EVALUATION OF DIFFERENT IMPUTATION METHODS FOR HANDLING MISSING AIR POLLUTION DATA IN BURSA, TURKEY

Saliha Çelikcan BİLGİN
saliha.celikcanbilgin@gmail.com, https://orcid.org/0009-0006-2512-4802
Istanbul Technical University, Graduate School, Atmospheric Sciences Master Program, İstanbul, Türkiye

Prof. Dr. Hüseyin TOROS
toros@itu.edu.tr, https://orcid.org/0000-0002-3028-6308
Istanbul Technical University, Faculty of Aeronautics and Astronautics, Meteorological Engineering, İstanbul, Türkiye

Prof. Dr. Turgay Tugay BİLGİN
turgay.bilgin@btu.edu.tr, https://orcid.org/0000-0002-9245-5728
Bursa Technical University, Faculty of Engineering and Natural Sciences, Computer Engineering, Bursa, Turkiye

ABSTRACT: *Air pollution data integrity is paramount for effective environmental management and public health protection. This study aims to address the issue of data integrity related to air pollution in Bursa and evaluates various advanced imputation techniques to effectively complete missing data. Five different methods, including linear regression, K-Nearest Neighbors (KNN), decision trees, random forests, and linear interpolation, are compared for filling gaps in the $PM_{2.5}$ dataset. Our findings demonstrate that these advanced imputation techniques enhance the completeness and accuracy of air pollution data.*

The significance of our study underscores the critical role of accurately and reliably completing missing air pollution data in environmental management and public health. These findings can guide researchers in selecting the most appropriate imputation method for the analysis of air pollution data. Additionally, this study establishes a foundation for addressing the issue of missing data in future research and the formulation of environmental policies.

KEYWORDS: *Air Pollution, Data Imputation, Imputation Techniques*

Date of Submission: 14.11.2023 Date of acceptance: 24.12.2023
https://doi.org/10.5281/zenodo.10446681

1. INTRODUCTION

Air pollution is a critical environmental issue with far-reaching consequences for public health and the environment. Accurate and complete data on air quality is essential for monitoring and managing pollution levels effectively. However, missing or incomplete data can be a common challenge in environmental research. In this article, we explore various imputation methods to address missing air pollution data in Bursa province, Turkey.

Bursa, located in northwestern Turkey, is one of the country's most populous and industrialized provinces. Rapid urbanization and industrial growth have led to increased air pollution concerns in the region. Monitoring air quality is vital to assess the impact of pollutants on public health and the environment.

In this study, the $PM_{2.5}$ data of Bursa station, which has the same name as Bursa province, were taken as a sample. The data obtained from this station were used as input for advanced techniques such as linear regression, K-nearest neighbor (KNN), decision trees, random forests and linear interpolation. For the linear regression method, $PM_{2.5}$ data from Uludağ University station, which is close to Bursa station, were included in the study. By comparing and analyzing these methods, we aim to provide insights into the most appropriate approach to load air pollution data in this region.

Various data imputation methods can be used in meteorological air pollution datasets. One approach is the use of ensemble-imputation-classification frameworks, which involve filling missing records using single and multiple imputation procedures and then using an ensemble of diverse classifiers to find the air pollution level (Narasimhan & Vanitha, 2023). Another method is based on multivariate time series clustering, which imputes missing air pollutant concentration data by considering spatial and temporal similarities between monitoring stations (Alahamade et al., 2021).

Neural networks, such as NARX neural networks, can also be employed for imputing missing data in air pollutant sensors (Calle et al., 2020). Additionally, low rank matrix completion (LRMC) can be used as a spatial interpolation method to impute missing values in air pollutants data sets (Liu et al., 2020). Other techniques include mean imputation, spline interpolation, simple moving average, exponentially weighted moving average, and Kalman smoothing on structural time series and autoregressive integrated moving average models (Wijesekara & Liyanage, 2020). One study compared five imputation methods, including mean imputation, conditional mean imputation, K-Nearest Neighbor imputation, multiple imputation, and Bayesian Principal Component Analysis imputation, and found that they were able to reconstruct the dataset with good performance in terms of completeness, errors, and bias (Quinteros et al., 2019). Another study aimed to reduce uncertainty in air quality assessment by experimenting with different approaches to estimate missing pollutant data at monitoring stations and within time series (Alahamade et al., 2020). Additionally, a study evaluated four imputation methods, including Mean Top Bottom, Linear Regression, Multiple Imputation, and Nearest Neighbor, and found that Mean Top Bottom was the most appropriate method for filling in missing values in air pollutants data (Zakaria & Noor, 2018). These methods are used to fill in missing values in the air pollutants data obtained from continuous ambient air quality monitoring stations. The performance of these imputation methods can be evaluated using measures such as Mean Absolute Error, Root Mean Squared Error, Coefficient of Determination, and Index of Agreement (Hadeed et al., 2020).

2. METHODOLOGY

To mitigate the impact of missing data and ensure the completeness of air pollution records, researchers and environmental scientists rely on imputation methods. Imputation techniques are statistical approaches used to estimate or fill in missing values based on available information and patterns within the dataset. These methods play a crucial role in data preprocessing, enabling more robust analyses and modeling of air quality trends.

In this article, we delve into an exploration of common imputation methods employed in addressing mis sing air pollution data. Each of these methods has distinct characteristics and advantages tailored to accommodate the challenges posed by incomplete datasets. We have examined five imputation methods, each tailored to address the challenges of missing air pollution data.

Linear Regression Imputation: Linear regression estimates missing values by fitting a linear model based on available data. The imputed value $((Y_i))$ for the missing data point at position (i) can be calculated as follows:

$$[Y_i = \beta_0 + \beta_1 X_i + \epsilon_i] \quad (1)$$

Where:

- (Y_i) is the imputed value.
- (β_0) is the intercept of the linear model.
- (β_1) is the slope of the linear model.
- (X_i) represents the available data for other relevant variables.
- (ϵ_i) is the error term.

K-Nearest Neighbors (KNN) Imputation: K-nearest neighbors imputation estimates missing values by averaging the values of the K most similar data points. The formula for KNN imputation is:

$$[Y_i = \frac{1}{K}\sum_{j=1}^{K} Y_{N_j}] \quad (2)$$

Where:

- (Y_i) is the imputed value.
- (K) is the number of nearest neighbors.
- (Y_{N_j}) represents the observed values of the (j)-th nearest neighbor.

Decision Tree Imputation: Decision tree imputation uses a decision tree algorithm to predict missing values. The imputed value (Y_i) is determined based on the decision tree rules. Decision tree-based imputation is a method that leverages decision tree models to estimate missing data points. Decision trees are versatile algorithms that can capture complex relationships and non-linear patterns within datasets. This approach is particularly effective when dealing with datasets where variables interact in intricate and non-linear ways, allowing for more accurate imputations in such scenarios.

Random Forest Imputation: Random forests are an ensemble learning method that combines multiple decision trees. The imputed value (Y_i) is the average prediction from all the decision trees in the random forest. Random Forest Imputation leverages this forest structure to predict missing data points and often stands out for its enhanced predictive performance in the dataset.

Linear Interpolation: Linear interpolation estimates missing values by assuming a linear relationship between consecutive data points. The formula for linear interpolation between two adjacent data points (X_a) and (X_b) at times (t_a) and (t_b) is:

$$[Y_i = \frac{(X_b - X_a)}{(t_b - t_a)}(t_i - t_a) + X_a] \quad (3)$$

Where:

- (Y_i) is the imputed value at time (t_i).
- (X_a) and (X_b) are the known values at times (t_a) and (t_b).
- (t_i) is the time for which imputation is needed.

3. FINDINGS

To evaluate the performance of these imputation methods, we collected $PM_{2.5}$ data from Bursa station. We introduced artificial missing values into the dataset to simulate real-world scenarios. In addition, we also utilized air pollution data from the Uludag University station, located close to the Bursa station, and used this data as a correlation dataset with no missing values.

The 24-hour $PM_{2.5}$ data of 2 stations obtained from the National Air Quality Monitoring Network (UHKİA) of the Ministry of Environment, Urbanization and Climate Change of the Republic of Turkey dated May 14, 2022, were used (*Hava Kalitesi - İstasyon Veri İndirme | T.C. Çevre, Şehircilik ve İklim Değişikliği Bakanlığı*, n.d.).We removed 7 consecutive hours of data (from 04:00 to 10:00) from the Bursa station from the data set.

In our comparative analysis, we placed a particular emphasis on understanding the shape and trend of missing data patterns within the air pollution dataset. This involved closely examining the distribution of missing values over time or space and identifying any discernible trends or patterns. Understanding the temporal or spatial characteristics of missing data is crucial for selecting the most appropriate imputation method, as different methods may be better suited to handle specific missing data profiles.

Additionally, we paid special attention to instances of large consecutive missing data stretches within the dataset. These extended gaps in the data can significantly impact the integrity of the dataset and, subsequently, the accuracy of imputations. Identifying and addressing such consecutive missing data is essential to ensure the reliability of the imputation results and the overall quality of the air pollution dataset.

By focusing on both the shape and trends of missing data, as well as addressing consecutive missing data, our analysis aimed to provide a comprehensive assessment of the performance and suitability of different imputation methods in handling the unique challenges posed by the air pollution dataset in Bursa station.

Figure 1 shows the comparison of 7 consecutive hours of missing $PM_{2.5}$ values of Bursa station with the actual values. Accordingly, it is seen that the closest data imputation process to the actual values of the data at Bursa station is realized with the Linear Regression method using the data at Uludag University station. The results of the other methods are far from a data imputation in accordance with the actual values.

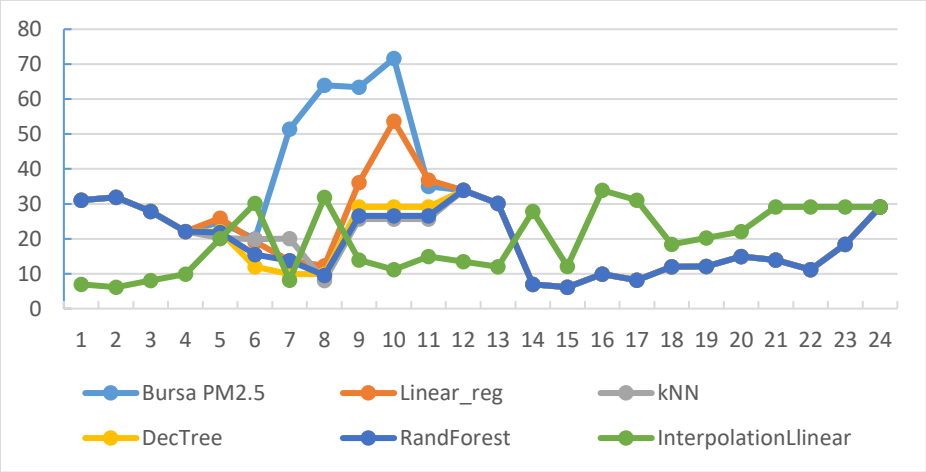

Figure 1: *Comparison of Different Method Results with Actual Values*

4. DISCUSSIONS

In the course of our comprehensive investigation, we harnessed air pollution data from two key monitoring stations: Bursa station and Uludag University station, with a specific focus on $PM_{2.5}$ concentrations. At Bursa station, we created a new test dataset by removing the data points between 04:00 and 10:00 on May 14, 2022. We then started the task of estimating the missing values in this specific time interval. We used the Uludag University data, which is located close to Bursa station, in a linear regression method. In the pursuit of our research objectives, we deployed an array of sophisticated imputation methodologies, including linear regression, K-nearest neighbor (KNN), Decision Tree, Random Forest, and Linear Interpolation. These methodologies were carefully selected due to their efficacy in handling missing data scenarios.

Our comparative analysis was predicated on the alignment of imputed missing data with actual observed data. The results of these experiments provide valuable insights into the performance of each imputation method, as shown in Figure 1.

Notably, our findings reveal that the Linear Regression method emerged as the most proficient in terms of replicating the form and trend of the missing data. This outcome underscores the utility of linear regression in addressing missing data challenges in air pollution datasets. However, it is essential to recognize that the choice of the most suitable imputation method should be contingent upon the specific characteristics of the dataset and the research objectives at hand.

The empirical evidence presented in this study contributes to the broader discourse on imputation techniques for air pollution data, offering a nuanced perspective on their efficacy and applicability in real-world environmental monitoring and management contexts.

5. CONCLUSION

Researchers need to particularly consider factors such as the characteristics of data in their study domains and the requirements for accuracy. The conducted comparative analysis has demonstrated that advanced methods, including linear regression, K-Nearest Neighbors (KNN), decision trees, random forests, and linear interpolation, can offer higher accuracy and flexibility compared to simple statistical methods.

Researchers should carefully evaluate the specific needs of their studies and the requirements of their datasets when selecting a data imputation method. Additionally, it is essential to acknowledge the limitations and potential biases introduced by imputation and to document the imputation process transparently in research publications.

As a result, it is observed that advanced imputation methods play a critical role in enhancing the reliability and completeness of air pollution data, both in the province of Bursa and from a broader perspective. This circumstance contributes to the more effective preservation of environmental management and public health, providing opportunities for the development of more efficient strategies in these areas.

REFERENCES

Alahamade, W., Lake, I., Reeves, C. E., & De La Iglesia, B. (2021). Evaluation of multi-variate time series clustering for imputation of air pollution data. *Geoscientific Instrumentation, Methods and Data Systems Discussions*, *2021*, 1–23.

Alahamade, W., Lake, I., Reeves, C. E., & De La Iglesia, B. (2020). Clustering Imputation for Air Pollution Data. In E. A. De La Cal, J. R. Villar Flecha, H. Quintián, & E. Corchado (Eds.), *Hybrid Artificial Intelligent Systems* (Vol. 12344, pp. 585–597). Springer International Publishing. https://doi.org/10.1007/978-3-030-61705-9_48

Calle, M., Orellana, M., & Ortega-Chasi, P. (2020). NARX Neural Network for Imputation of Missing Data in Air Pollution Datasets. In G. Rodriguez Morales, E. R. Fonseca C., J. P. Salgado, P. Pérez-Gosende, M. Orellana Cordero, & S. Berrezueta (Eds.), *Information and Communication Technologies* (Vol. 1307, pp. 226–240). Springer International Publishing. https://doi.org/10.1007/978-3-030-62833-8_18

Hadeed, S. J., O'Rourke, M. K., Burgess, J. L., Harris, R. B., & Canales, R. A. (2020). Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Science of the Total Environment*, *730*, 139140.

Hava Kalitesi—İstasyon Veri İndirme | T.C. Çevre, Şehircilik ve İklim Değişikliği Bakanlığı. (n.d.). Retrieved October 26, 2023, from https://sim.csb.gov.tr/STN/STN_Report/StationDataDownloadNew

Liu, X., Wang, X., Zou, L., Xia, J., & Pang, W. (2020). Spatial imputation for air pollutants data sets via low rank matrix completion algorithm. *Environment International*, *139*, 105713.

Narasimhan, D., & Vanitha, M. (2023). Machine Learning Approach-based Big Data Imputation Methods for Outdoor Air Quality Forecasting. *Journal of Scientific & Industrial Research*, *82*(03), 338–347.

Quinteros, M. E., Lu, S., Blazquez, C., Cárdenas-R, J. P., Ossa, X., Delgado-Saborit, J.-M., Harrison, R. M., & Ruiz-Rudolph, P. (2019). Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile. *Atmospheric Environment*, *200*, 40–49.

Wijesekara, W. M. L. K. N., & Liyanage, L. (2020). Comparison of Imputation Methods for Missing Values in Air Pollution Data: Case Study on Sydney Air Quality Index. In K. Arai, S. Kapoor, & R. Bhatia (Eds.), *Advances in Information and Communication* (Vol. 1130, pp. 257–269). Springer International Publishing. https://doi.org/10.1007/978-3-030-39442-4_20

Zakaria, N. A., & Noor, N. M. (2018). Imputation methods for filling missing data in urban air pollution data formalaysia. *Urbanism. Arhitectura. Constructii*, *9*(2), 159.