The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM)
ISSN: 2602-3199

The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 2024

Volume 27, Pages 122-129

IConTech 2023: International Conference on Technology

Review of Data Imputation Techniques in Time Series Data: Comparative Analysis

Yousef Jaradat
Al-Zaytoonah University of Jordan

Mohammad Masoud
Al-Zaytoonah University of Jordan

Ahmad Manasrah
Al-Zaytoonah University of Jordan

Mohammad Alia
Al-Zaytoonah University of Jordan

Ismael Jannoud
Al-Zaytoonah University of Jordan

Abstract: This review paper offers a thorough analysis of different data imputation methods that can be applied to time series data. Time series data is an essential element in various analytical and predictive models used in different domains. Time series data frequently experiences missing values as a result of diverse factors, such as system errors, human influences, or inherent gaps in data collection. The presence of these missing values significantly undermines the precision and dependability of models constructed using this data. This paper classifies imputation methods into basic and advanced techniques, providing a comprehensive examination of each. The simplicity and effectiveness of basic techniques, such as mean or median imputation and linear interpolation, are discussed in specific scenarios. The study investigates the efficacy of advanced techniques, such as ARIMA statistical models, K-Nearest Neighbors machine learning approaches, and Long Short-Term Memory networks deep learning techniques, in managing intricate and extensive time series datasets. The paper emphasizes a comparative approach, assessing each method's complexity, accuracy, and computational demands. The paper concludes by emphasizing the need for continuous innovation in imputation techniques to meet the growing complexity of time series data across various domains. It advocates for a collaborative approach that combines domain expertise with advanced data science methods to develop tailored, efficient, and accurate imputation strategies.

Keywords: Time Series, Data Imputation, Machine Learning, Deep Learning, Computational complexity

Introduction

Time series data, which consists of sequential measurements taken over time, is a crucial component of analysis in diverse scientific and commercial domains. The data sets, observed at consecutive time intervals, are essential for uncovering patterns, trends and predicting future events (Box et al., 2015). Time series data, which is frequently encountered in finance (Alshehadeh et al., 2023), meteorology (Sun et al., 2024), and healthcare (Morid et al., 2023), plays a vital role in making well-informed decisions and predictions. Nevertheless, the dependability and precision of conclusions drawn from such data are directly dependent on its quality. Ensuring high data quality in time series is crucial, as any inaccuracies or missing values can result in misleading analyses

and incorrect predictions. Missing values in time series data can frequently compromise its integrity. These missing values can occur due to various factors, such as equipment malfunction, human error, or gaps in data collection (Donders et al., 2006). The absence of these values has a significant effect, especially in time-critical fields such as stock market analysis or weather forecasting, where accuracy is crucial. Hence, it is imperative to guarantee optimal data quality by employing efficient data imputation techniques for conducting time series analysis.

Data imputation is the process of substituting missing data with alternative values. The application of imputation in time series data is particularly challenging and complex due to the sequential nature of the data. Time series data differs from cross-sectional data in that it exhibits temporal dependencies, indicating that each data point may be influenced by its preceding and subsequent points in time (Hyndman & Athanasopoulos, 2018). Specialized imputation techniques are necessary for this dependency to properly acknowledge and utilize these temporal relationships.

The range of imputation techniques available for time series data is extensive, encompassing both basic methods and advanced algorithms. Common techniques for handling missing values include mean, median, and mode imputation. More advanced methods, such as linear interpolation, involve filling in missing values by estimating them based on the linear patterns observed in neighboring data points. Advanced methodologies encompass statistical models such as Autoregressive Integrated Moving Average (ARIMA), as well as machine learning techniques, including but not limited to K-Nearest Neighbors (KNN) and deep learning algorithms such as Long Short-Term Memory (LSTM) networks (Fang & Wang, 2020; Jaradat et al., 2021). These advanced techniques utilize the inherent patterns and relationships within the data, providing more precise and contextually appropriate imputation. The importance of these imputation techniques has increased significantly due to the expanding volume and complexity of time series data. This review paper seeks to clarify the different imputation methods, their suitability, and efficacy in the domain of time series data. The paper aims to present a comprehensive perspective on current methodologies and their effects on the quality and dependability of time series analysis by examining a range of techniques, from basic to advanced. Furthermore, it will explore practical applications in various industries, showcasing the importance of effective data imputation in extracting precise insights from time series data.

In summary, data imputation in time series is not merely a corrective measure for missing data but a crucial aspect that ensures the integrity and usability of the data. The ongoing developments in this field highlight the growing importance of sophisticated imputation techniques in the face of increasingly complex and large-scale time series datasets.

The remainder of the paper is structured as follows. Understanding missing data in time series is described in Section II. Section III summarizes the different categories of data imputation in time series data. Section IV provides a comparative analysis metrics for data imputation in time series. Section V concludes the paper.

Understanding Missing Data in Time Series

In the field of time series analysis, dealing with missing data is an unavoidable and substantial obstacle that analysts and researchers must address. Comprehending the characteristics of missing data, its influence on the analysis of time series, and the inherent difficulties in dealing with it is essential for effective data imputation and subsequent analysis. Missing data in time series can be classified into three main types (Donders et al., 2006, Ahn et al., 2022): Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR).

- *Missing Completely at Random (MCAR)*: This occurs when the probability of missing data is the same for all observations, implying that the missingness is independent of both observed and unobserved data. For example, if temperature readings from a weather station are missing due to random equipment failures, the missingness is likely MCAR.

- *Missing at Random (MAR)*: In this scenario, the missingness is related to the observed data but not the missing data itself. For instance, in a financial time series, if stock prices are recorded only on days when the market is up, the missing data on down days is MAR, as it depends on the market's performance (an observed variable).

- *Not Missing at Random (NMAR)*: Here, the missingness is related to the unobserved data. For example, if patients in a clinical study tend to drop out as their health deteriorates, the missing data (dropout) is related to the unobserved health status of the patient.

Understanding these categories is crucial for choosing the suitable imputation technique, as each category of missing data necessitates a distinct approach for precise imputation. Missing data can significantly distort the analysis of time series. It can lead to biased estimates, reduced statistical power, and it can make the data analysis more complex and less accurate. The impact is more pronounced in time series data due to its sequential nature and the temporal correlations present. For instance, missing data in a time series can disrupt the continuity needed for trend analysis and forecasting, leading to misleading conclusions.Imputing missing data in time series presents unique challenges (Moritz & Bartz-Beielstein, 2017; Ribeiro & Castro, 2021):

- *Maintaining Temporal Structure*: Time series data is defined by its sequential order and temporal dependencies. Imputation methods must preserve these characteristics to maintain the integrity of the data.

- Handling Seasonality and Trends: Time series often exhibit seasonality and trends. Imputation methods need to account for these patterns, which requires sophisticated techniques, especially when large portions of data are missing.

- Varying Missing Data Patterns: The pattern of missingness can vary greatly in time series data. Methods that work well for random, sporadic missingness might not be effective for systematic missingness, such as data missing at regular intervals.

- Computational Complexity: Some of the more effective methods for imputing missing time series data, like state-space models or multiple imputation, can be computationally intensive, especially for large datasets.

Imputation Techniques in Time Series Data

Imputation techniques in time series data can be broadly classified into two types (Morid et al., 2023; Ahn et al., 2022): Basic imputation techniques and advanced imputation techniques as shown in Figure 1.

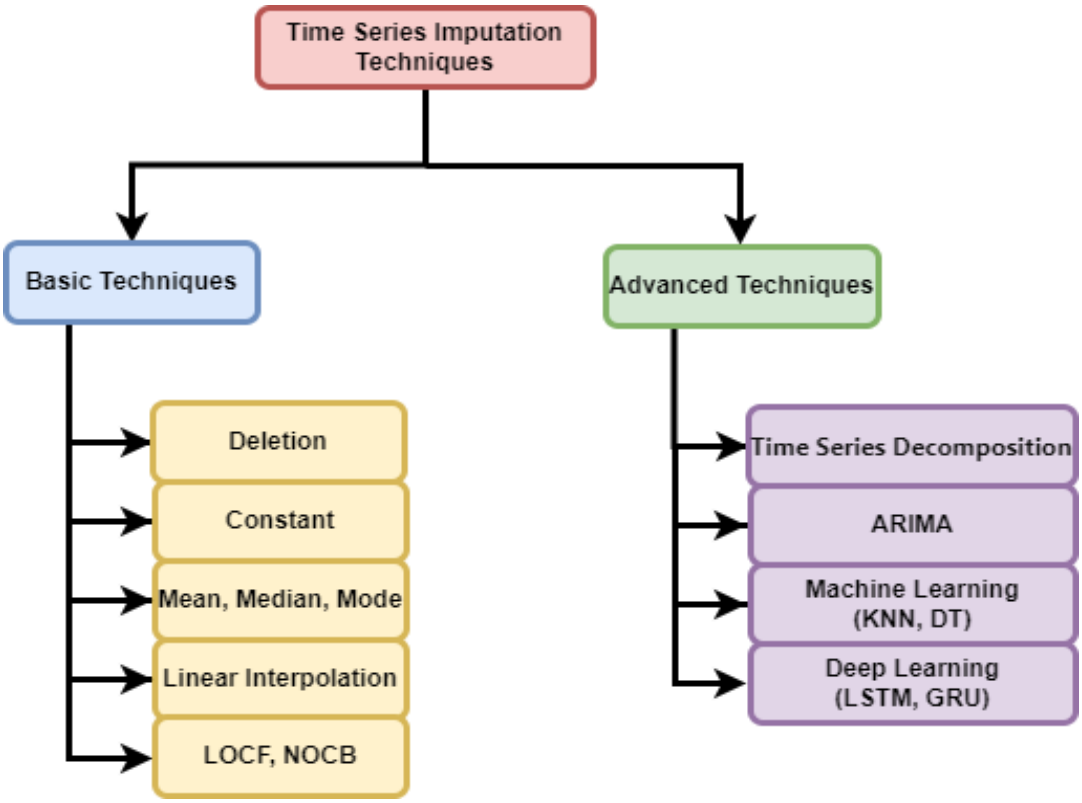

Figure 1. Time series imputation techniques

A) **Basic Imputation Techniques in Time Series Data** (Pratama et al., 2013; John et al., 2019).

- *Deletion*: This basic imputation method involves eliminating any data points in a time series that are missing values. It is particularly useful for large datasets with a small percentage of missing data. The process is straightforward and doesn't require the estimation of missing values. However, deletion can result in a significant loss of data and is not advisable for time series with a considerable amount of missing data. This method assumes that data is Missing Completely at Random (MCAR), which can introduce bias if this assumption is incorrect. Furthermore, deletion reduces the overall size of the dataset, potentially impacting the model's statistical power.

- *Constant Imputation*: This technique substitutes missing values in the data with a fixed constant. This constant might be a generic number such as zero, or a value that holds particular relevance in the analysis context. Advantages include its simplicity and the preservation of the original dataset size. However, it may lead to biased results and fails to consider the dynamics of the time series. When building models, using constant imputation can alter the natural pattern of the time series, potentially resulting in inaccurate outcomes in further analyses.

- *Mean, Median, and Mode Imputation*: This approach replaces missing values with the mean, median, or mode of the existing data. It is straightforward to apply and offers an improvement over constant imputation by utilizing information from the dataset. Nonetheless, these methods can be deceptive if the data is not MCAR and do not consider the specific dynamics of time series. In terms of model building, although these methods are an advancement over constant imputation, they can still result in skewed models if the assumption of MCAR for the missing data is not met.

- *Linear Interpolation*: This technique estimates missing values by creating a straight line between two neighboring known values and using this line for interpolation. It is advantageous as it considers the sequence of the data and is apt for datasets with linear trends. However, it is not appropriate for non-linear time series as it presumes that the data points between known values alter at a consistent rate. Regarding model building, while linear interpolation can be useful for linear time series, it may result in considerable bias when applied to non-linear series.

- *Last Observation Carried Forward (LOCF) and Next Observation Carried Backward (NOCB)*: In LOCF missing values are imputed using the last observed value. While in NOCB the next available observation is used to fill the missing value. These methods are straightforward and help to maintain the chronological integrity of the data. However, they can create lagged effects and are not suitable for time series with significant variability or trending behaviors. In terms of model building, while LOCF and NOCB maintain the data's time sequence, they might not accurately represent the true nature of the data, particularly in fluctuating time series.

B) **Advanced Imputation Techniques in Time Series Data** (Morid et al., 2023, Ribeiro & Castro, 2021).

- *Time Series Decomposition*: This method divides the time series into its constituent parts: trend, seasonal, and residual components, allowing for separate imputation of missing values in each. It is beneficial as it addresses the inherent patterns within the data, particularly useful for time series with seasonal variations. However, it is more intricate and relies on clearly identifiable trends and seasonality. Regarding its impact on model building, decomposition can notably enhance the precision of models in time series that exhibit seasonal and trend patterns, but it may not perform as well with data that is irregular or lacks seasonal characteristics.

- *ARIMA-based Imputation*: In this approach, missing values are predicted using the Autoregressive Integrated Moving Average (ARIMA) model, which relies on the existing data points. This technique capitalizes on the autocorrelations present in the data, making it ideal for stationary time series. However, it is a sophisticated method and necessitates that the time series be stationary. In terms of its effect on model building, while ARIMA-based imputation can improve accuracy in suitable situations, its effectiveness is constrained by the requirement for the data to be stationary.

- *Machine Learning Approaches*: This category includes techniques like K-Nearest Neighbors (KNN) and Decision Trees (DT), which predict missing values by identifying similar patterns or applying decision rules extracted from the dataset. These methods are adept at recognizing intricate patterns and are effective for

handling non-linear data. However, they demand a sizeable dataset and are computationally demanding. Regarding their influence on model building, these approaches are versatile and potent, yet they run the risk of overfitting the data and necessitate significant computational power.

Table 1. Summarizes the pros, cons and the impact on model building for all imputation techniques.

	Pros	*Cons*	*Impact on Model Building*
Deletion	- Simple and easy to implement. - No need to estimate missing values.	- Can lead to significant data loss - Not suitable for time series with substantial missing data. -Risks bias as it assumes the data is MCAR.	- It can lead to biased models if the assumption of MCAR is violated. - It also reduces the dataset size, which can affect the statistical power of the model.
Constant	- Simple and easy to implement. - Maintains the dataset size	- Can introduce bias -Doesn't take into account the time series dynamics	- Can distort the inherent structure of the time series data, leading to misleading results in subsequent analysis.
Mean, Median, Mode	- Easy to implement -Better than constant imputation as they use information from the data	- Can be misleading if the data is not MCAR - Does not account for time series dynamics.	-While these methods are better than constant imputation, they can still lead to biased models if the missing data is not MCAR.
Linear interpolation	- Takes into account the order of the data - Suitable for data with linear trends	- Not suitable for non-linear time series - Assumes that the data points between known values change at a constant rate.	- Can be effective for time series with linear trends, but it can introduce significant bias in non-linear series.
LOCF, NOCB	- Simple to implement - Maintains the data's temporal structure	- Can introduce lagged effects. - Inappropriate for time series with high variability or trends.	- Both LOCF and NOCB can preserve the time sequence but may not accurately reflect the underlying data process, especially in volatile time series.
Time Series Decomposition	- Accounts for underlying patterns in the data. - Effective for seasonal time series.	- More complex; requires a well-defined trend and seasonality.	- Can significantly improve model accuracy in seasonal and trended time series but may be less effective in irregular or non-seasonal data.
ARIMA	- Utilizes the autocorrelations in the data - Suitable for stationary time series.	- Complex - Requires the time series to be stationary.	-ARIMA-based imputation can enhance model accuracy in appropriate contexts but is limited by its assumptions about stationarity.
Machine Learning (KNN, DT)	-Can capture complex patterns. -Suitable for non-linear data.	-Requires a large dataset. -Computationally intensive.	-These techniques are powerful and flexible but may overfit the data. -Require substantial computational resources.
Deep Learning (LSTM, GRU)	-Highly effective in capturing long-term dependencies; -Suitable for complex and large datasets.	-Computationally expensive. -Requires expertise to tune and interpret.	-These methods can significantly improve model performance, especially in complex time series, but are resource-intensive and require careful tuning.

- *Deep Learning Methods*: Utilizing advanced neural networks like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) (Park et al., 2023), these methods are ideal for modeling sequences that exhibit long-term dependencies. They are particularly adept at identifying long-range patterns and work well with complex, large datasets. However, these methods are resource-heavy and necessitate specialized knowledge for effective tuning and interpretation. In terms of their impact on model building, while they can greatly enhance the performance of models, especially in intricate time series, they demand considerable resources and meticulous calibration

Comparative Analysis of Imputation Techniques

When comparing basic and advanced imputation techniques, several key factors emerge (Jadhav et al., 2019, Hamzah et al., 2021):

1. *Complexity vs. Simplicity*: Basic techniques like mean, median, mode imputation, and LOCF are simple to implement but often fail to capture the complexities of time series data. In contrast, advanced techniques like ARIMA, machine learning approaches, and deep learning methods are more complex but can handle intricate patterns in data more effectively.

2. *Applicability to Data Structure*: Basic techniques are generally more suited to datasets where the missing data is random and does not form a significant portion of the dataset. Advanced techniques are preferred in scenarios where the data exhibits strong temporal dependencies, non-linear patterns, or seasonality.

3. *Accuracy and Bias*: Basic methods can introduce bias, particularly if the missing data is not random. Advanced techniques, with their sophisticated algorithms, tend to provide more accurate imputations but at the cost of increased complexity and the risk of overfitting.

4. *Computational Resources*: Advanced techniques, especially deep learning methods, require substantial computational resources and expertise, which may not be feasible in all scenarios. Basic methods are computationally inexpensive and easier to implement.

5. *Flexibility and Adaptability*: Advanced techniques are more adaptable to different types of data and missingness patterns. They can be tailored to the specific characteristics of the dataset, whereas basic techniques offer limited flexibility.

6. *Impact on Model Building*: The choice of imputation technique has a direct impact on the subsequent model building. Basic techniques might be sufficient for simple models and analyses, but advanced techniques are often necessary for complex, high-stakes analyses where accuracy is paramount.

Selecting between fundamental and sophisticated imputation methods for time series data hinges on several considerations. These include the characteristics of the missing data, dataset complexity, the level of accuracy sought in the imputation, available computational power, and the particular demands of the ensuing analysis. Basic methods are straightforward and user-friendly, whereas advanced techniques yield higher precision and are more apt for intricate datasets. Nonetheless, the latter demand greater computational effort and expertise. The optimal strategy typically involves striking a balance among these elements, customized to the unique requirements and limitations of the specific project.

Conclusion

This review has comprehensively explored the diverse landscape of data imputation techniques in time series data, offering insights into both basic and advanced methodologies, their applications, and a comparative analysis of their effectiveness. The journey from simple imputation methods like mean, median, mode, and deletion to more sophisticated techniques such as ARIMA modeling, machine learning approaches, and deep learning methods like LSTMs and GRUs, reflects the evolving complexity and growing needs of time series analysis in a data-driven world.

A key takeaway from this review is the critical importance of understanding the nature of the missing data and the specific requirements of the time series dataset. The choice between basic and advanced imputation methods should not be made lightly; it requires careful consideration of the dataset's characteristics, the pattern of

missingness, and the ultimate goal of the analysis. Basic methods, while easier to implement and less resource-intensive, may fall short in datasets where the missing data is not random or where complex temporal dynamics are at play. In contrast, advanced techniques, though more accurate and capable of handling intricate patterns, come with the cost of increased computational complexity and the need for specialized expertise.

Scientific Ethics Declaration

The authors declare that the scientific ethical and legal responsibility of this article published in EPSTEM Journal belongs to the authors.

Acknowledgements or Notes

* This article was presented as an oral presentation at the International Conference on Technology (www.icontechno.net) held in Alanya/Turkey on May 02-05, 2024.

* This research was funded by Al-Zaytoonah University of Jordan. Project number 2023-2022/17/41.

References

Ahn, H., Sun, K., & Kim, K. P. (2022). Comparison of missing data imputation methods in time series forecasting. *Computers, Materials & Continua, 70*(1), 767-779.

Alshehadeh, A., Alia, M., Jaradat, Y., Injadat, E., & Al-khawaja, H. (2023). Big data analytics techniques and their impacts on reducing information asymmetry: Evidence from Jordan. *International Journal of Data and Network Science, 7*(3), 1259-1266.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology, 59*(10), 1087-1091.

Fang, C., & Wang, C. (2020). Time series data imputation: A survey on deep learning approaches. arXiv preprint arXiv:2011.11347.

Hamzah, F. B., Hamzah, F. M., Razali, S. M., & Samad, H. (2021). A comparison of multiple imputation methods for recovering missing data in hydrological studies. *Civil Engineering Journal, 7*(9), 1608-1619.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence, 33*(10), 913-933.

Jaradat, Y., Masoud, M., Jannoud, I., Manasrah, A., & Alia, M. (2021). A tutorial on singular value decomposition with applications on image compression and dimensionality reduction. In *2021 International Conference on Information Technology* (ICIT) (pp. 769-772). IEEE.

John, C., Ekpenyong, E. J., & Nworu, C. C. (2019). Imputation of missing values in economic and financial time series data using five principal component analysis approaches. *CBN Journal of Applied Statistics (JAS), 10*(1), 3.

Morid, M. A., Sheng, O. R. L., & Dunbar, J. (2023). Time series prediction using deep learning methods in healthcare. *ACM Transactions on Management Information Systems, 14*(1), 1-29.

Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: Time series missing value imputation. *Journal of Statistical Software, 9*(1), 207.

Park, J., Müller, J., Arora, B., Faybishenko, B., Pastorello, G., Varadharajan, C., ... & Agarwal, D. (2023). Long-term missing value imputation for time series data using deep neural networks. *Neural Computing and Applications, 35*(12), 9071-9091.

Pratama, I., Permanasari, A. E., Ardiyanto, I., & Indrayani, R. (2016). A review of missing values handling methods on time-series data. In *2016 International Conference on Information Technology Systems and Innovation* (ICITSI) (pp. 1-6). IEEE.

Ribeiro, S. M., & de Castro, C. L. (2021). Missing data in time series: A review of imputation methods and case study. *Learning and Nonlinear Models-Revista Da Sociedade Brasileira De Redes Neurais-Special Issue: Time Series Analysis and Forecasting Using Computational Intelligence, 19*(2).

Sun, Y., Deng, K., Ren, K., Liu, J., Deng, C., & Jin, Y. (2024). Deep learning in statistical downscaling for deriving high spatial resolution gridded meteorological data: A systematic review. *ISPRS Journal of Photogrammetry and Remote Sensing, 208,* 14-38.

Author Information

Yousef Jaradat
Al-Zaytoonah University of Jordan
Amman, Jordan
Contact e-mail: *y.jaradat@zuj.edu.jo*

Mohammad Masoud
Al-Zaytoonah University of Jordan
Amman, Jordan

Ahmad Manasrah
Al-Zaytoonah University of Jordan
Amman, Jordan

Mohammad Alia
Al-Zaytoonah University of Jordan
Amman, Jordan

Ismael Jannoud
Al-Zaytoonah University of Jordan
Amman, Jordan

To cite this article:

Jaradat, Y., Masoud, M., Manasrah, A., Alia, M., & Jannoud, I. (2024). Review of data imputation techniques in time series data: Comparative analysis. *The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 27,* 122-129.