapping

1. Introduction

In recent years, the study of air pollution, its environmental drivers, and its health consequences has become a central research topic in both atmospheric science and public health. A substantial body of evidence demonstrates the profound impact of chronic and acute exposure to particulate matter on human well-being. Numerous epidemiological studies have documented a wide spectrum of adverse outcomes, including respiratory diseases [1–7], neurodegenerative disorders, cardiometabolic impairments, as well as elevated risks of premature mortality and cancer [8]. This growing corpus of research underscores the necessity of precise, high-resolution air quality monitoring and the development of modern analytical methods capable of capturing both spatial and temporal processes.

Parallel to health impact studies, numerous investigations have focused on long-term trend analysis and descriptive characterization of air pollution using both reference-grade stations—typically operating on gravimetric principles—and a rapidly expanding network of low-cost sensors (LCSs). Although LCS devices exhibit lower precision, their affordability

Check for updates

Academic Editor: Gianluigi Ferrari

Received: 14 January 2026
Revised: 20 February 2026
Accepted: 27 February 2026
Published: 4 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

and scalability enable dense spatial coverage, producing high-granularity datasets that support localized assessments of pollution variability. Studies of this type have been conducted across diverse geographical contexts, from Poland (e.g., focusing on the Kraków metropolitan area [9–11]) to the Czech Republic [12], Serbia [13], China [14], and the United States [15], reflecting a global shift toward distributed sensing and data-driven environmental monitoring.

Another vibrant line of research involves forecasting pollution concentrations using a wide range of modeling paradigms, from classical time-series approaches such as ARIMA models to advanced machine learning and hybrid statistical–computational frameworks [16]. Complementary to these predictive efforts is a growing methodological emphasis on diagnostic and explanatory analyses leveraging Explainable AI (XAI) or geostatistical models. For example, Danek et al. [8] employed Geographically Weighted Regression (GWR) to characterize the spatially varying influence of meteorological drivers, while Zareba and Danek [17] introduced an integrated deep learning and geostatistical–XAI framework capable of capturing seasonal variability and temporal dynamics of meteorological mechanisms responsible for smog episodes.

More recently, the field has witnessed an accelerated growth in applications of large language models (LLMs), primarily within literature synthesis, decision support pipelines, and health-oriented recommendation systems. A particularly innovative contribution in this domain is the work of Cogiel et al. [18], who demonstrated the use of multimodal AI (MLLMs) combined with visual context engineering to automate the interpretation of pollution intensity maps. This methodological advancement is especially important given the explosive growth of IoT-based LCS networks: as sensor density and sampling frequency increase, the manual interpretation of spatiotemporal pollution maps becomes impractical. These findings highlight that for MLLMs to generalize effectively, the cartographic and visual quality of input maps—including interpolation techniques, color encoding, normalization, and spatial representation—is crucial.

Traditional geostatistical methods—particularly kriging—remain widely used for spatial interpolation; however, their practical application in contemporary environmental monitoring workflows is often constrained by several methodological and computational limitations. Kriging requires explicit modeling of the spatial covariance structure through a variogram [19], a process that is both statistically delicate and highly dependent on expert judgment. Inaccurate or poorly parameterized variograms can propagate substantial errors into interpolated surfaces, especially in heterogeneous or meteorologically dynamic environments. Moreover, kriging's underlying assumptions of second-order stationarity and isotropy, while mathematically convenient, rarely reflect the actual complexity of urban atmospheric systems. When spatial gradients or directional transport processes (e.g., valley channeled winds or boundary layer inversions) violate these assumptions, kriging may produce overly smoothed estimates under strong nonstationarity or directional transport. Its computational cost also increases rapidly with dataset size, making the method less suitable for high-density sensor networks or real-time mapping applications. Even advanced variants such as universal kriging or anisotropic kriging only partially mitigate these issues and often introduce additional layers of model complexity. As a result, despite its theoretical elegance, kriging may produce misleading spatial patterns when applied in contexts characterized by strong nonstationarity, irregular sampling geometry, or rapidly evolving pollution phenomena.

In contrast, Random Forest (RF) [20] regression has emerged as a robust and flexible complementary data-driven approach for spatial prediction. Multiple methodological extensions have been proposed, including the foundational approach by Hengl et al. [21], spatially attenuated variants accounting for distance-dependent relationships [22], and

formulations designed to accommodate multi-resolution predictors and high-dimensional feature spaces [23]. In this study, the approach of Hengl et al. [21] is adopted due to the relatively limited spatial extent of the study area, the uniform distribution of point measurements, and the consistent raster data resolution. RF models provide stable predictions across nonlinear feature interactions, and their ability to incorporate a wide array of explanatory variables—such as topographic morphology or meteorological conditions (temperature, humidity, and wind speed)—offers a substantial advantage over classical geostatistics. Furthermore, embedded measures of predictor importance align RF with the broader goals of Explainable AI, allowing for transparent assessment of environmental drivers. This study tests the hypothesis that Random Forest-based spatial mapping differs from ordinary kriging in its ability to preserve local-scale spatial variability in sparse sensor observations.

Within the regional air pollution typology developed by Morawiec et al. [24], Kraków is classified as part of the southern urban–industrial macroregion, which is characterized by persistently elevated pollutant concentrations and intensive anthropogenic pressure. This classification provides a structured spatial context for the present study and identifies Kraków as a representative and critical case for air quality analysis. Despite the implementation of stringent local mitigation measures, including prohibitions on coal and wood combustion within municipal boundaries, the city has experienced chronic air pollution exposure over multiple decades. In contrast, adjacent municipalities operate under less restrictive regulatory regimes, resulting in uneven emission controls and fostering complex transboundary pollution processes that significantly affect the spatial and temporal variability of air quality within Kraków. Spatial pollution mapping is therefore essential for disentangling local emission contributions from regional inflow patterns. Kraków constitutes a particularly compelling case study not only because it lies within the highly regulated environmental framework of the European Union but also because it is situated within a national energy system whose characteristics resemble those of fast-developing regions in Asia or South America [25]. The city's topography—a valley bordered by several towns lacking strict emission controls—creates a natural testbed for investigating pollutant dispersion, cold-air pooling, and boundary layer dynamics. The application of RF-based mapping to time-series data clearly illustrates how these processes modulate the spatial and temporal distributions of particulate matter across the metropolitan region.

The contribution of this study lies in the application of a rigorously validated Random Forest-based mapping framework for generating high-resolution air pollution maps in a complex urban environment. Specifically, the study contributes (i) a validated spatial mapping framework integrating dense low-cost sensor data with meteorological and topographic predictors, (ii) an explicit assessment of intra-day temporal variability and spatial robustness using leave-one-out validation, and (iii) a transparent evaluation of predictor importance to enhance interpretability. The framework integrates dense low-cost sensor data with meteorological and topographic predictors, evaluates intra-day temporal variability, and quantifies predictor importance, enabling the identification of fine-scale spatial patterns that are not readily captured by classical geostatistical interpolation. The study provides a concise and reproducible workflow for RF-based air quality mapping and illustrates its practical value for spatial analysis of $PM_{2.5}$ in Kraków. Rather than introducing a completely new prediction algorithm, the study focuses on application-level validation, spatial robustness, and interpretability in the specific context of dense urban sensor networks.

2. Materials and Methods

2.1. Sensors and Area Characterization

Air quality assessment in Poland is embedded within the regulatory framework defined by the European Union's (EU) Ambient Air Quality Directive (2008/50/EC) [26]. This legislation specifies permissible concentrations of particulate matter and prescribes standardized measurement protocols to ensure comparability across member states. While the EU maintains a substantial network of reference air monitoring stations, their spatial distribution is intentionally sparse due to the high cost and strict maintenance requirements of reference instruments. Kraków exemplifies this limitation: despite chronic pollution challenges, the city operates only a handful of reference-grade monitors for PM_{10} and $PM_{2.5}$. However, their limited number prevents capturing fine-scale spatial variability across the urban landscape.

To overcome these spatial gaps, the current study integrates a dense network of Airly LCS (www.airly.org, accessed on 1 September 2025) optical devices. These sensors operate on light-scattering principles to estimate particulate concentrations and can be deployed extensively across the city, providing high-resolution coverage impossible with reference stations alone. Although LCS measurements are not certified for regulatory reporting under EU law, numerous intercomparison studies have demonstrated that once corrected for environmental influences such as humidity, temperature, and local microclimate, LCS data can approximate reference station readings with high accuracy. For instance, Airly sensors used in this study have been reported to achieve a root mean square error (RMSE) of 3–6 $\mu g/m^3$ for $PM_{2.5}$ and 5–8 $\mu g/m^3$ for PM_{10}, with correlation coefficients (R^2) ranging from 0.85 to 0.93 under typical urban conditions.

The Airly sensors used in this study employ an MCERTS-certified optical system for real-time measurement of PM_{10}, $PM_{2.5}$, and $PM_{1.0}$. While the manufacturer does not publicly disclose the internal sensor model [27], independent field evaluations report the integration of a Plantower PMS5003 laser-based particulate matter sensor [28]. Measurements are calibrated at the system level against reference-grade stations and disseminated after the application of proprietary internal correction and quality control procedures [29].

Each Airly device is continuously recalibrated using proprietary machine learning models that account for local environmental conditions, nearby reference measurements, and temporal drift. While the precise calibration algorithms are not publicly disclosed, similar ML-based correction techniques have been shown to reduce raw measurement errors by 30–50% and substantially improve agreement with reference stations. The data accessed through the Airly API are already postprocessed, ensuring that the measurements used in this study represent corrected and reliable particulate concentrations rather than raw sensor outputs.

Dense LCS networks have previously enabled analyses that would be impossible using reference stations alone. For example, during the spring 2020 COVID-19 lockdown, detailed studies combining LCS observations with chemical composition analyses revealed that the majority of particulate matter reaching Kraków originated from solid fuel heating in surrounding municipalities rather than from transportation, which was drastically reduced. Carbonaceous aerosols from coal combustion accounted for approximately 50% of PM_{10} during the heating season, with secondary inorganic aerosols contributing around 20%, metals 3–4%, and other unidentified components the remainder. Seasonal variability was observed, with transportation emissions playing a larger role during summer months. These findings align with regional emission inventories, which identify residential solid fuel combustion as the primary contributor to approximately half of annual PM_{10} and $PM_{2.5}$ emissions.

In this study, the high spatial density and quantitative reliability of LCS data were essential for the RF-based mapping framework. The network comprised 51 Airly sensors distributed across the Kraków metropolitan area (Figure 1). Topographic information was integrated from the Copernicus Digital Elevation Model (DEM) [30], while meteorological variables including temperature, humidity, and wind speed were obtained from the Open-Meteo database [31]. Meteorological data, originally at a 5 km × 5 km resolution, were rescaled to a 500 m × 500 m grid to match the model resolution. The combination of LCS, topographic, and meteorological data allowed for fine-grained, data-driven reconstruction of particulate matter distribution, capturing microscale spatial patterns and enabling a robust assessment of predictor importance in an urban setting. By providing dense, temporally resolved measurements, the LCS network formed the backbone of this study's high-resolution air pollution mapping approach, bridging the gap between sparse reference stations and the needs of modern machine learning models.

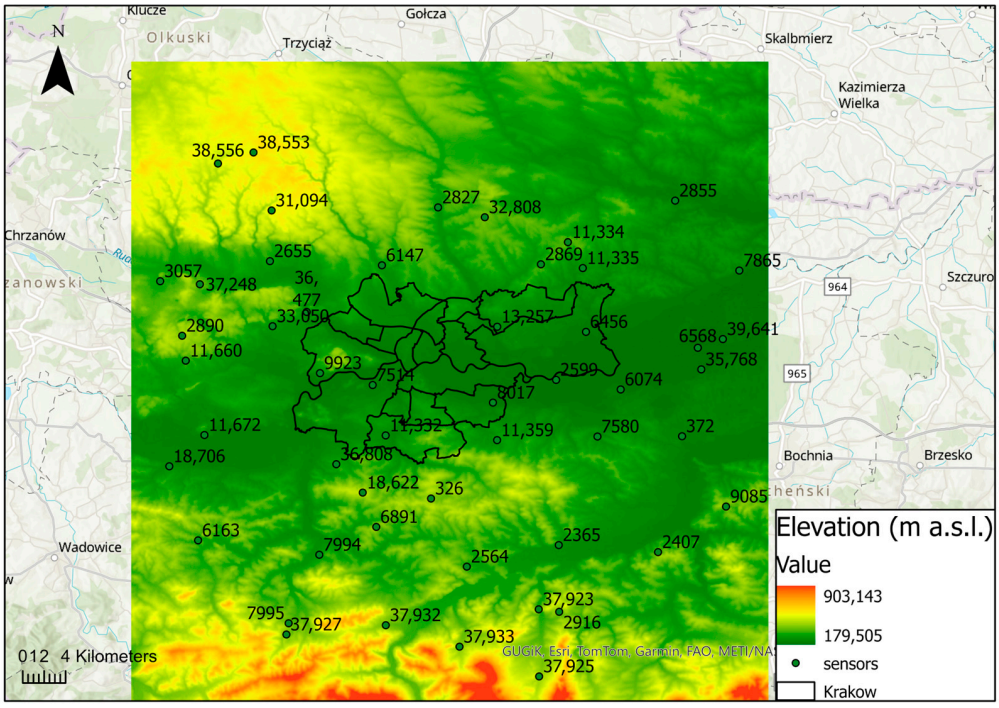


Figure 1. Elevation map of Kraków area with sensor locations.

The analysis was conducted for 23–24 March 2022, focusing on five representative hours (23 March at 6:00, 16:00, 20:00, and 23:00 and 24 March at 1:00). The selected study period represents a typical winter smog episode (heating related characterized by diurnal temperature fluctuations around the freezing point, low wind, and intensified residential combustion) and was used to validate the methodological performance of the proposed approach by illustrating its temporal robustness and spatial stability under different daily emission regimes, rather than to draw wider climatological conclusions. Extension to longer time periods is beyond the scope of this study.

It is important to note that air quality data were retrieved via the Airly API as hourly averaged $PM_{2.5}$ concentrations, using values already corrected by the data provider. Raw sensor data and API keys cannot be shared due to provider restrictions, but the described workflow is transferable to other dense sensor networks.

2.2. Random Forest

Random Forest regression was applied as a data-driven framework for spatial mapping of $PM_{2.5}$ concentrations. The method was selected due to its ability to model nonlinear

relationships and to integrate heterogeneous predictors without requiring assumptions of stationarity or predefined spatial covariance structures.

For each analyzed hour, point measurements of $PM_{2.5}$ from low-cost sensors were combined with coincident meteorological observations, including air temperature, relative humidity, and wind speed. In addition, elevation was extracted from a Digital Elevation Model at sensor locations and included as a topographic predictor. All variables were assembled into a tabular dataset representing the conditions at a given time step.

Spatial dependency was introduced explicitly through distance-based predictors. For each sensor location, Euclidean distances to all other sensors were computed and added as individual predictors. This approach allows the Random Forest model to learn spatial relationships directly from the data rather than relying on an explicit variogram model as in classical geostatistics. Similar distance-based representations have been used in established Random Forest frameworks for spatial prediction to represent spatial proximity and autocorrelation [21]. While this increases the dimensionality of the predictor space, Random Forest is relatively robust to correlated features, and model performance was evaluated using leave-one-out validation to reduce potential location-specific effects. In this study, the distance-based representation provides a simple and reproducible way to account for spatial proximity in a dense sensor network. More compact spatial encodings were not explored and remain beyond the scope of this work.

Meteorological variables were prepared as raster layers for each time step. Gridded temperature, humidity, and wind speed fields were projected to a common coordinate system, resampled to the target grid resolution, and spatially smoothed to reduce local artifacts. These raster predictors were then used as inputs for spatial prediction across the interpolation grid.

The Random Forest model was fitted using the ranger implementation [32] using 500 trees, considering one-third of the available predictors at each split, and a minimum node size of five. Model training was performed separately for each selected hour to capture short-term variability in pollution patterns. Predictions were generated for all grid cells by combining raster based meteorological predictors with distance-based spatial features, producing continuous $PM_{2.5}$ concentration surfaces. All variables used as inputs to the Random Forest model, including the target variable and meteorological, topographic, and spatial predictors, are summarized in Table 1.

Table 1. Target variable and predictors used for training the Random Forest model.

Category	Variable	Description/Source
Target variable	$PM_{2.5}$ concentration	Hourly average $PM_{2.5}$ concentration from Airly low-cost sensors
Meteorological predictor	Air temperature	Hourly air temperature from the Open-Meteo database
Meteorological predictor	Relative humidity	Hourly relative humidity from the Open-Meteo database
Meteorological predictor	Wind speed	Hourly wind speed from the Open-Meteo database
Topographic predictor	Elevation	Terrain elevation derived from the Copernicus Digital Elevation Model
Spatial predictors	Euclidean distances to sensor locations	Distance-based spatial predictors representing spatial proximity between sensors

Model performance at measurement locations was evaluated using the coefficient of determination R^2. Spatial robustness was assessed using leave-one-out cross-validation, where models were repeatedly trained with individual sensors removed and the variability of predictions was analyzed across the domain.

To enhance interpretability and support explainable spatial analysis, permutation-based variable importance was calculated for non-spatial predictors [33]. For each variable, importance was calculated by randomly permuting its values and quantifying the resulting decrease in model performance while keeping all other predictors unchanged. Final importance values were estimated by repeatedly refitting the model and averaging permutation scores for temperature, humidity, wind speed, and elevation. This procedure allowed the identification of dominant environmental controls and their temporal variability.

The resulting Random Forest maps were finally compared with surfaces generated using ordinary kriging to evaluate differences in spatial structure, smoothness, and interpretability.

2.3. Ordinary Kriging

Ordinary kriging (OK) is the most commonly applied variant of kriging, which is a geostatistical interpolation method. It provides estimates of values at unsampled locations within a study area based on a known variogram and the information from neighboring observations [34]. The OK estimator is expressed as:

$$Z^*(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) \quad (1)$$

where Z^* is the predicted value at location x_0, $Z(x_i)$ is the observed value at location x_i, λ_i denotes the kriging weight associated with the i-th observation, and n is the number of neighboring sample points used in the estimation.

It assumes a constant but unknown mean across the study area, which makes it particularly suitable for environmental applications such as air pollution mapping [35–38].

Kriging interpolations for all maps were generated using the Geostatistical Wizard tool in ArcGIS Pro 3.0.3, which served solely as an implementation environment, while the final interpolation relied on manually specified parameters rather than default settings. A single spatial covariance structure was assumed for all analyzed time steps. An exponential semivariogram model was used with a nugget of 20, partial sill of 314, and range of 20 km. To ensure temporal comparability, identical color scales were applied to all maps using the global minimum and maximum $PM_{2.5}$ concentrations across the analyzed time steps.

3. Results

Figure 2 shows the diurnal evolution of meteorological conditions at all sensor locations on the analyzed day, with individual time series colored according to site elevation. Air temperature exhibits a coherent daily cycle across the study area, with minimum values during early morning hours and a rapid increase after sunrise, reaching a pronounced maximum in the afternoon. Differences between sites are relatively small during daytime but become more distinct during nighttime and early morning, indicating elevation dependent thermal stratification and the presence of local inversion conditions. Relative humidity follows an inverse pattern to temperature, with high and spatially consistent values during nighttime and early morning hours, followed by a marked decrease around midday. Evening hours show renewed divergence between sites, particularly between lower and higher elevations. Wind speed displays the highest spatial variability among the analyzed variables. Overall wind conditions remain weak to moderate throughout most of the day, with a distinct minimum around midday and a pronounced increase during evening hours. This evening intensification is spatially heterogeneous and more pronounced at higher

elevations, suggesting locally driven circulation patterns. The combined temporal behavior of temperature, humidity, and wind speed reflects meteorological conditions favorable for pollutant accumulation and limited dispersion, providing important context for the spatial patterns of $PM_{2.5}$ modeled in subsequent analyses.

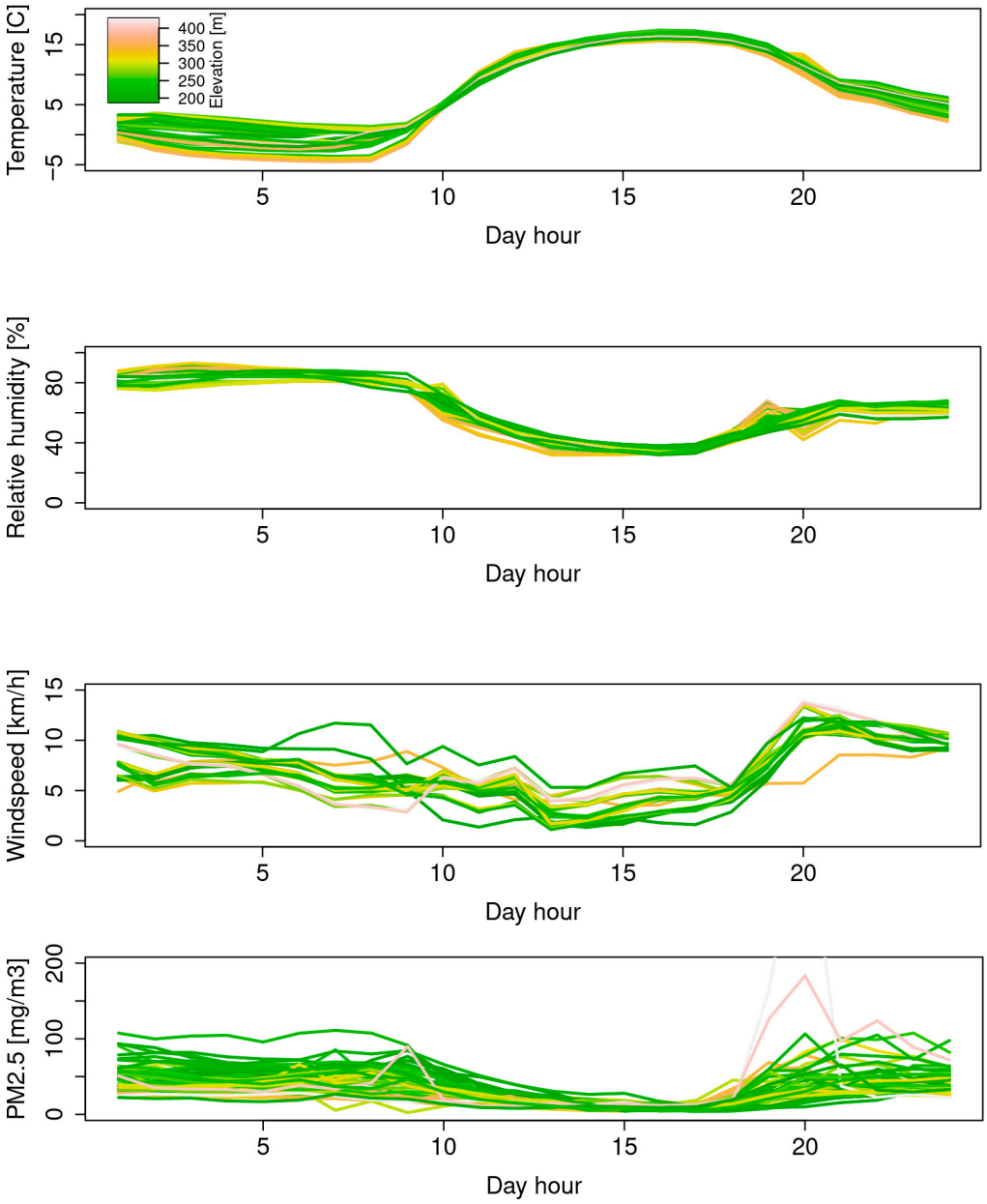


Figure 2. Diurnal variability of meteorological conditions at sensor locations on the analyzed day. Panels show hourly time series of air temperature, relative humidity, wind speed, and $PM_{2.5}$ for all monitoring sites. Individual lines represent measurements from single sensors, with colors indicating elevation extracted from the Digital Elevation Model (green—lowest; orange—highest).

The maps were generated using the Random Forest algorithm for 23 March at 6:00, 16:00, 20:00, and 23:00 and 24 March at 1:00. Figure 3a–e illustrate the model predictions of $PM_{2.5}$ concentrations at these specific hours. To better understand the meteorological background of these processes, Figures 4–6 present smoothed spatial distributions of temperature, humidity, and wind speed at the same hours used as a predictors in main spatial modeling runs. The role of meteorological and topographic predictors in shaping pollution patterns was further assessed using Random Forest permutation importance

analysis (Figure 7). Among the non-spatial predictors, elevation was consistently the most influential variable across all hours, reflecting the significance of terrain-induced accumulation. Temperature and humidity also played important roles, particularly in the afternoon and evening, while wind speed contributed less to the model but still showed localized effects. All maps were generated using ArcGIS Pro software 3.0.3, ESRI.

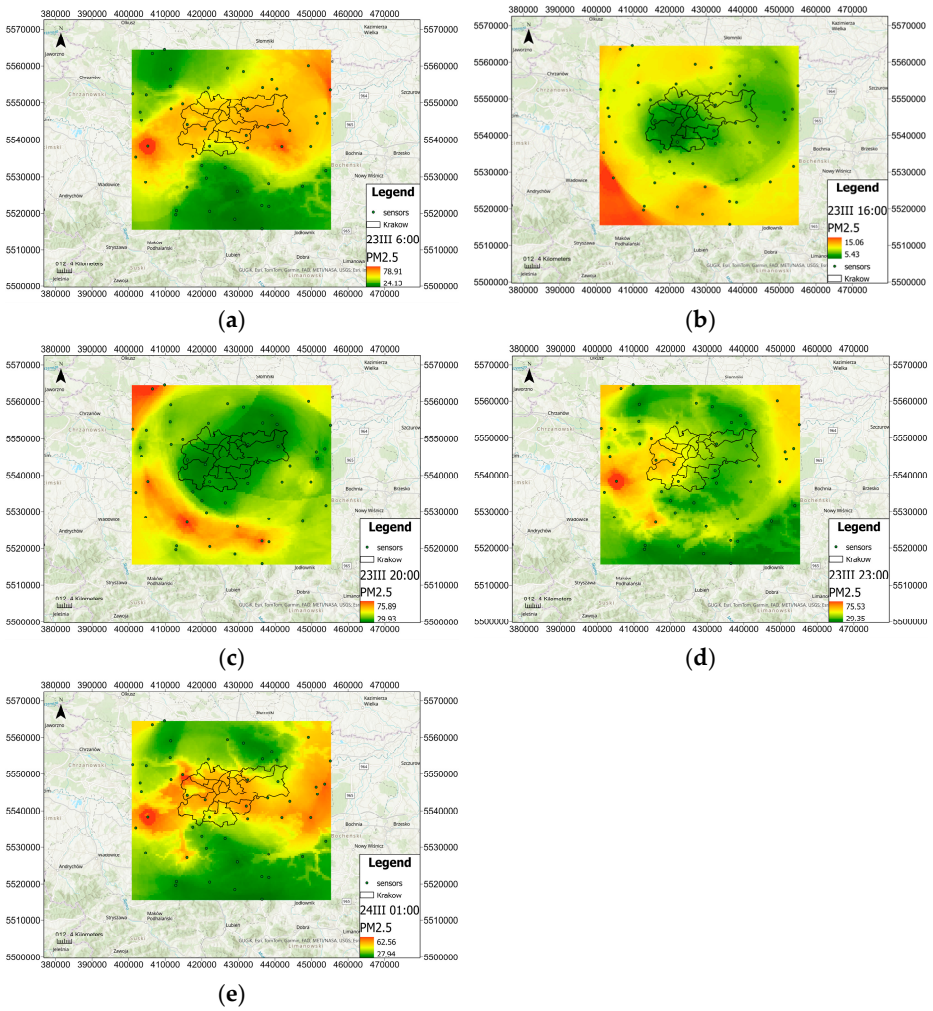


Figure 3. Random Forest model predictions of $PM_{2.5}$ concentrations in the Kraków area on March 23 at 6:00 (**a**), 16:00 (**b**), 20:00 (**c**), and 23:00 (**d**) and on 24 March at 1:00 (**e**). Note the systematic accumulation of $PM_{2.5}$ in river valleys and terrain depressions, particularly during evening and nighttime hours (20:00–1:00), as well as the development of a ring-shaped ("bagel") pollution pattern surrounding the relatively cleaner city center.

Geostatistical maps produced for the same hours exhibit smoother interpolations; however, they lack the capability to effectively incorporate additional predictors. These maps are shown in Figure 8, and they represent $PM_{2.5}$ concentrations generated using ordinary kriging.

The agreement between Random Forest predictions and point measurements was high, with R^2 values typically exceeding 0.9 across the analyzed hours. These R^2 values refer to the agreement between predicted and observed $PM_{2.5}$ concentrations at sensor locations and are reported as a measure of overall model fit rather than out-of-sample predictive performance. Out-of-sample evaluation was performed separately using leave-one-out cross-validation, with the focus on spatial robustness and stability of the predicted concentration fields rather than pointwise error metrics. As illustrated in Figure 9 for the example of 6:00, the standard deviation of LOO predictions remains low over most of the study area,

indicating that the resulting concentration fields are robust to the removal of individual sensors. Slightly higher uncertainty is confined to localized zones, mainly near isolated measurement points or areas with lower sensor density. Overall, the spatial distribution of LOO error confirms that the RF-based mapping framework produces stable and consistent $PM_{2.5}$ patterns while allowing for the identification of locally influential observations.

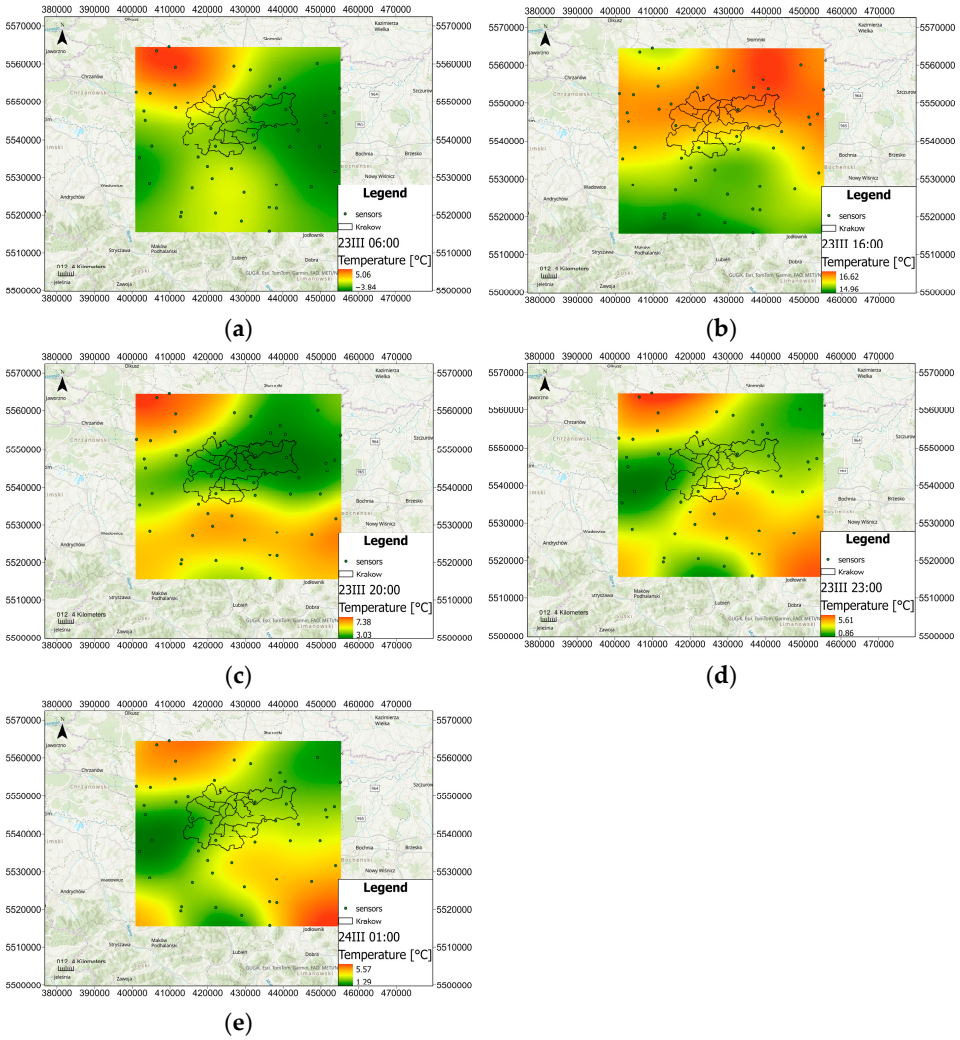


Figure 4. Smoothed Open-Meteo temperature data around Kraków area on 23 March at 6:00 (**a**), 16:00 (**b**), 20:00 (**c**), 23:00 (**d**), and on 24 March at 1:00 (**e**).

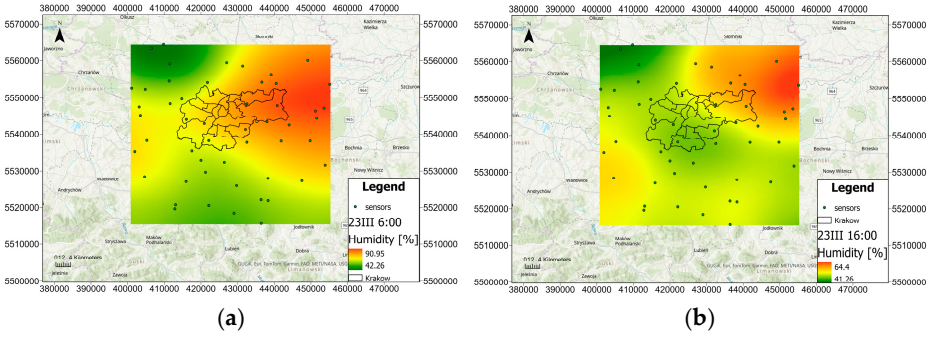


Figure 5. *Cont.*

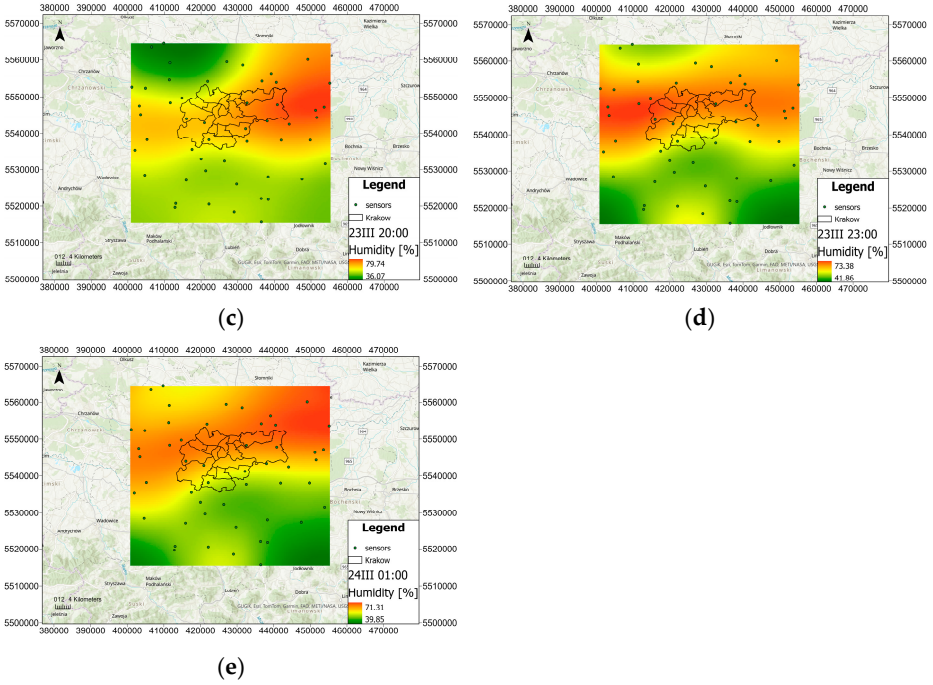


Figure 5. Smoothed Open-Meteo relative humidity data around Kraków area on 23 March at 6:00 (**a**), 16:00 (**b**), 20:00 (**c**), and 23:00 (**d**) and 24 March at 1:00 (**e**).

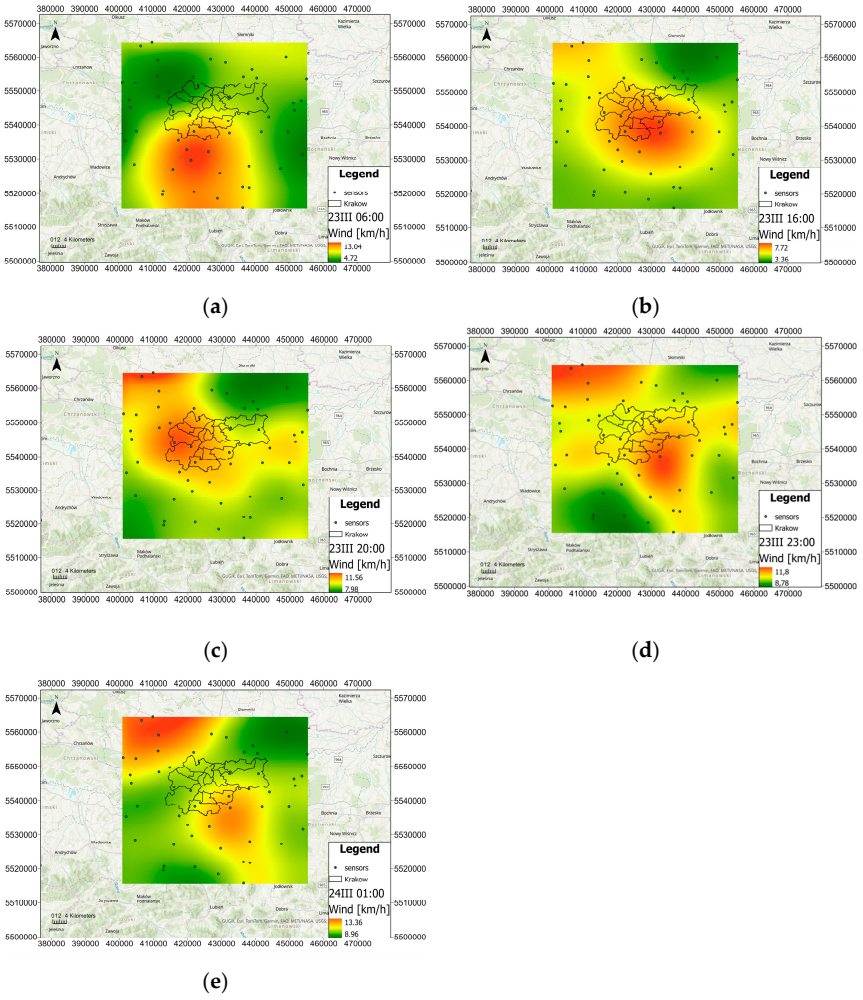


Figure 6. Smoothed Open-Meteo wind speed data around Kraków area on 23 March at 6:00 (**a**), 16:00 (**b**), 20:00 (**c**), and 23:00 (**d**) and 24 March at 1:00 (**e**).

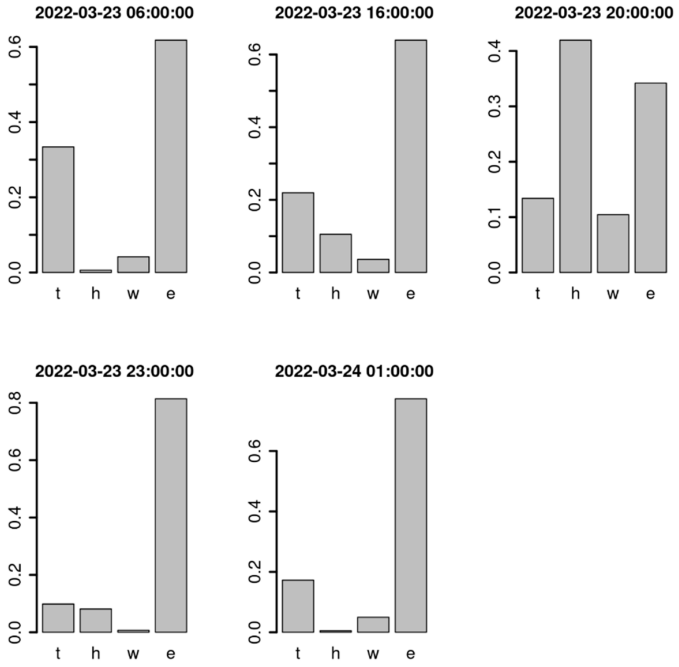
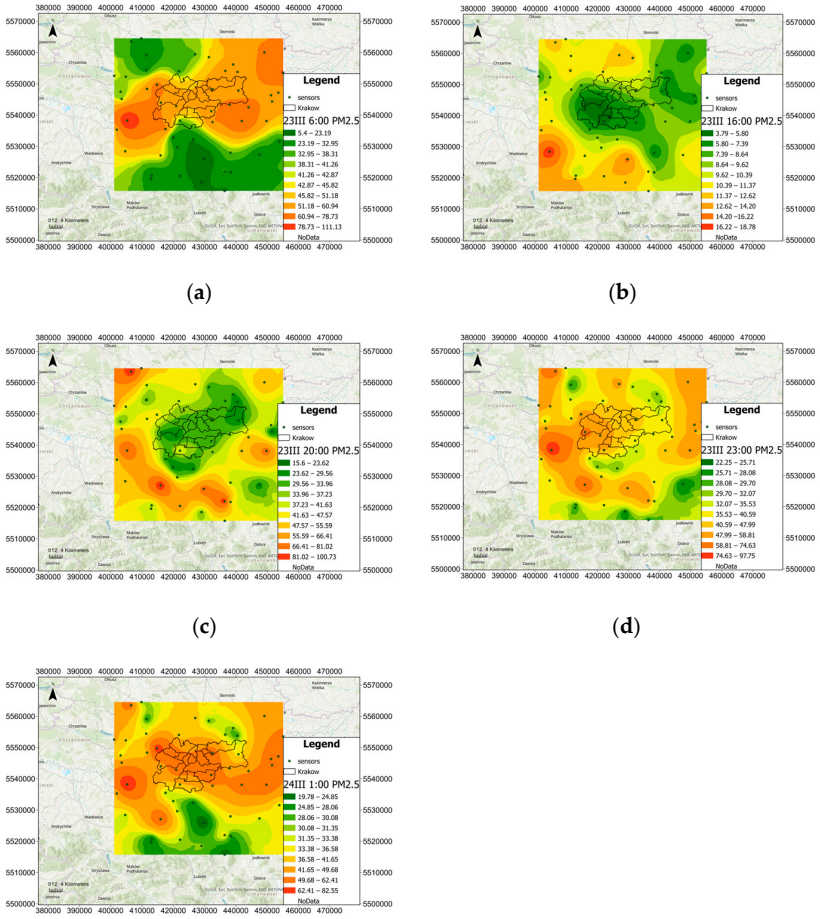


Figure 7. Permutation importance of non-spatial predictors (temperature—t, humidity—h, wind speed—w, and elevation—e) for 6:00, 16:00, 20:00, 23:00, and 1:00 of the following day.



(**a**) (**b**)

(**c**) (**d**)

(**e**)

Figure 8. Spatial interpolation of $PM_{2.5}$ concentration in the studied area on 23 March at 6:00 (**a**), 16 (**b**), 20 (**c**), and 23 (**d**) and 24 March at 1:00 (**e**), generated using kriging. Compared with the Random Forest results, the kriging surfaces appear smoother and less structured, with reduced visibility of fine-scale valley transport pathways and local concentration contrasts.

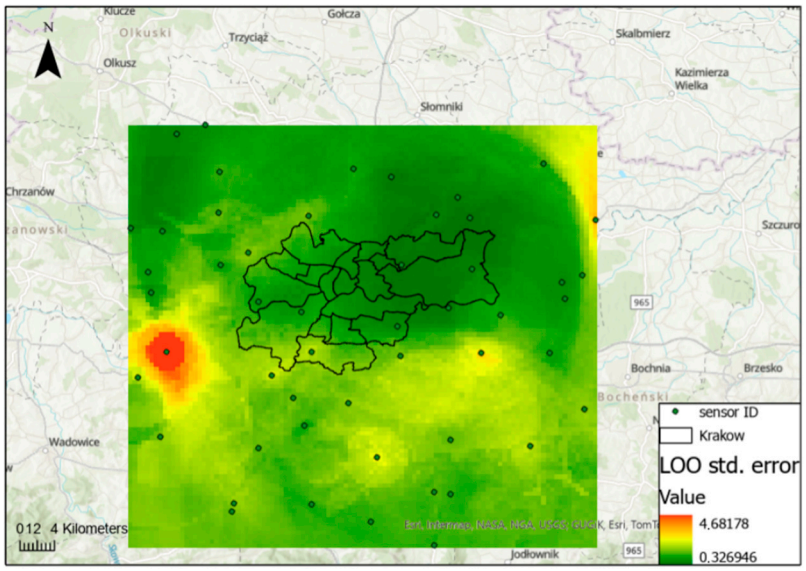


Figure 9. Spatial distribution of LOO standard error for the interpolation model with sensor locations in the Kraków area.

4. Discussion

The spatiotemporal patterns of $PM_{2.5}$ obtained with the RF model reflect the combined effects of topography, meteorological conditions, and emission dynamics. The analysis of five selected time points provides insight into how predictors interact to shape air quality in the Kraków region under inversion and low-wind conditions. In the context of this study, emission dynamics are interpreted indirectly through observed concentration patterns, while the model itself is designed for spatial reconstruction under given meteorological and topographic conditions rather than detailed emission modeling.

At 6:00, elevated $PM_{2.5}$ concentrations were observed in river valleys, indicating potential pollution transport pathways. Mountainous areas remained relatively clean despite intensive fossil fuel use in these regions, likely because residential heating activities had not yet started in the early morning. Elevation emerged as the dominant predictor, with the model highlighting increased pollution in terrain depressions. Temperature ranked second in importance, shaping finer details of the spatial distribution. By 16:00, despite unusually high temperatures for March, the first signs of the so-called "bagel" appeared, consistent with typical evening emission peaks reported for residential heating in the Kraków region [39]. Elevation remained the most important predictor, but the role of temperature increased, and humidity gained significantly in importance, influencing the emerging spatial patterns. At 20:00, the "bagel" became more pronounced. The evening temperature drop favored intensified fossil fuel combustion, while the city center of Kraków remained relatively clean. Humidity was the most influential predictor at this time, likely shaping $PM_{2.5}$ values in less polluted areas. Elevation continued to be important, but the importance of wind speed increased, suggesting that even weak airflow patterns may have contributed to the redistribution of pollutants. By 23:00, pollutant levels decreased in regions dominated by manual heating systems and began to flow into Kraków along the Vistula River valley from the west. Elevation once again dominated predictor importance, while temperature and humidity jointly ranked second. Their influence likely differed by location: humidity-shaping conditions in Kraków as the receptor area and temperature in emission-producing regions. Interestingly, their spatial distributions were highly correlated during this time, reinforcing their combined role in pollution dynamics. At 1:00, pollutant transport along the Skawinka Valley led to accumulation within Kraków. The situation

closely resembled that of 6:00 on the previous day, with a nearly identical hierarchy of predictor importance.

RF models provide not only a high agreement with point measurements but also spatial stability, remaining robust to the removal of individual measurement points (Figure 9). Traditional maps, although effective for interpolation, do not account for important predictors, which limits their interpretability.

The present study has several limitations that should be acknowledged. First, the analysis is based on selected hours representing a single winter smog episode and therefore does not aim to provide generalized conclusions across multiple seasons or pollution regimes. Second, the modeling framework does not explicitly incorporate traffic intensity or detailed emission inventories. Emission dynamics are interpreted indirectly through observed concentration patterns and meteorological data. Third, $PM_{2.5}$ measurements were obtained from low-cost sensors using manufacturer's correction algorithms, and raw data processing procedures are proprietary. While these constraints do not affect the methodological demonstration of the spatial mapping framework, they limit the extent to which source-specific or long-term conclusions can be drawn.

5. Conclusions

This study demonstrated that RF-based spatial modeling provides an effective and interpretable framework for high-resolution mapping of $PM_{2.5}$ concentrations in complex urban terrain. By integrating dense low-cost sensor measurements with meteorological and topographic predictors, the proposed approach achieved high predictive accuracy. Comparison with ordinary kriging showed that traditional geostatistical interpolation produces smoother concentration maps, but it cannot include additional environmental predictors and therefore provides limited insight into the processes driving pollution patterns.

RF preserved fine-scale spatial structures, particularly along river valleys, whereas kriging produced smoother and less detailed concentration fields. The RF model reproduced observed concentrations with coefficients of determination of 0.85–0.95 and showed high spatial robustness in leave-one-out validation (error below 5%).

The presented results demonstrate that spatial prediction based on machine learning can complement classical geostatistical mapping. The methodology is transferable to other cities with dense sensor networks, despite restrictions on raw data sharing.

Author Contributions: Conceptualization, E.W., M.Z., and T.D.; methodology, E.W., M.Z., and T.D.; software, E.W., M.Z., and T.D.; validation, E.W., M.Z., and T.D.; formal analysis, E.W., M.Z., and T.D.; investigation, E.W., M.Z., and T.D.; resources, E.W., M.Z., and T.D.; data curation, E.W., M.Z., and T.D.; writing—original draft preparation, E.W., M.Z., and T.D.; writing—review and editing, E.W., M.Z., and T.D.; visualization, E.W., and T.D.; supervision, T.D.; project administration, E.W., M.Z., and T.D.; funding acquisition, E.W., M.Z., and T.D. All authors have read and agreed to the published version of the manuscript.

Funding: The study was financed by the AGH University of Krakow, Faculty of Geology, Geophysics and Environmental Protection as part of the statutory projects, and by the program "Excellence initiative-research university for the AGH University of Krakow".

Data Availability Statement: Datasets from Airly sensors were analyzed in this study and can be found here: (https://map.airly.org/, accessed on 11 November 2025). API documentation from Airly is available here: (https://developer.airly.org/en/docs, accessed on 11 November 2025). The Airly sensor data analyzed in this study were obtained under a free academic API key issued by Airly S.A. in accordance with the Airly API Service Terms. These terms restrict the redistribution of the raw data and prohibit the sharing of the API key with third parties. Consequently, the underlying data are not publicly available. Researchers can obtain current data directly from Airly by registering

for an academic research API key at (https://map.airly.org/ (accessed on 11 November 2025)) (see AirlyAPI Terms of Service).

Acknowledgments: During the preparation of this manuscript, the authors used OpenAI models (o3, GPT-5.2) and DeepL for translation and language proofreading. The authors have reviewed and edited all outputs and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

OK	Ordinary Kriging
RF	Random Forest
LOO	Leave One Out
LCS	Low-Cost Sensor

References

1. Adamkiewicz, G.; Liddie, J.; Gaffin, J.M. The Respiratory Risks of Ambient/Outdoor Air Pollution. *Clin. Chest Med.* **2020**, *41*, 809–824. [CrossRef]
2. Chen, J.; Zeng, Y.; Lau, A.K.H.; Guo, C.; Wei, X.; Lin, C.; Huang, B.; Lao, X.Q. Chronic Exposure to Ambient PM2.5/NO2 and Respiratory Health in School Children: A Prospective Cohort Study in Hong Kong. *Ecotoxicol. Environ. Saf.* **2023**, *264*, 114558. [CrossRef]
3. Danesh Yazdi, M.; Wang, Y.; Di, Q.; Wei, Y.; Requia, W.J.; Shi, L.; Sabath, M.B.; Dominici, F.; Coull, B.A.; Evans, J.S.; et al. Long-Term Association of Air Pollution and Hospital Admissions Among Medicare Participants Using a Doubly Robust Additive Model. *Circulation* **2021**, *143*, 1584–1596. [CrossRef]
4. Dominici, F.; Peng, R.D.; Bell, M.L.; Pham, L.; McDermott, A.; Zeger, S.L.; Samet, J.M. Fine Particulate Air Pollution and Hospital Admission for Cardiovascular and Respiratory Diseases. *JAMA* **2006**, *295*, 1127–1134. [CrossRef] [PubMed]
5. Lin, S.; Xue, Y.; Thandra, S.; Qi, Q.; Hopke, P.K.; Thurston, S.W.; Croft, D.P.; Utell, M.J.; Rich, D.Q. PM2.5 and Its Components and Respiratory Disease Healthcare Encounters—Unanticipated Increased Exposure–Response Relationships in Recent Years after Environmental Policies. *Environ. Pollut.* **2024**, *360*, 124585. [CrossRef] [PubMed]
6. Ren, Z.; Liu, X.; Liu, T.; Chen, D.; Jiao, K.; Wang, X.; Suo, J.; Yang, H.; Liao, J.; Ma, L. Effect of Ambient Fine Particulates (PM2.5) on Hospital Admissions for Respiratory and Cardiovascular Diseases in Wuhan, China. *Respir. Res.* **2021**, *22*, 128. [CrossRef]
7. Hamanaka, R.B.; Mutlu, G.M. Particulate Matter Air Pollution: Effects on the Respiratory System. *J. Clin. Investig.* **2025**, *135*, e194312. [CrossRef] [PubMed]
8. Raaschou-Nielsen, O.; Andersen, Z.J.; Beelen, R.; Samoli, E.; Stafoggia, M.; Weinmayr, G.; Hoffmann, B.; Fischer, P.; Nieuwenhuijsen, M.J.; Brunekreef, B.; et al. Air pollution and lung cancer incidence in 17 European cohorts: Prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE). *Lancet Oncol.* **2013**, *14*, 813–822. [CrossRef]
9. Danek, T.; Węglińska, E.; Zareba, M. The influence of meteorological factors and terrain on air pollution concentration and migration: A geostatistical case study from Krakow, Poland. *Sci. Rep.* **2022**, *12*, 11050. [CrossRef]
10. Zareba, M.; Weglinska, E.; Danek, T. Air pollution seasons in urban moderate climate areas through big data analytics. *Sci. Rep.* **2024**, *14*, 3058. [CrossRef] [PubMed]
11. Sabal, M.; Danek, T.; Zaręba, M.; Węglińska, E. Explainable AI for effective management of urban heat sources. *Sci. Rep.* **2025**, *15*, 40616. [CrossRef] [PubMed]
12. Hůnová, I. Ambient Air Quality in the Czech Republic: Past and Present. *Atmosphere* **2020**, *11*, 214. [CrossRef]
13. Arnaut, F.; Đurđević, V.; Kolarski, A.; Srećković, V.A.; Jevremović, S. Improving Air Quality Data Reliability through Bi-Directional Univariate Imputation with the Random Forest Algorithm. *Sustainability* **2024**, *16*, 7629. [CrossRef]
14. Choi, S.-M.; Choi, H. Statistical Modeling for PM_{10}, $PM_{2.5}$ and PM_1 at Gangneung Affected by Local Meteorological Variables and PM_{10} and $PM_{2.5}$ at Beijing for Non- and Dust Periods. *Appl. Sci.* **2021**, *11*, 11958. [CrossRef]
15. Cisneros, R.; Schweizer, D.; Amiri, M.; Zarate-Gonzalez, G.; Gharibi, H. Long-Term Fine Particulate Matter ($PM_{2.5}$) Trends and Exposure Patterns in the San Joaquin Valley of California. *Atmosphere* **2025**, *16*, 721. [CrossRef]
16. Zaręba, M.; Cogiel, S.; Węglińska, E.; Danek, T. Evaluating Machine Learning Models for Air Quality Error Mapping in Kraków, Poland. *Misc. Geogr. Reg. Stud. Dev.* **2026**, *30*, 24–32. [CrossRef]
17. Zareba, M.; Danek, T. A novel methodology for Explainable Artificial Intelligence integrated with geostatistics for air pollution control and environmental management. *Ecol. Inform.* **2025**, *92*, 103450. [CrossRef]

18. Cogiel, S.; Zareba, M.; Danek, T.; Arnaut, F. Automating Air Pollution Map Analysis with Multi-Modal AI and Visual Context Engineering. *Atmosphere* **2026**, *17*, 2. [CrossRef]
19. Oliver, M.A.; Webster, R. *Basic Steps in Geostatistics: The Variogram and Kriging*; Springer: Berlin/Heidelberg, Germany, 2015. [CrossRef]
20. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
21. Hengl, T.; Nussbaum, M.; Wright, M.N.; Heuvelink, G.B.M.; Gräler, B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ* **2018**, *6*, e5518. [CrossRef]
22. Georganos, S.; Grippa, T.; Gadiaga, A.N.; Linard, C.; Lennert, M.; Vanhuysse, S.; Mboga, N.; Wolff, E.; Kalogirou, S. Geographical random forests: A spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto Int.* **2019**, *36*, 121–136. [CrossRef]
23. Talebi, H.; Peeters, L.J.M.; Otto, A.; Tolosana-Delgado, R. A Truly Spatial Random Forests Algorithm for Geoscience Data Analysis and Modelling. *Math. Geosci.* **2022**, *54*, 1–22. [CrossRef]
24. Morawiec, T.; Zaręba, M.; Danek, T.; Chuchro, M. Air Pollution Macro-Regions Identification Using Machine Learning and Spatio-Temporal Analysis. *PLoS ONE* **2026**, *21*, e0340191. [CrossRef] [PubMed]
25. Zareba, M. Assessing the Role of Energy Mix in Long-Term Air Pollution Trends: Initial Evidence from Poland. *Energies* **2025**, *18*, 1211. [CrossRef]
26. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. Available online: http://data.europa.eu/eli/dir/2008/50/oj (accessed on 1 January 2026).
27. Airly Pure. Available online: https://airly.org/en/solutions/airly-pure/ (accessed on 1 January 2026).
28. Vogt, M.; Schneider, P.; Castell, N.; Hamer, P. Assessment of Low-Cost Particulate Matter Sensor Systems against Optical and Gravimetric Methods in a Field Co-Location in Norway. *Atmosphere* **2021**, *12*, 961. [CrossRef]
29. Airly Air Quality Monitors Comply with MCERTS Performance Standards for Indicative Ambient Particulate Monitoring. Available online: https://airly.org/en/airly-air-quality-monitors-comply-with-mcerts-performance-standards-for-indicative-ambient-particulate-monitoring/ (accessed on 1 January 2026).
30. European Space Agency (ESA); Airbus. Copernicus DEM; Copernicus Data Space Ecosystem: 2022. Available online: https://dataspace.copernicus.eu/explore-data/data-collections/copernicus-contributing-missions/collections-description/COP-DEM (accessed on 26 November 2025). [CrossRef]
31. Zippenfenig, P. *Open-Meteo.com Weather API [Computer Software]*; Zenodo, CERN: Geneva, Swtizerland, 2023. [CrossRef]
32. Wright, M.N.; Ziegler, A. Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C^{++} and R. *J. Stat. Softw.* **2017**, *77*, 1–17. [CrossRef]
33. Strobl, C.; Boulesteix, A.; Kneib, T.; Augustin, T.; Zeileis, A. Conditional variable importance for random forests. *BMC Bioinf.* **2008**, *9*, 307. [CrossRef]
34. Wackernagel, H. Ordinary Kriging. In *Multivariate Geostatistics*; Springer: Berlin/Heidelberg, Germany, 1995. [CrossRef]
35. Sówka, I.; Badura, M.; Pawnuk, M.; Szymański, P.; Batog, P. The use of the GIS tools in the analysis of air quality on the selected University campus in Poland. *Arch. Environ. Prot.* **2020**, *46*, 100–106. [CrossRef]
36. Felici-Castell, S.; Fernandez-Vargas, E.; Segura-Garcia, J.; Perez-Solano, J.J.; Fayos-Jordan, R.; Lopez-Ballester, J. Accurate Estimation of Air Pollution in Outdoor Routes for Citizens and Decision Making. *Appl. Sci.* **2023**, *13*, 9930. [CrossRef]
37. Salsabilla, S.; Syaharni, A.F.; Chamidah, N. Prediction of PM2.5 in DKI Jakarta Using Ordinary Kriging Method. *Enthusiastic Int. J. Appl. Stat. Data Sci.* **2023**, *3*, 48–58. [CrossRef]
38. Shin, J.; Woo, J.; Choe, Y.; Min, G.; Kim, D.; Kim, D.; Lee, S.; Yang, W. Spatiotemporal Exposure Assessment of $PM_{2.5}$ Concentration Using a Sensor-Based Air Monitoring System. *Atmosphere* **2024**, *15*, 664. [CrossRef]
39. Węglińska, E.; Sabal, M.; Zareba, M.; Danek, T. Income, Heating Technologies and Behavioral Patterns as Drivers of Particulate Matter Emissions in the Kraków Metropolitan Area. *Energies* **2026**, *19*, 283. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.