

Environmental engineering Aplinkos inžinerija

A SENSITIVITY ANALYSIS OF DATA IMPUTATION IN DAILY PM10 PREDICTION

Mina Adel Shokry FAHIM , Jūratė SUŽIEDELYTĖ VIŠOCKIENĖ 

Vilnius Gediminas Technical University, Vilnius, Lithuania

- received 30 April 2026
- accepted 1 June 2026

Abstract. Reliable daily prediction of particulate matter (PM10) depends not only on model structure and predictor selection, but also on how missing data are treated during preprocessing. In air-quality applications, gaps are common in both ground-based monitoring records and satellite-derived predictors due to instrument downtime, quality screening, and cloud contamination. Despite this, data imputation is often treated as a purely technical step. The study presents a sensitivity analysis of data imputation strategies and their impact on daily PM10 prediction using a Random Forest (RF) framework over Lithuania for the period 2019–2024. Using a consistent spatiotemporal dataset integrating ground observations, meteorological variables, satellite-derived atmospheric composition, and temporal features, multiple gap-filling approaches were evaluated under identical model configurations and temporal data splits. The examined imputation methods include inverse distance weighting (IDW), Gaussian process regression (GPR), station mean and median, and Multivariate Imputation by Chained Equations (MICE). Model performance was assessed using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). Results indicate that spatially informed imputation IDW consistently achieved the highest reconstruction accuracy under moderate missingness, while GPR became competitive under extreme data loss. Using the selected IDW-completed dataset, the RF model achieved predictive performance (Test: $R^2 = 0.615$, RMSE = $6.37 \mu\text{g}/\text{m}^3$, MAE = $4.78 \mu\text{g}/\text{m}^3$).

Keywords: PM₁₀ prediction, data imputation, random forest, air quality modelling, spatiotemporal analysis.

 Corresponding author. E-mail: mina-adel-shokry.fahim@vilniustech.lt

1. Introduction

Air is fundamental to sustaining human life, and preserving its quality requires the adoption of effective techniques for monitoring, management, and control to support informed decision-making. However, air pollution poses significant risks to both human health and natural ecosystems (Bernacki & Scherer, 2025). Its sources originate from anthropogenic activities such as industrial emissions, vehicular traffic, and military operations, as well as from natural processes, including volcanic eruptions and forest fires (Bai et al., 2018). Consequently, the need for accurate measurement and robust modeling of air pollutants has increased substantially over recent years, driven by growing environmental concerns and public health demands, while numerous air quality sensors have been deployed across different regions to monitor atmospheric conditions. The resulting time-series data often contain missing values. These gaps may arise due to electricity, sensor malfunctions, or data corruption (Díaz-González et al., 2025). Consequently, the application of appropriate data imputation techniques is essential to estimate missing values with reasonable accuracy. Complete and continuous time-series datasets are

critical for enabling predictive models to effectively analyze patterns and produce reliable air quality forecasts. Despite the rapid development of air-quality forecasting systems, the treatment of missing observations remains inconsistently addressed in the literature. Many studies employ machine-learning algorithms while applying simple preprocessing solutions, often without examining how these choices influence the final predictive skill. For instance, in Petrosani, Romania, a study demonstrates that machine-learning approaches can effectively model PM_{2.5} dynamics in a university micro-environment by integrating IoT sensor data, conventional monitoring observations, and meteorological variables. Among the tested algorithms, the RF model achieved the strongest predictive performance ($R^2 = 0.82764$), indicating moderate but reliable explanatory power under real-world conditions. Overall, the research provides a practical and transferable framework for campus-scale air-quality management and targeted pollution mitigation strategies (Panaite et al., 2024), while in Malaysia, a national-scale study demonstrates the application of RF and Support Vector Regression models to estimate PM_{2.5} concentrations by integrating satellite-derived aerosol optical depth, ground-based pollutant measurements,

and meteorological parameters. Across diverse site types and seasonal conditions, the developed models achieved moderate predictive performance, with overall accuracies ranging from ($R^2 = 0.46$ to 0.76) (Zaman et al., 2021), models based on RF have repeatedly demonstrated strong capability in capturing nonlinear relationships between emissions, meteorology, and atmospheric chemistry, yet the uncertainty introduced during gap filling is rarely quantified in a systematic manner.

Recent research has begun exploring more sophisticated reconstruction strategies, including spatial interpolation, neighbor-based similarity, and probabilistic time-series approaches (Baruah et al., 2024). However, comparisons are frequently conducted under different datasets, feature sets, or validation schemes, making it difficult to isolate the real contribution of the imputation method itself. Moreover, relatively few investigations evaluate performance across multiple levels of data availability, even though operational monitoring networks commonly experience varying degrees of incompleteness from day to day, therefore, a clear methodological gap persists, the absence of a controlled and reproducible framework in which alternative gap-filling strategies are tested under identical modelling conditions before the forecasting stage.

The present study addresses this limitation by explicitly separating reconstruction accuracy from prediction accuracy and by quantifying how preprocessing uncertainty propagates into daily PM₁₀ forecasts. Through systematic masking experiments and consistent model configurations, the work provides show the comparison between imputation techniques can be made long before training begins, thereby offering practical guidance for the design of reliable air-quality prediction systems.

2. Materials and methods

The study was conducted in Lithuania and was designed to quantify how different data imputation strategies influence daily PM₁₀ prediction. The methodological workflow Figure 1 consists of three main components: multi-source data acquisition, methodology, and accuracy assessment, which introduction of missingness, systematic comparison of imputation techniques, and downstream RF prediction under identical model configurations.

2.1. Data acquisition

All datasets were aligned to a unified daily temporal resolution and spatially linked to individual monitoring stations. The integration process ensured consistency across heterogeneous data sources, including ground observations, meteorological variables, satellite atmospheric products, and engineered temporal predictors. Quality control procedures were applied prior to modelling to remove invalid or flagged records and to ensure that subsequent comparisons reflect methodological differences rather than inconsistencies in raw data.

A summary of all datasets is provided in Table 1.

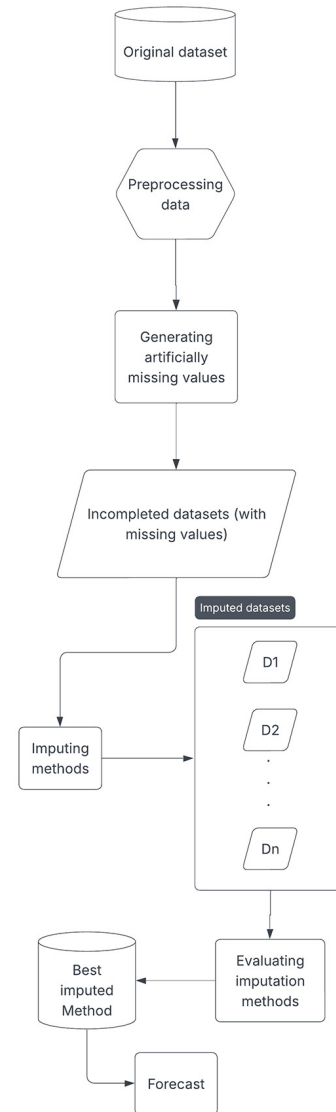


Figure 1. Workflow of the imputation sensitivity and forecasting framework (source: the authors)

Table 1. Dataset catalogue (source: the authors)

Data	Variables	Unit
Ground air quality data	PM ₁₀	µg/m ³
Meteorological data	Air temperature, feels-like temperature, wind speed, wind gust, relative humidity, cloud cover, sea-level pressure, precipitation	°C, m/s, %, hPa, mm
Satellite atmospheric data	NO ₂ , CO, SO ₂ , O ₃ , HCHO columns, Absorbing Aerosol Index (AAI)	mol/m ² , unitless
Station characteristics	Latitude, longitude, altitude	°, m
Temporal features	year, month, day, weekday, season, lag1, lag2, lag3, roll7	unitless, µg/m ³

2.1.1. Ground air quality observations

Hourly PM₁₀ concentrations constitute the primary target variable of this study. Measurements were obtained from European Environment Agency (n.d.) air quality reporting, which provides standardized observations from certified monitoring stations, raw records were subjected to quality control procedures. Invalid or unverified entries were excluded based on the provided data flags. After cleaning, the time series were aggregated to daily averages for each station, forming continuous station-based PM₁₀ records across the study period. Because temporary data gaps are unavoidable in operational monitoring networks, several imputation strategies were later evaluated and compared within the modelling framework.

2.1.2. Meteorological data

Meteorological conditions modulate pollutant dispersion, transport, and accumulation (Chhabra, 2024; Johnson, 2018; Karami et al., 2025). Hourly observations of air temperature, feels-like temperature, wind speed, wind gust, cloud cover, sea-level pressure, relative humidity, and precipitation were retrieved from the national Meteo service (Meteo.Lt – Lithuanian Hydrometeorological Service, n.d.). The variables were aggregated to daily statistics consistent interpolated using IDW, with regard to each air quality station location with the temporal resolution of the air quality dataset.

2.1.3. Satellite atmospheric data

Satellite-based atmospheric composition products were derived from Sentinel-5P TROPOMI observations (Earth Engine Data Catalog, n.d.; TROPOMI: TROPOspheric Monitoring Instrument, n.d.). The dataset includes daily indicators of tropospheric NO₂, CO, SO₂, O₃, formaldehyde (HCHO), and the Absorbing Aerosol Index (AAI), which was assessed using GEE. The remaining observations were spatially associated with each monitoring station location. Due to orbital coverage limitations and cloud interference, residual gaps may occur; these were handled during the imputation assessment phase using statistical and machine-learning approaches.

2.1.4. Station geographic characteristics

Each monitoring location was described by its geographic coordinates (latitude and longitude) and altitude obtained from EEA metadata; it was essential for representing the spatial relationships between.

2.1.5. Temporal predictors

To capture systematic temporal behavior in air pollution, calendar-based variables were derived from the daily index, including year, month, day of month, weekday, and season. Short-term persistence effects were represented using lagged PM₁₀ concentrations for the previous one to three days (lag1–lag3). In addition, a 7-day rolling mean (roll7) was computed to characterize recent accumula-

tion patterns. Such temporal descriptors are widely used in air-quality modelling to improve the representation of delayed atmospheric responses and emission continuity (Karapetyan & Galoyan, 2025; Zhao et al., 2021).

2.2. Methodology

A structured sensitivity-analysis framework was established before the training of the RF algorithm. A fully validated dataset of daily PM₁₀ concentrations was first prepared and treated as the reference ground truth. Controlled data gaps were then introduced by randomly masking predefined proportions of records, ranging from minor to severe data loss. Because the original measurements remained known, the ability of each reconstruction approach to recover the removed values could be evaluated objectively and under reproducible conditions, from complete reference multiple incomplete datasets were derived to represent alternative availability scenarios. Within each dataset, a specific share of observations was randomly replaced with missing entries (Hua et al., 2024), while still preserving access to the authentic value for benchmarking. Accuracy assessment was performed exclusively at the masked positions, preventing biased estimates.

A diverse set of methods of gap-filling techniques was examined in order to contrast assumptions of temporal persistence, spatial interaction, and multivariate dependency. Station-based statistical reconstructions relied on historical summaries such as mean or median values. Spatial structure was captured through IDW, where nearby monitors exert stronger influence than distant ones (Deligiorgi et al., 2011; Wong et al., 2004). Methods exploiting similarity among multiple predictors included k-nearest neighbour imputation, while iterative reconstruction through chained regression equations was implemented via MICE (Azifah Mohd Pauzi et al., 2021; Chen et al., 2022). Finally, GPR modelled the evolution of concentrations as a probabilistic model.

2.3. Accuracy indicators

The gap-filling strategy does not introduce systematic bias into the modelling framework; structured accuracy assessment was conducted. The evaluation aimed to quantify how closely each imputation technique can reconstruct known observations under controlled missing-data scenarios. Missingness was introduced into the daily PM₁₀ records by randomly masking a predefined proportion of observations. Several levels of data removal were examined in order to simulate a wide range of operational data availability conditions, from limited gaps to severe data loss. The original unaltered measurements served as the reference (ground truth) against which reconstructed values were compared.

The assessment included simple statistical, temporal, spatial, and machine-learning-based methods. These comprised persistence-based approaches, such as station summary statistics, such as mean or median, multivariate

similarity methods, including k-nearest neighbour imputation, iterative multivariate MICE, GPR applied along the temporal dimension of individual stations, and spatial interpolation via IDW. Reconstruction skill was quantified using three complementary indicators, MAE which measures the average magnitude of deviations without emphasizing extreme values, RMSE, which penalizes large deviations more strongly (Willmott & Matsuura, 2005), R^2 , which evaluates the proportion of variance in the reference observations explained by the imputed estimates as explained in Equations (1–3) (Fahim & Visockienė, 2026):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\text{PM}_{10}^{\text{obs}}(i) - \text{PM}_{10}^{\text{pred}}(i) \right)^2}; \quad (1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n \left(\text{PM}_{10}^{\text{obs}}(i) - \text{PM}_{10}^{\text{pred}}(i) \right)^2}{\sum_{i=1}^n \left(\text{PM}_{10}^{\text{obs}}(i) - \overline{\text{PM}_{10}^{\text{obs}}} \right)^2}; \quad (2)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n \left| \text{PM}_{10}^{\text{obs}}(i) - \text{PM}_{10}^{\text{pred}}(i) \right|, \quad (3)$$

where n is the number of observations, and $\text{PM}_{10}^{\text{obs}}(i)$ and $\text{PM}_{10}^{\text{pred}}(i)$ are the observed and predicted values, respectively, and $\overline{\text{PM}_{10}^{\text{obs}}}$ is the mean of the observations.

The best-performing strategy was subsequently adopted to reconstruct the full dataset before training the RF prediction model. After gap filling, the completed dataset was merged with meteorological, satellite, spatial, and temporal predictors. RF was then trained using consistent hyperparameters and temporal splits.

3. Results

3.1. Accuracy under controlled missingness

The controlled masking experiments reveal clear and systematic differences in performance across imputation methods and missingness levels (10–90%). As shown in Figures 2–4, accuracy progressively deteriorates as the proportion of missing daily PM₁₀ observations increases. For the best-performing approaches, RMSE rises from approximately 7.51 $\mu\text{g}/\text{m}^3$ for IDW at 10% missingness to around 11.09 $\mu\text{g}/\text{m}^3$ at 90% missingness Figure 4. A similar decline is observed for R^2 , which decreases from 0.628 at 10% to 0.158 at 90% in Figure 3, indicating a reduced ability to reproduce variance as data availability becomes more limited. MAE follows a consistent upward trend in Figure 2, confirming that absolute reconstruction deviations increase under more severe data loss.

Across most scenarios, spatial interpolation using IDW demonstrates the strongest overall performance. IDW achieves the lowest RMSE at 10% (7.51 $\mu\text{g}/\text{m}^3$, $R^2 = 0.628$), 20% (7.55, 0.616), 30% (7.77, 0.605), 40% (7.95, 0.575), 50% (8.12, 0.565), 60% (8.38, 0.498), 70% (9.11, 0.435), and 80% (9.47, 0.378) missingness. At 90%, the lowest RMSE is obtained by GPR (11.04 $\mu\text{g}/\text{m}^3$, $R^2 = 0.166$), while IDW remains very close (11.09 $\mu\text{g}/\text{m}^3$, $R^2 = 0.158$), the slight crossover at extreme missingness suggests that when spatial support weakens due to simultaneous gaps across

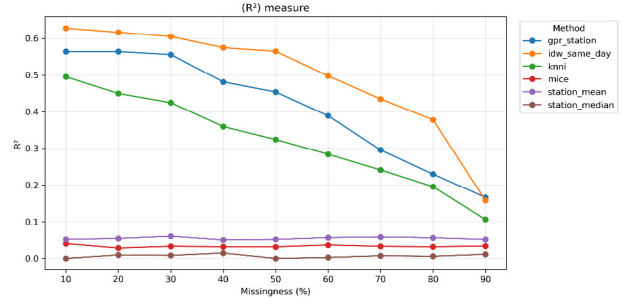


Figure 2. R^2 of daily PM₁₀ reconstruction across missingness levels (source: the authors)

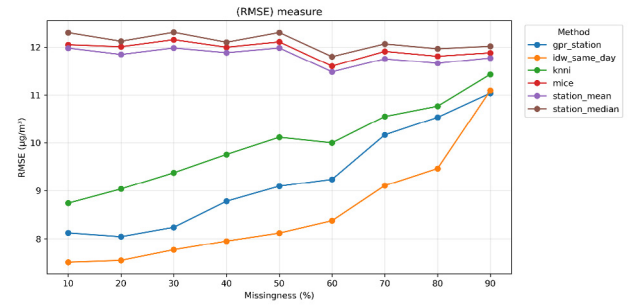


Figure 3. RMSE of daily PM₁₀ reconstruction across missingness levels (source: the authors)

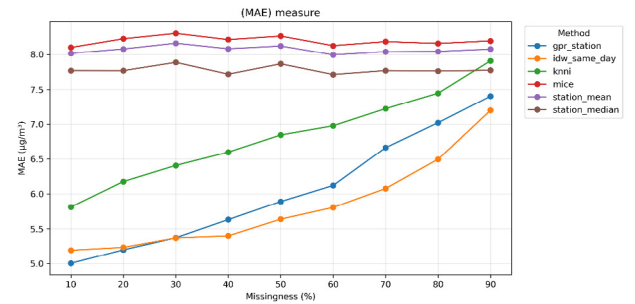


Figure 4. MAE of daily PM₁₀ reconstruction across missingness levels (source: the authors)

stations, temporally adaptive models such as GPR become comparatively competitive.

In contrast, simpler statistical approaches consistently show inferior reconstruction capability. Station mean and station median imputation produce relatively stable but high errors RMSE generally between ~11.5 and 12.3 $\mu\text{g}/\text{m}^3$ and very low explained variance R^2 typically below 0.06, reflecting limited ability to capture daily variability. The KNNI method provides intermediate performance, with RMSE increasing from 8.74 $\mu\text{g}/\text{m}^3$ R^2 of 0.495 at 10% missingness to 11.43 $\mu\text{g}/\text{m}^3$ with R^2 of 0.106 at 90%. The MICE approach remains close to the simple-statistics group in this configuration, with RMSE predominantly around 11.6–12.2 $\mu\text{g}/\text{m}^3$ and R^2 near 0.03–0.04 across missingness levels. Overall, Figures 2–4 illustrate a consistent ranking pattern: spatially informed methods perform best under moderate data loss, temporally adaptive modelling gains relevance at extreme missingness, and deterministic summary-based methods show limited robustness as data availability declines.

3.2. RF Prediction to accuracy

To quantify how preprocessing decisions propagate into forecasting performance, the RF prediction stage was applied to the original daily PM₁₀ dataset from the 15 monitoring stations. A complete daily calendar was constructed for the period 1 January 2019 to 31 December 2024, resulting in 2,192 days per station and a total of 32,880 station-day records. After rebuilding the full time series, 1,162 values ($\approx 3.5\%$) were missing and subsequently reconstructed using the IDW approach, selected earlier as the most robust method under controlled missingness scenarios, to ensure a consistent and spatially coherent completion strategy prior to model training, while keeping the modelling setup identical across experiments, after completing the dataset, the RF model was trained using the same predictor set and the same fixed hyperparameter configuration, with a temporal split into training, validation, and testing subsets with Train = 23,010; Validation = 6,570; Test = 3,300 samples. under identical model settings. Performance statistics for the RF model showing validation and test subsets. On the validation data, the model achieved MAE = 4.68 $\mu\text{g}/\text{m}^3$, RMSE = 6.90 $\mu\text{g}/\text{m}^3$, and $R^2 = 0.570$, while on the independent test data it improved slightly to MAE = 4.78 $\mu\text{g}/\text{m}^3$, RMSE = 6.37 $\mu\text{g}/\text{m}^3$, and $R^2 = 0.615$.

The agreement between observed and predicted concentrations is further illustrated in Figure 5, which compares RF predictions against the corresponding observations for the test period. The scatter distribution concentrates around the 1:1 line, indicating that the model captures the overall variability of daily PM₁₀, although dispersion increases at higher concentrations, consistent with the known difficulty of reproducing episodic pollution peaks. Together, with Figure 3 demonstrate that the selected IDW-based completion strategy provides a stable basis for RF prediction, and that imputation choices remain a consequential component of the overall modelling pipeline, influencing both reconstructed inputs and downstream forecast accuracy.

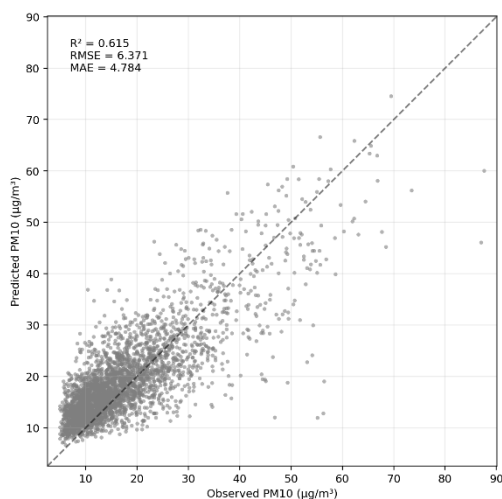


Figure 5. Observed vs. predicted daily PM₁₀ concentrations using RF (test dataset) (source: the authors)

4. Conclusions

The study compared between data imputation strategies influence daily PM₁₀ prediction within a controlled RF framework over Lithuania for the period 2019–2024. By explicitly separating reconstruction accuracy from forecasting performance, the analysis isolated the contribution of preprocessing decisions to overall model skill, the controlled masking experiments demonstrated a clear and consistent decline in reconstruction accuracy as missingness increased from 10% to 90%. Spatially informed interpolation using IDW consistently outperformed other techniques under low to moderate data loss, achieving the lowest RMSE and highest R^2 across most scenarios. Under extreme missingness (90%), GPR became competitive, suggesting that temporally adaptive probabilistic approaches gain relevance when spatial support weakens. In contrast, simple deterministic methods such as station mean and median showed limited ability to reproduce daily variability, resulting in low explained variance and persistently higher errors.

When propagated into the forecasting stage, the choice of imputation method demonstrated measurable effects on predictive performance. Using the IDW-completed dataset, the RF model achieved moderate but stable generalization capacity of Test dataset $R^2 = 0.615$, RMSE = 6.37 $\mu\text{g}/\text{m}^3$, MAE = 4.78 $\mu\text{g}/\text{m}^3$. the findings highlight that data imputation is not merely a technical preprocessing step but a methodological decision that directly affects downstream predictive skill. Spatially aware gap-filling approaches provide robust performance under realistic operational conditions, while the sensitivity framework introduced in this study offers a reproducible template for evaluating preprocessing uncertainty in air-quality modelling. Future research may extend this framework to other pollutants, alternative machine-learning architectures, and spatial generalization scenarios, thereby strengthening the reliability and transparency of environmental forecasting systems.

Author contributions

M. A. S. F. and J. S. V. contributed to the conceptualization of the study. M. A. S. F. developed the methodology and performed the data curation and visualization. Formal analysis was carried out by both M. A. S. F. and J. S. V., while the investigation was conducted by J. S. V. Resources were provided by M. A. S. F. and J. S. V. M. A. S. F. prepared the original draft of the manuscript, and both authors participated in reviewing and editing the final version. J. S. V. supervised the study.

Disclosure statement

The authors declare that they have no competing financial, professional, or personal interests with any other parties.

References

- Azifah Mohd Pauzi, N., Bee Wah, Y., Mohd Deni, S., & Khatijah Nor Abdul Rahim, S. (2021). Comparison of single and MICE imputation methods for missing values: A simulation study. *Pertanika Journal of Science & Technology*, 29(2), 979–998. <https://doi.org/10.47836/pjst.29.2.15>
- Bai, L., Wang, J., Ma, X., & Lu, H. (2018). Air pollution forecasts: An overview. *International Journal of Environmental Research and Public Health*, 15(4), Article 780. <https://doi.org/10.3390/ijerph15040780>
- Baruah, A., Bousiotis, D., Damayanti, S., Bigi, A., Ghermandi, G., Ghaffarpasand, O., Harrison, R. M., & Pope, F. D. (2024). A novel spatiotemporal prediction approach to fill air pollution data gaps using mobile sensors, machine learning and citizen science techniques. *Npj Climate and Atmospheric Science*, 7, Article 310. <https://doi.org/10.1038/s41612-024-00859-z>
- Bernacki, J., & Scherer, R. (2025). A comprehensive review of data-driven techniques for air pollution concentration forecasting. *Sensors*, 25(19), Article 6044. <https://doi.org/10.3390/s25196044>
- Chen, M., Zhu, H., Chen, Y., & Wang, Y. (2022). A novel missing data imputation approach for time series air quality data based on logistic regression. *Atmosphere*, 13(7), Article 1044. <https://doi.org/10.3390/atmos13071044>
- Chhabra, P. (2024). *The effect of different meteorological factors on the concentrations of air pollutants*. EarthArXiv. <https://doi.org/10.31223/X5F98K>
- Deligiorgi, D., Philippopoulos, K., Deligiorgi, D., & Philippopoulos, K. (2011). Spatial interpolation methodologies in urban air pollution modeling: Application for the eater Area of Metropolitan Athens, Greece. In F. Nejadkoorki, *Advanced air pollution*. IntechOpen. <https://doi.org/10.5772/17734>
- Díaz-González, L., Trujillo-Urbe, I., Pérez-Sansalvador, J. C., & Lakouari, N. (2025). Handling missing air quality data using bidirectional recurrent imputation for time series and random forest: A case study in Mexico City. *AI 2025*, 6(9), Article 208. <https://doi.org/10.3390/ai6090208>
- Earth Engine Data Catalog. (n.d.). *Datasets tagged tropomi in Earth Engine*. Retrieved January 15, 2026, from <https://developers.google.com/earth-engine/datasets/tags/tropomi>
- European Environment Agency. (n.d.). Retrieved February 1, 2026, from <https://www.eea.europa.eu/en>
- Fahim, M. A. S., & Visockienė, J. S. (2026). An assessment of the multi-input spatiotemporal RF-XGBoost hybrid framework for PM10 estimation in Lithuania. *Sustainability* 2026, 18(4), Article 2022. <https://doi.org/10.3390/su18042022>
- Hua, V., Nguyen, T., Dao, M. S., Nguyen, H. D., & Nguyen, B. T. (2024). The impact of data imputation on air quality prediction problem. *PLOS One*, 19(9), Article e0306303. <https://doi.org/10.1371/journal.pone.0306303>
- Johnson, H. (2018). *PM2.5 pollution and temperature inversions: A case study in St. Louis, MO*. Iowa State University Digital Repository. <https://doi.org/10.31274/cc-20240624-1133>
- Karami, S., Ghassabi, Z., Khoddam, N., & Habibi, M. (2025). Investigating meteorological factors influencing pollutant concentrations and copernicus atmosphere monitoring service (CAMS) model forecasts in the Tehran Metropolis. *Atmosphere*, 16(3), Article 264. <https://doi.org/10.3390/atmos16030264>
- Karapetyan, I., & Galoyan, E. (2025). *Air pollution dynamics in Yerevan* (DS 299 Capstone in data science project report). https://cse.aau.am/wp-content/uploads/2025/06/Air_pollution-1.pdf
- Meteo.lt – Lithuanian Hydrometeorological Service. (n.d.). Retrieved February 2, 2026, from <https://www.meteo.lt/>
- Panaite, F. A., Rus, C., Leba, M., Ionica, A. C., & Windisch, M. (2024). Enhancing air-quality predictions on university campuses: A machine-learning approach to PM2.5 forecasting at the University of Petroșani. *Sustainability*, 16(17), Article 7854. <https://doi.org/10.3390/su16177854>
- TROPOMI: TROPospheric Monitoring Instrument. (n.d.). Retrieved February 3, 2026, from <https://www.tropomi.eu/>
- Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1), 79–82. <https://doi.org/10.3354/cr030079>
- Wong, D. W., Yuan, L., & Perlin, S. A. (2004). Comparison of spatial interpolation methods for the estimation of air quality data. *Journal of Exposure Science & Environmental Epidemiology*, 14, 404–415. <https://doi.org/10.1038/sj.jea.7500338>
- Zaman, N. A. F. K., Kanniah, K. D., Kaskaoutis, D. G., & Latif, M. T. (2021). Evaluation of machine learning models for estimating PM2.5 concentrations across Malaysia. *Applied Sciences*, 11(16), Article 7326. <https://doi.org/10.3390/app11167326>
- Zhao, F., Liang, Z., Zhang, Q., Seng, D., & Chen, X. (2021). Research on PM2.5 spatiotemporal forecasting model based on LSTM neural network. *Computational Intelligence and Neuroscience*, 2021(1), Article 1616806. <https://doi.org/10.1155/2021/1616806>

TRŪKSTAMŲ DUOMENŲ UŽPILDYMO ĮTAKO ANALIZĖ PROGNOZUOJANT PAROS PM10 KONCENTRACIJAS

Santrauka. Patikima paros kietųjų dalelių (PM₁₀) koncentracijos prognozė priklauso ne tik nuo modelio struktūros ir prognozuojamųjų kintamųjų parinkimo, bet ir nuo to, kaip duomenų paruošimo etape tvarkomi trūkstami duomenys. Oro kokybės tyrimuose duomenų spragos dažnai susidaro tiek antžeminių matavimų įrašuose, tiek iš palydovinių duomenų išvestuose kintamuosiuose dėl prietaisų veikimo sutrikimų, kokybės filtravimo procedūrų ar debesuotumo poveikio. Nepaisant to, trūkstamų duomenų užpildymas dažnai laikomas tik techniniu duomenų apdorojimo žingsniu. Šiame tyrime pateikiama trūkstamų duomenų užpildymo metodų jautrumo analizė ir jų įtaka paros PM₁₀ koncentracijos prognozėms vertinti, taikant atsitiktinių miškų (angl. *Random Forest*, RF) metodą Lietuvos teritorijoje 2019–2024 m. laikotarpiu. Naudojant nuoseklų erdvinį ir laiko duomenų rinkinį, integruojant antžeminius stebėjimus, meteorologinius kintamuosius, iš palydovinių duomenų gautus atmosferos sudėties rodiklius bei laiko požymius, buvo vertinami keli trūkstamų duomenų užpildymo metodai, taikant vienodas modelio konfigūracijas ir identišką duomenų skaidą laiko atžvilgiu. Analizuoti metodai apėmė atvirkštinio atstumo svorių metodą (IDW), Gauso procesų regresiją (GPR), stoties vidurkio ir medianos metodus bei daugiamatės grandinės lygtis (MICE). Modelių tikslumas vertintas pagal determinacijos koeficientą (R²), vidutinę kvadratinę paklaidą (RMSE) ir vidutinę absoliučiąją paklaidą (MAE). Tyrimo rezultatai parodė, kad, taikant IDW metodą, kai yra vidutinis trūkstamas duomenų kiekis, gaunamas didžiausias atkūrimo tikslumas, o esant itin dideliems duomenų praradimams gerų rezultatų pasiekta GPR metodu. Taikant IDW metodą su papildytu duomenų rinkiniu ir RF modeliu pasiektos šios prognozavimo tikslumo rodiklių imtys: R² = 0,615, RMSE = 6,37 μg/m³ ir MAE = 4,78 μg/m³.

Reikšminiai žodžiai: PM₁₀ prognozavimas, duomenų priskyrimas, atsitiktinių miškų metodas, oro kokybės modeliavimas, erdvinė ir laiko analizė.