Review Article *Open Access*

Journal of Pioneering Artificial Intelligence Research

ISSN: 3069-0846

DOI: doi.org/10.63721/25JPAIR0113

Analysis and Comparison of Missing Values Imputation Methods for Atmospheric Pollution Data

Jakub Jasin´ski

Faculty of Electrical Engineering Warsaw University of Technology Warsaw, Poland

Citation: Jakub Jasin´ski(2025) Analysis and Comparison of Missing Values Imputation Methods for Atmospheric Pollution Data J of Pion Artf Research 1(3), 1-6. WMJ/JPAIR-113

Abstract

Missing values frequently occur in real-world time series datasets, significantly affecting the precision and reliability of data analysis and machine learning models. This research project aims to explore the types of missing data occurrences and examine various imputation methods. The approaches considered will range from simple statistical techniques to more complex methods such as regression models, neural networks, and LSTM models. The effectiveness of these imputation techniques will be assessed using atmospheric pollution data, with a particular focus on PM10 and PM2.5 levels. Each method's performance will be evaluated based on accuracy, consistency, and its impact on subsequent predictive models. The findings indicate that LSTM models are the most effective, while regression and MLP models, though less accurate, offer faster alternatives. Conversely, mean imputation results in the highest error values.

***Corresponding author:** Jakub Jasin´ski, Faculty of Electrical Engineering Warsaw University of Technology Warsaw, Poland.

Submitted: 19.09.2025 **Accepted:** 27.09.2025 **Published:** 18.10.2025

Keywords: Missing Data, Data Imputation, Time Series, Regression, Lstm, Atmospheric Pollution, PM10, PM2.5

Introduction

Air pollution is widely recognized as one of the leading threats to both environmental stability and public health today. It stems from various human activities — industrial processes, transportation, agriculture —and natural events like wildfires and dust storms. Regular exposure to polluted air has been shown to cause various health problems, especially affecting the lungs and heart. Fine particles, known as PM2.5, are particularly dangerous because they are small enough to enter deep into the lungs and even reach the bloodstream. This can trigger inflammation throughout the

body and worsen conditions such as asthma, chronic bronchitis, or cardiovascular diseases [1]. In recent years, research has also suggested a link between air pollution and damage to the brain, potentially contributing to cognitive decline and developmental disorders in children [2].

Beyond its health effects, air pollution also harms the environment. It plays a role in acid rain, reduces crop yields, and accelerates climate change. For these reasons, gaining a deeper understanding of how pollution works and how it affects us is key to creating solutions that protect people and the planet alike.

The Polish air quality index is one of the most important indicators for determining the level of air quality in Poland. It is calculated based on 1-hour results of measurements of the following air concentrations: sulfur dioxide (SO2), nitrogen dioxide (NO2), particulate matter (PM10), particulate matter (PM2.5), and ozone (O3). The map presented in Figure1 displays air quality levels across Poland using a color-coded index ranging from green (very good) to dark red (very bad). The highest concentrations of poor air quality (red and orange markers) are observed in the southern and southwestern regions, particularly near Wroclaw and the Czech border.

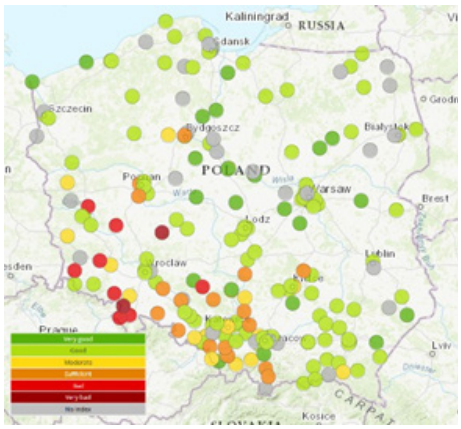

Figure 1: Polish Air Quality Index

Imputing missing values in air pollution data is essential to ensure accurate monitoring of environmental conditions and detection of harmful pollution levels. Incomplete data can lead to incorrect health risk assessments and hinder timely public health responses. Predictive models for air quality and climate heavily depend on consistent data, and missing values can reduce their accuracy and reliability. Additionally, com- plete datasets support better policy-making and more effective strategies for pollution control and public safety.

Literature Review

Missing Values in Time Series

Missing values in time series refer to absent or unrecorded observations at specific time points in a temporal dataset, which can disrupt the continuity and integrity of time- dependent analysis. These gaps may arise due to sensor failures, transmission errors, or irregular data collection and pose significant challenges for forecasting, modelling, and anomaly detection [3].

Type of Missing Data

We can distinguish three types of missing data [4].

- **Missing Completely at Random (MCAR)** — Missing values occur independently of both observed and unobserved data. In this case, the probability of data being missing is the same for all observations.Example: A random sensor logging error causes PM10 values to be missed every 1000th reading, regardless of environmental conditions or pollution levels.
- **Missing at Random (MAR)** — The missingness is related only to observed data. This means other variables in the dataset can explain the probability of missing values.Example: PM2.5 measurements are frequently missing during rainy weather, which is logged in the dataset. Thus, the missingness can be modelled using weather conditions.
- **Missing Not at Random (MNAR)** — The probability of missingness depends on the unobserved data, making it more challenging to model or impute.Example: NO2 concentrations become extremely high during industrial incidents, and the sensors tend to malfunction precisely at these critical levels. The miss- ing data depends on the unmeasured extreme values themselves.

Dealing with Missing Values

Handling missing values is a critical step in data prepro- cessing, as it can significantly influence the outcome of any analysis or model [3]. Below

- **Ignoring** - One of the simplest approaches is

ignoring missing values during analysis. This method involves proceeding with computations without accounting for the missing entries. However, we must be sure that omitting these values will not cause the models to malfunction.

- **Deletion** - This method involves removing either the variables (columns) or observations (rows) that contain missing values. While it ensures a clean dataset, deletion can result in significant information loss and reduced statistical power, particularly if missingness is widespread.
- **Imputation** - It refers to filling in missing values with estimated ones based on available data. This approach retains the dataset's structure and size. It can range from basic statistical methods to advanced techniques involving regression models, machine learning, or deep learning algorithms.

Methodology

Mean Imputation

In a mean substitution, the mean value of a variable is used in place of the missing data value for that same variable. The theoretical background of the mean substitution is that the mean is a reasonable estimate for a randomly selected observation from a normal distribution. However, with missing values that are not strictly random, especially in the presence of great inequality in the number of missing values for the different variables, the mean substitution method may lead to inconsistent bias[5]

Regression

Regression imputation using an iterative imputer with Bayesian Ridge regression estimates missing values by mod- elling each variable as a linear function of the others in multiple rounds. Bayesian Ridge adds regularization through priors, improving stability and handling multicollinearity. This iterative approach refines estimates with each pass, but it may still introduce bias or underestimate variability, especially if data relationships are nonlinear or the model assumptions are violated [6].

Neural Networks

Using a Multi-Layer Perceptron (MLP) for data imputa- tion involves framing the problem as a supervised learning task, where the MLP learns to predict missing values based on patterns in available historical data. In time series, a common approach is to use a sliding window to transform sequences of past observations into input features for the model [7]. The MLP, being a feedforward neural network, captures complex nonlinear relationships in the data but does not inherently model temporal dependencies, which can be partially addressed through input engineering. Although MLPs are simpler than recurrent models like LSTM, they are efficient and effective for datasets with short-term dependencies and relatively low missingness.

LSTM Recurrent Deep Network

LSTM networks introduce a memory cell and a set of gating mechanisms—input, output, and forget gates—that regulate the flow of information. These gates allow the network to retain or discard information over time, making LSTM models particularly effective for tasks involving time-series forecast- ing, natural language processing, and signal classification [8]. LSTM networks are widely used for imputing missing val- ues in time series by transforming the data into lagged input- output pairs, enabling the model to predict missing values based on temporal context. They are particularly effective in capturing both short- and long-term dependencies and outperform traditional methods under moderate to high miss- ingness. However, LSTMs require careful tuning and more computational resources and struggle with input sequences that contain missing values unless preprocessed [9].

Dataset

The GIOS Air Quality Archive [10] provides access to a comprehensive database of air pollution measurements collected across Poland by the State Environmental Monitoring network. This includes two key pollutants of health and environmental concern:

- PM10 (Particulate Matter $\leq$ 10 micrometers): Coarse particles that can penetrate the respiratory system, con- tributing to respiratory issues and cardiovascular diseases.
- PM2.5 (Particulate Matter $\leq$ 2.5 micrometers): Fine particles capable of reaching the lungs and bloodstream, strongly linked to heart and lung conditions.

These data are included as time series measured every hour at about 170 stations (depending on the year) in 2000-2023. An analysis was carried out to realize how much data is missing in the data below, resulting in the histograms presented in Figures 2 and 3.

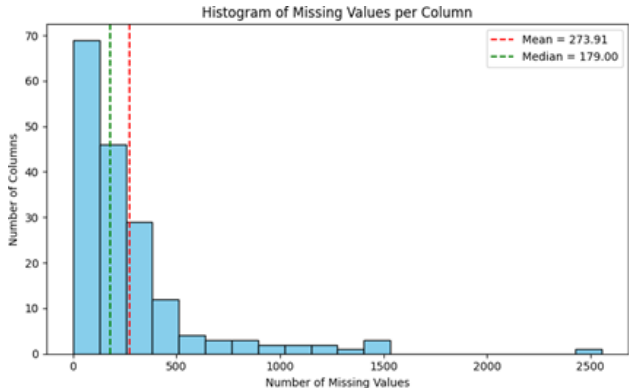

Figure 2: Missing Values Count in PM10 data of year 2023

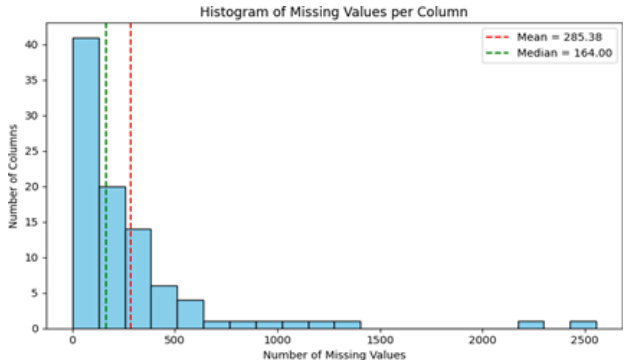

Figure 3: Missing Values Count in PM2.5 data of year 2023

As seen in both cases, most of the data contains less than 250 missing data. The average number of missing data is about 280 in both pollution cases, which is about 3% of the year. However, there are also cases of stations where the number of missing data exceeds 1460, equivalent to 2 months of no measurements and may significantly impact subsequent data analyses.

Another critical aspect of missing data analysis is the time intervals in which data is missing. It is much easier to predict a single missing value than a longer interval because of the possibility of taking information from the context of the surrounding measurements.

Table 1: Comparison of Maximum Missing Interval Statistics (Pm10 & Pm2.5)

Pollution	PM10	PM2.5
Max Interval (hrs)	2426.00	2426.00
Min Interval (hrs)	1.00	2.00
Mean Interval (hrs)	140.36	189.36
Median Interval (hrs)	64.00	74.50
Count of Stations	176.00	92.00

As shown in Table I both datasets show significant missing intervals, with some gaps lasting over 100 days, which can severely impact imputation accuracy. Long and irregular gaps challenge standard methods like mean imputation or iterative imputation significantly when multiple stations are affected simultaneously.

Results

To check the accuracy of the previously mentioned methods, randomly selected values from the available time series will be marked as artificially missing when using the imputation method. Then, the new values will be compared with metrics: SMAPE, MAE, and RMSE [11].

Tables II–IV present the performance of four imputation methods: Mean, Regression, MLP, and LSTM, applied to PM10 and PM2.5 datasets under varying levels of artificially induced missingness: 0.5%, 3%, and 20%. These levels correspond to typical missing durations in air quality datasets, with 0.5% (2 days) representing near-median gap lengths, 3% (11 days) approximating the average number of missing observations, and 20% (2 months) capturing the most extended missing intervals.

Model	PM10			PM2.5		
	SMAPE (%)	MAE (µg/m3)	RMSE (µg/m3)	SMAPE (%)	MAE (µg/m3)	RMSE (µg/m3)
Mean	31.93	5.52	9.12	32.1	4.5	7.24
Reg	20.81	3.05	4.78	15.77	2.37	5.63
MLP	19.07	2.78	3.75	16.31	2.15	4.2
LSTM	9.88	2.09	2.83	9.06	1.95	2.6

At minimal missingness, the LSTM model yields the most accurate predictions across all metrics for both pollutants. This demonstrates the model's capability to effectively utilize recent temporal context for short-gap recovery, outperforming statistical approaches.

Table 3: Performance Metrics for Pm10 and Pm2.5 and 3% Missing Data

Model	PM10			PM2.5		
	SMAPE (%)	MAE (µg/m3)	RMSE (µg/m3)	SMAPE (%)	MAE (µg/m3)	RMSE (µg/m3)
Mean	32.24	5.56	9.14	34.77	5.67	9.74
Reg	21.35	3.33	5.82	15.37	2.29	4.79
MLP	21.57	3.1	4.68	15.62	2.5	4.83
LSTM	11.76	2.56	3.38	11.34	2.42	3.25

When the missing rate increases to a level of dataset average loss, LSTM continues to show superior accuracy, particularly in SMAPE and RMSE. At the same time, MLP also performs competitively, highlighting the benefits of non-linear learning in moderate data loss scenarios.

Table 4: Performance Metrics for Pm10 and Pm2.5 with 20% Missing Data

Model	PM10			PM2.5		
	SMAPE (%)	MAE (µg/m3)	RMSE (µg/m3)	SMAPE (%)	MAE (µg/m3)	RMSE (µg/m3)
Mean	33.31	5.65	9.41	36.28	5.86	10.72
Reg	21.98	3.36	5.67	16.06	2.35	4.9
MLP	21.76	3.47	5.96	17.05	2.63	5.31
LSTM	14.37	3.11	4.18	14.12	3.1	3.1

With a missing rate representative of extreme data gaps, all methods experience reduced accuracy, yet LSTM maintains its lead across all metrics. This confirms its effectiveness in learning long-term temporal dependencies, making it the most robust method under severe missingness conditions.

Conclusions

The article presents and compares different solutions of the system forecasting missing data of air pollution PM10 and PM2.5. An essential point of this research is the analysis of the occurrence of missing data, their length and quantity, and testing methods on different configurations of missing data.

The experimental results reveal that LSTM models consistently deliver the best performance, particularly in moderate to severe missingness scenarios. LSTM's capacity to learn and leverage temporal dependencies allows it to achieve the lowest errors across nearly all conditions, validating its suitability for time series

imputation tasks in environmental datasets.

In contrast, mean imputation performs poorly across all levels of missingness, exhibiting the highest errors and failing to capture even basic temporal structure. This result highlights the risks of relying on naive statistical methods in contexts where the data exhibits strong time-dependent behavior.

Regression-based and MLP imputations, while not as effective as neural models, consistently outperform mean imputations and show particular promise in low and moderate missingness scenarios. It represents a viable lightweight alternative when computational resources are limited, or model interpretability is prioritized.

Additionally, during the testing process, analyzing individual errors and imputation processes for individual research stations, it can be noticed that using the same method on data from different stations, the error results can differ significantly. To carry out further research, one can focus on more complex relationships in terms of patterns of missing data occurrence, which can help capture the models' ability to apply more effective imputation.

References

1. CA Pope, DW Dockery (2006) "Health effects of fine particulate air pollution: Lines that connect," Journal of the Air & Waste Management Association 56: 709-742.
2. ML Block, L Calder (2009) on-Garcidue˜nas, "Air pollution: Mechanisms of neuroinflammation and cns disease," Trends in Neurosciences 32: 506-516.
3. I Pratama, A E Permanasari, I Ardiyanto, and R. Indrayani (2020) "A review of missing values handling methods on time-series data," in 2020 3rd International Conference on Information and Communications Technology (ICOIACT), IEEE 286-291.
4. CM Salgado, C Azevedo, H Proenca, SM Vieira (2016) "Missing data," Secondary analysis of electronic health records 143-162.
5. H Kang (2013) "The prevention and handling of the missing data," Korean Journal of Anesthesiology 64: 402-406.
6. RJ Little, D B Rubin (2019) Statistical Analysis with Missing Data, 3rd. Hoboken, https://onlinelibrary.wiley.com/doi/book/10.1002/9781119482260.
7. J M Jerez, I Molina, P J Garc´ ıa-Laencina, et al. (2010) "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," Arti f icial Intelligence in Medicine 50: 105-115.
8. S Hochreiter, J Schmidhuber (1997) "Long short-term memory," Neural Computation 9: 1735-1780.
9. Z.Che, S.Purushotham, K Cho, D Sontag and Y Liu (2018) "Recurrent neural networks for multivariate time series with missing values," in Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI) 689-696.
10. Chief Inspectorate of Environmental Protection, Air quality measurement data archive, https://powietrze.gios.gov.pl/pjp/archives.
11. A Botchkarev (2019) "A new typology design of performance metrics to measure errors in machine learning regression algorithms," Interdisciplinary Journal of Information, Knowledge, and Management 14: 045-076.