Multiscale Gap-Filling in Particulate Matter Time Series Using Machine Learning and Dynamic Modeling

Ruslan Safarov
Lecturer Researcher, L. N. Gumilyov Eurasian National University, Astana, Kazakhstan

Yuriy Nossenko
Margulan University, Pavlodar, Kazakhstan

Eldar Kopishev
L. N. Gumilyov Eurasian National University, Astana, Kazakhstan

Zhuldyz Bexeitova
L. N. Gumilyov Eurasian National University, Astana, Kazakhstan

Abstract:

Accurate and continuous monitoring of fine particulate matter (PM2.5) is essential for environmental assessment and public health protection. However, real-world air quality datasets often suffer from significant data gaps due to sensor failures, harsh climate conditions, and infrastructure limitations. This study presents a comprehensive evaluation of 46 gap-filling methods for PM2.5 time series, ranging from basic statistical imputations to advanced machine learning and dynamic models. Using high-resolution data from Pavlodar, Kazakhstan, we systematically assessed model performance across five representative gap lengths (5–72 hours) using synthetic and real-world missing patterns. Tree-based sequence-to-sequence models, particularly XGBoost with bidirectional architecture, demonstrated the best accuracy for medium-length gaps (e.g., 5.23±0.29 μg/m³ MAE for 12-hour gaps), outperforming traditional methods by over 60%. The added value of multivariate inputs – such as wind speed, direction, temperature, and humidity – increased with gap length, improving performance by up to 18% for 72-hour gaps. We further developed novel dynamic models capable of automatically adapting to variable gap durations, enabling reliable imputation for gaps up to 191 hours without retraining. The reconstructed time series revealed that over 60% of monitored hours exceeded WHO PM2.5 daily limits, highlighting persistent air quality challenges. Our results offer practical guidance for environmental data managers and demonstrate the importance of scalable, adaptive imputation strategies in modern air quality monitoring systems.

Keywords:

pm2.5, gap-filling, time series, machine learning, dynamic modeling.