Sigma J Eng Nat Sci, Vol. 43, No. 1, pp. 199–212, February, 2025

Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2025.00015

Research Article

Performance of imputation techniques: A comprehensive simulation study using the transformer model

İsmail YENİLMEZ[1,*]

[1]*Department of Statistics, Eskişehir Technical University, Eskişehir, 26555, Türkiye*

ARTICLE INFO

Article history
Received: 06 May 2024
Revised: 08 July 2024
Accepted: 03 October 2024

Keywords:
ARIMA; GARCH; Kalman Filtering; Missing Data; Monte Carlo Simulation

ABSTRACT

This study addresses the critical challenge of handling missing data in time series analysis, which is maintaining the accuracy and reliability of financial forecasting and other predictive models. The study aims to assess various imputation techniques' and estimation methods' performance. The purpose of using imputed data is to enhance the robustness and accuracy of time series analyses, especially when dealing with incomplete datasets. We compared eight different imputation methods to identify the most effective approach. We also compared the performance of the Transformer model, Autoregressive Integrated Moving Average, and Generalized Autoregressive Conditional Heteroskedasticity methods in time series analysis using both complete and imputed datasets. The study employed a comprehensive approach, utilizing the Transformer model, Autoregressive Integrated Moving Average, and Generalized Autoregressive Conditional Heteroskedasticity for time series analysis. Eight imputation methods—last observation carried forward, next observation carried backward, mean imputation, linear interpolation, seasonal decomposition, moving average, regression imputation, and Kalman filtering—were evaluated. Monte Carlo simulations and an application were conducted on generated and real data-driven datasets with different proportions of missing data to assess the performance of these methods. The findings suggest that imputation techniques, such as mean imputation, considered conventional, and Kalman filtering, can significantly enhance the accuracy of time series models, particularly when integrated with innovative models like the Transformer. Moreover, the last observation carried forward, seasonal decomposition, and moving average did not provide better results in any scenario. Simulation-based synthetic data and application-based real data also revealed that the Transformer model outperformed traditional methods in scenarios with complete data (the original dataset) and new datasets generated through imputation at different rates. The results obtained from the real data-driven application support the findings from the simulation results. In addition to the simulation findings, the application results show that mean imputation performs well in cases with low levels of imputation, while Kalman filtering proves more successful when imputing a high proportion of missing data. This work goes beyond previous studies by systematically comparing a wide range of imputation methods within a unified framework, incorporating both traditional and modern time series models. A comprehensive evaluation of estimation techniques and imputation strategies applicable to time series analysis is presented, exploring appropriate combinations of estimation methods and imputation techniques.

Cite this article as: Yenilmez İ. Performance of imputation techniques: A comprehensive simulation study using the transformer model. Sigma J Eng Nat Sci 2025;43(1):199–212.

***Corresponding author.**
*E-mail address: ismailyenilmez@eskisehir.edu.tr

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

Published by Yıldız Technical University Press, İstanbul, Turkey

INTRODUCTION

In time series analysis, handling missing data is a critical challenge, as the accuracy and reliability of the results depend heavily on the effectiveness of imputation methods. These methods not only address gaps in the data but also improve the ability to predict future values. Traditional models like the Autoregressive Integrated Moving Average (ARIMA) [1] and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) [2] have been extensively used for forecasting and are well-established in time series analysis. For instance, ARIMA is effective in modeling time series data by predicting future values based on past observations, but its performance can be compromised in the presence of missing data [3]. Similarly, GARCH models, known for their ability to model volatility, also struggle with incomplete datasets, potentially leading to biased estimates and inaccurate forecasts [2].

To address these limitations, advanced imputation techniques have been developed. Kalman filtering, a state-space model, has been successfully applied to estimate missing values by leveraging dynamic linear models [4]. Moreover, recent research has explored even more sophisticated imputation methods, such as the generalized m-parameter Mittag-Leffler function, which has shown promise in handling complex differential and integral equations [5]. Despite these advancements, there is still a need for a systematic evaluation of how these methods perform in different scenarios, particularly when integrated with newer models like Transformers, which have shown significant potential in time series analysis [6].

Related Work

This section presents a comprehensive review of existing studies on imputation techniques and their application in time series analysis. ARIMA and GARCH models are widely used in time series forecasting due to their robustness and efficacy in handling various types of data. These models have been extensively applied across multiple domains. For instance, [7] employed GARCH models to capture the volatility in energy markets, demonstrating their ability to model fluctuations in energy prices effectively. Similarly, [8] utilized ARIMA and GARCH models for traffic modeling and prediction in telecommunication networks, emphasizing the models' robustness in handling complex network traffic data. Furthermore, [9] illustrated the integration of ARIMA and GARCH models for forecasting the USD/EUR exchange rate, highlighting the enhanced prediction accuracy achieved through this combination.

The hybrid ARIMA-GARCH model is particularly effective in financial forecasting, where both linear patterns and volatility need to be accounted for. [10] demonstrated the superiority of this hybrid model in gold price forecasting, where it significantly improved forecasting accuracy by addressing both linear trends and volatility. Additionally, comparisons between traditional time series models and machine learning models have garnered attention in the literature. Studies by [11,12], and [13] have highlighted the potential of neural network models in capturing complex patterns in data, particularly in finance and economics.

The Transformer model, a deep learning approach leveraging self-attention mechanisms, has gained prominence in time series analysis due to its ability to capture long-term dependencies. Unlike traditional models like ARIMA and GARCH, which rely on past values and variances, the Transformer model utilizes self-attention to weigh the importance of different time steps. This makes it particularly effective for complex and irregular time series data [14,15]. However, the Transformer model is not without limitations. One significant challenge is its computational inefficiency, especially concerning the self-attention mechanism, which scales quadratically with the length of the input sequence. This can become a bottleneck when dealing with long time series, leading to substantial time and memory complexity [3, 16]. Additionally, standard Transformers may struggle with capturing local dependencies in time series data, which is crucial for accurate forecasting and anomaly detection.

Handling missing data is a critical aspect of time series analysis, as it directly impacts the accuracy and reliability of the results. Various imputation methods are employed to address this challenge, each with its strengths and limitations. Commonly used techniques include Last Observation Carried Forward (LOCF), Next Observation Carried Backward (NOCB), Mean Imputation, Linear Interpolation, Seasonal-Trend Decomposition using Regression (STR), Moving Average, Regression Imputation, and Kalman Filtering [17-20]. The choice of imputation method is critical in time series analysis, where preserving temporal dependencies and trends is essential.

Research by [21,22] emphasizes that assumptions of linearity and stationarity in time series data may not always hold, making the selection of imputation methods even more crucial. More recent advancements in imputation methods are highlighted by [23], who introduced a modified genetic algorithm for the Travelling Salesman Problem, featuring novel crossover and mutation operators that could be adapted for time series imputation. Additionally, [24,25] presented numerical methods for solving complex differential equations, which could enhance imputation accuracy in datasets with unique structural characteristics. The dual hesitant fuzzy set theoretic approach in fuzzy reliability analysis, as discussed by [26,27], offers a theoretical foundation that could further improve the accuracy and robustness of imputation methods, particularly in systems characterized by uncertainty.

In comparing the performance of ARIMA, GARCH, and Transformer models across different imputation scenarios, it is essential to consider how each method handles missing data. Studies by [28,29] have shown that Transformer models generally outperform recurrent neural networks (RNNs) across different imputation methods, particularly with Stineman interpolation. This suggests that while traditional models like ARIMA and GARCH are robust in many scenarios, Transformer models may offer superior performance in handling incomplete time series data, especially when advanced imputation techniques are applied.

In conclusion, the effectiveness of imputation methods in time series analysis depends on the specific characteristics of the data and the research objectives. The methods reviewed in this section provide a comprehensive framework for understanding the efficacy of different imputation techniques in various scenarios, particularly when applied to traditional time series models and Transformer models.

Although there is a wealth of research on time series forecasting and missing data imputation, several gaps remain. First, while traditional models like ARIMA and GARCH have been well-studied, their comparative performance against modern models like Transformers, especially in the context of missing data, is not fully understood. Additionally, the effectiveness of various imputation techniques when applied to these models has not been comprehensively evaluated, particularly across different levels of data completeness and sample sizes. This gap in the literature necessitates a thorough investigation into which imputation techniques best complement specific forecasting models under varying conditions.

To address these gaps, this study systematically compares the performance of ARIMA, GARCH, and Transformer models in handling missing data across different scenarios. By evaluating eight different imputation methods—ranging from conventional techniques like mean imputation and linear interpolation to more advanced methods like Kalman filtering—this research aims to identify the most effective combinations of models and imputation techniques. The study employs a comprehensive simulation approach, analyzing different sample sizes and varying levels of missing data to provide a robust framework for selecting the most appropriate methods in time series forecasting.

This study advances the existing literature by offering a detailed, comparative analysis of traditional and modern time series models in conjunction with a wide range of imputation methods. By integrating advanced techniques and considering various scenarios, this research not only fills critical gaps in the literature but also introduces a new framework for evaluating the effectiveness of different approaches to missing data in time series analysis. The findings have the potential to significantly improve the accuracy and reliability of time series forecasting, particularly in fields where data completeness is a challenge.

For this purpose, the study examines eight imputation methods through a comprehensive simulation, considering three different sample sizes and varying imputation rates. This analysis provides valuable insights into possible combinations of estimation methods and imputation techniques. The inclusion of advanced imputation techniques, such as those discussed by [5], [30,31], and [27], enriches the discussion by integrating novel approaches. Furthermore, the study compares different methods while also exploring the potential for transitioning from conventional to more sustainable methods, as discussed by [32], suggesting new avenues for enhancing imputation techniques using AI-based methods.

In the rest of the study, ARIMA and GARCH, which are traditional methods frequently used in time series analysis, and Transformer models, which are innovative methods, are presented in the Method section. Additionally, information is provided about eight different imputation methods. The structural, metric, and scenario parameters of the simulation study and the information and steps for application are introduced in the Analysis section. In the Results section, all findings are presented in detail for all cases. In the Conclusion section, the findings are discussed.

MATERIALS AND METHODS

This section outlines the methodologies employed in this study, including the models and imputation techniques used for analysis.

Estimators

The ARIMA model and the GARCH model (are popular methods for time series forecasting and volatility modeling.

ARIMA is a widely used time series analysis model for forecasting and understanding time-dependent data. The ARIMA model is denoted as ARIMA (*p, d, q*), where *p* represents the autoregressive order, *d* represents the differencing order, and *q* represents the moving average order. The model's equations involve the autoregressive terms, moving average terms, and the differencing operator, which are used to capture the temporal dependencies and trends in the data [33,34].

The ARIMA model equation can be represented in Eq.1:

$$\begin{aligned}Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} \\ + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t\end{aligned} \tag{1}$$

where Y_t is the value of the time series at time *t*. *c* is the constant term or intercept. ϕ_i ($i = 1,2,\ldots, p$) are the autoregressive parameters representing the effect of past values on the current value. Y_{t-i} are the lagged values of the time series. ϕ_j ($i = 1,2,\ldots, q$) are the moving average parameters representing the effect of past errors on the current value. ϵ_t is the error term at time *t*, assumed to be white noise with mean zero and constant variance. The *p* and *q* parameters represent the order of the autoregressive and moving average components, respectively. The Integrated (*I*) component indicates the number of differences needed to make the time series stationary.

GARCH is a model used to analyses and forecast the volatility of time series data. The GARCH model is denoted as GARCH (*p, q*), where *p* represents the order of the GARCH terms, and *q* represents the order of the ARCH terms. The model's equations involve the conditional variance, which captures the time-varying volatility in the data [35].

The basic GARCH model equation can be represented in Eq.2:

$$\sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{2}$$

where σ_t^2 is the conditional variance of the time series at time t. ω is the constant term or intercept of the GARCH model. α_1 is the coefficient of the lagged squared error term, representing the persistence of volatility shocks. $\epsilon_{t\text{-}1}^2$ is the squared error term at time t - 1. β_1 is the coefficient of the lagged conditional variance term, representing the decay of past volatility shocks. $\sigma_{t\text{-}1}^2$ is the conditional variance at time t - 1.

The Transformer model utilizes self-attention mechanisms to capture dependencies across the entire sequence [6]. The equations for the Transformer model are:

$$\begin{gathered}\left[Z_1^l = LayerNorm\left(X^{(l)} + MultiHeadAttention\left(X^{(l)}\right)\right)\right]\\ \left[Z_2^l = LayerNorm\left(Z_1^{(l)} + MultiHeadAttention\left(Z_1^{(l)}\right)\right)\right]\\ \left[Z^{(l+1)} = LayerNorm\left(Z_2^{(l)} + FeedForward\left(Z_2^{(l)}\right)\right)\right]\end{gathered} \tag{3}$$

where $Z^{(l)}$ represents the output of the l -th layer, and *LayerNorm* is a normalization layer, *MultiHeadAttention* is the multi-head attention mechanism, *FeedForward* is a feedforward neural network layer. These equations highlight the fundamental elements of the Transformer model: the multi-head attention mechanism, which enables the model to attend to various segments of the input sequence, and the feedforward neural network, which processes the aggregated attention data. Layer Normalization is employed to normalize the inputs of each layer, aiding in the stabilization and acceleration of deep neural network training.

The Transformer model is a type of neural network architecture that has been applied to time series analysis. It utilizes self-attention mechanisms to capture dependencies between different time steps in the data. The equations of the Transformer model involve the self-attention mechanism, which allows the model to weigh the importance of different time steps when making predictions [6]. An adaptation of the Transformer model equation for time series analysis includes the following components: input representation, positional encoding, transformer encoder, transformer decoder, and output layer.

Scaled dot-product Attention:

$$attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

In the equation above: Q represents the query. K represents the key. V represents the value in the attention mechanism. d_k is the dimension of the keys (K^T: The transpose of the key matrix K).

Positional Encoding for time series:

$$\begin{gathered}PE_{(pos,2i)} = sin\left(\frac{pos}{1000^{\frac{2i}{d_{model}}}}\right)\\ PE_{(pos,2i+1)} = cos\left(\frac{pos}{1000^{\frac{2i}{d_{model}}}}\right)\end{gathered} \tag{5}$$

In the equations above: $PE_{(pos,2i)}$ and $PE_{(pos,2i+1)}$ represent the positional encoding for even and odd indices, respectively. pos represents the position. d_{model} is the dimension of the model.

Imputation Techniques

The imputation techniques utilized in the comparison study include:

a. Last Observation Carried Forward (LOCF): Fills missing values with the last observed value [36].
b. Next Observation Carried Backward (NOCB): Fills missing values with the next observed value [37].
c. Mean Imputation: Replaces missing values with the mean of the available data [37].
d. Linear Interpolation: Estimates missing values based on linear interpolation between adjacent data points [38].
e. Seasonal Decomposition: Decomposes the time series into seasonal and trend components, filling missing values based on the decomposition [39].
f. Moving Average: Fills missing values with the average of neighboring data points within a specified window [40].
g. Regression Imputation: Predicts missing values using regression analysis based on available data [41].
h. Kalman Filtering: Utilizes Kalman filter algorithms to estimate missing values based on observed data and system dynamics [42].

These techniques offer diverse approaches to handling missing data in time series analysis, each with its strengths and limitations.

ANALYSIS

Simulation

An inclusive simulation study was conducted for the analysis of the study. Simulation inputs are presented in detail in Table 1. This detailed procedure outlines how to conduct the comparison of the specified models using various imputation methods on different types of data and sample sizes [43]. The RMSE (Root Mean Square Error) metric is used to evaluate the accuracy of each model under different conditions. RMSE is related to bias and variance in the context of model evaluation and is therefore considered the metric used in this study. With the abundance of scenarios and tables, all results are presented with RMSE for interpretability. The RMSE is represented as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (6)$$

where y_i is the actual value. $\hat{y}_i$ is the predicted value. n is the number of observations.

The simulation procedure systematically evaluates the performance of eight imputation techniques and various estimation models across different scenarios, including the Transformer, ARIMA, and GARCH models. The analysis stages are as follows:

- A sample time series dataset is generated for each combination of estimation model, imputation method, and sample sizes, with missing values introduced based on specified percentages and imputation methods.
- The missing values are then imputed using the chosen method, and each model is trained and evaluated using the imputed data.
- RMSE values are calculated for each model, with the process repeated for 1000 iterations to capture variability.
- The RMSE values are aggregated over the iterations to derive average RMSE values for each combination.
- Finally, comparisons are made across different models, imputation methods, data cases, and sample sizes to determine their relative performance.

All applications in this study were conducted in R using RStudio IDE. Steps of simulation is also shared in Table 2. This table outlines the simulation process, detailing each step and the corresponding operation performed. Tables 3-7 are derived from the results of the simulation study and are discussed in detail in the Results section.

Application

In addition to the simulation study, an application based on real data was conducted to evaluate the practical performance of different imputation methods and time series models. For this application, historical stock data for AAPL (Apple Inc.) was utilized. The analysis utilized data consisting of 756 observations, covering approximately three years of time series data. Thus, a supportive approach was adopted in the simulation studies with sample sizes of 200, 600, and 1000.

Table 1. Simulation arguments

Models to Compare	Imputation Methods	Imputation Methods	Data Cases	Sample	Evaluation Metric	Simulation Iterations
Transformer Model	Last Observation Carried Forward (LOCF)	Seasonal Decomposition	Original Data	200	Root Mean Squared Error (RMSE)	1000 iterations
ARIMA (AutoRegressive Integrated Moving Average) Model	Next Observation Carried Backward (NOCB)	Moving Average	10% Imputed Data	600		
GARCH (Generalized Autoregressive Conditional Heteroskedasticity) Model	Mean Imputation	Regression Imputation	25% Imputed Data	1000		
	Linear Interpolation	Kalman Filtering	40% Imputed Data			

Table 2. Simulation Steps

Steps	Description	Operation
1	Load Libraries	\`library(forecast)\`, \`library(fGarch)\`, etc.
2	Set Parameters	Iterations, repetitions, sample widths, imputation percentages, methods, model types
3	Generate Time Series Data	\`generate_time_series(n)\`
4	Introduce Missing Values	\`introduce_missing_values(ts_data, missing_percentage)\`
5	Impute Missing Values	\`impute_missing_values(ts_data, method)\`
6	Train and Evaluate Models	\`train_and_evaluate_models(ts_data, model_type)\`
7	Store Results	Store RMSE values in the matrix
8	Calculate Average RMSE and Output Results	\`aggregate(RMSE ~ Sample_Width + Imputation_Percentage + Imputation_Method + Model_Type, data = results_df, FUN = mean)\`

The analysis began with loading essential R libraries. Historical adjusted closing prices were obtained. To simulate missing data based on randomness, function based on uniform distribution was applied, creating data gaps at levels of 0%, 10%, 25%, and 40%. Eight imputation techniques were employed, such as LOCF, NOCB, mean imputation, linear interpolation, seasonal decomposition, moving average, regression imputation, and Kalman filtering. The data was then used to fit three models: The Transformer model, ARIMA, and GARCH. The performance of each model was assessed by calculating the RMSE for every combination of imputation method and missing data level. Results were systematically compiled into a table, presented as Table 8, and the findings are discussed in the results section. The application provides insight into the extent to which the simulation results align with those obtained from real-world data analysis and highlights the implications of the study's findings for practical applications.

RESULTS AND DISCUSSION

To analyses the results, we can examine the RMSE values for each combination of model, imputation method, and missing data percentage.

Table 3 shows the analysis results for the case where the number of samples is 200. In terms of estimators, across all missing data percentages (0%, 10%, 25%, 40%), the Transformer model generally performs better than ARIMA and GARCH models in terms of RMSE. This suggests that the Transformer model is more effective in predicting time series data compared to traditional ARIMA and GARCH models. Regarding imputation methods, among the imputation methods, the performance varies depending on the

Table 3. Simulation RMSE Results for *n*=200*

***n* = 200**	**0%**			**10%**			**25%**			**40%**		
	TraMod	**ARIMA**	**GARCH**	**TraMod**	**ARIMA**	**GARCH**	**TraMod**	**ARIMA**	**GARCH**	**TraMod**	**ARIMA**	**GARCH**
LOCF	0.5314	0.9403	0.9328	0.5726	0.9953	1.0475	0.5952	0.9976	1.0910	0.6004	1.0119	1.0990
NOCB	0.5271	0.9403	0.9328	0.5109	0.9900	1.0108	0.6005	0.9922	1.0135	0.7010	0.9923	1.0246
Mean	0.5191	0.9403	0.9328	0.5074	0.9462	0.9475	0.5686	0.9462	0.9494	0.6429	0.9503	0.9740
Linear	0.5296	0.9403	0.9328	0.6148	0.9595	0.9888	0.6162	0.9684	1.0271	0.6457	0.9711	1.0505
Seasonal	0.5192	0.9403	0.9328	0.5577	0.9664	0.9956	0.5683	0.9671	1.0368	0.6043	0.9706	1.0445
MovAve	0.5298	0.9403	0.9328	0.5952	0.9509	0.9524	0.6055	0.9534	0.9661	0.6160	0.9593	0.9777
Regression	0.5217	0.9403	0.9328	0.4839	0.9997	1.0615	0.5423	1.0009	1.0889	0.6083	1.0202	1.1433
KalmFil	0.5299	0.9403	0.9328	0.5654	0.9456	0.9470	0.5727	0.9471	0.9566	0.5800	0.9547	0.9592

* LOCF: Last Observation Carried Forward, NOCB: Next Observation Carried Backward, Mean: Mean Imputation, Linear: Linear Interpolation, Seasonal: Seasonal Decomposition, MovAve: Moving Average Imputation, Regression: Regression Imputation, KalmFil: Kalman Filter Imputation; TraMod: Transformer Model, ARIMA: AutoRegressive Integrated Moving Average, GARCH: Generalized AutoRegressive Conditional Heteroskedasticity.

Table 4. Simulation RMSE Results for *n*=600*

***n* = 600**	**0%**			**10%**			**25%**			**40%**		
	TraMod	**ARIMA**	**GARCH**	**TraMod**	**ARIMA**	**GARCH**	**TraMod**	**ARIMA**	**GARCH**	**TraMod**	**ARIMA**	**GARCH**
LOCF	0.4665	0.8377	0.8264	0.5328	0.9532	1.0015	0.5330	0.9621	1.0170	0.5451	0.9926	1.0253
NOCB	0.4596	0.8377	0.8264	0.3784	0.9454	0.9714	0.3796	0.9492	0.9755	0.4864	0.9875	0.9883
Mean	0.4613	0.8377	0.8264	0.4619	0.8550	0.8587	0.4702	0.8624	0.8665	0.5045	0.8995	0.9193
Linear	0.4433	0.8377	0.8264	0.5564	0.8824	0.9696	0.5668	0.9060	0.9787	0.6051	0.9144	0.9826
Seasonal	0.4329	0.8377	0.8264	0.4117	0.8805	0.9364	0.4601	0.8912	0.9719	0.5281	0.9131	0.9887
MovAve	0.4307	0.8377	0.8264	0.4875	0.8614	0.8674	0.5334	0.8770	0.8903	0.5698	0.9056	0.9214
Regression	0.4327	0.8377	0.8264	0.4408	0.9627	1.0148	0.4487	0.9658	1.0375	0.4545	0.9996	1.0486
KalmFil	0.4371	0.8377	0.8264	0.5413	0.8590	0.8563	0.5556	0.8778	0.8734	0.5616	0.8810	0.8806

* LOCF: Last Observation Carried Forward, NOCB: Next Observation Carried Backward, Mean: Mean Imputation, Linear: Linear Interpolation, Seasonal: Seasonal Decomposition, MovAve: Moving Average Imputation, Regression: Regression Imputation, KalmFil: Kalman Filter Imputation; TraMod: Transformer Model, ARIMA: AutoRegressive Integrated Moving Average, GARCH: Generalized AutoRegressive Conditional Heteroskedasticity.

Table 5. Simulation RMSE Results for n=1000*

***n* = 1000**	**0%**			**10%**			**25%**			**40%**		
	TraMod	**ARIMA**	**GARCH**	**TraMod**	**ARIMA**	**GARCH**	**TraMod**	**ARIMA**	**GARCH**	**TraMod**	**ARIMA**	**GARCH**
LOCF	0.3287	0.7119	0.6952	0.3741	0.8945	0.9362	0.4162	0.9164	0.9551	0.4729	0.9274	0.9644
NOCB	0.3427	0.7119	0.6952	0.2899	0.9038	0.8808	0.3229	0.9136	0.9193	0.3236	0.9175	0.9203
Mean	0.3337	0.7119	0.6952	0.3815	0.7459	0.7310	0.3953	0.7764	0.7774	0.4616	0.7805	0.7917
Linear	0.3534	0.7119	0.6952	0.3782	0.7601	0.5453	0.3934	0.8124	0.9309	0.8675	0.8171	0.8968
Seasonal	0.3672	0.7119	0.6952	0.3086	0.7979	0.7566	0.3712	0.8016	0.8816	0.3988	0.8098	0.9337
MovAve	0.3202	0.7119	0.6952	0.3124	0.7746	0.7987	0.4476	0.7886	0.8073	0.4637	0.800	0.8491
Regression	0.3657	0.7119	0.6952	0.2328	0.8866	0.9716	0.3147	0.9442	0.9831	0.3873	0.9515	1.0092
KalmFil	0.3257	0.7119	0.6952	0.4120	0.7689	0.7626	0.4930	0.7773	0.7802	0.5124	0.7814	0.7843

* LOCF: Last Observation Carried Forward, NOCB: Next Observation Carried Backward, Mean: Mean Imputation, Linear: Linear Interpolation, Seasonal: Seasonal Decomposition, MovAve: Moving Average Imputation, Regression: Regression Imputation, KalmFil: Kalman Filter Imputation; TraMod: Transformer Model, ARIMA: AutoRegressive Integrated Moving Average, GARCH: Generalized AutoRegressive Conditional Heteroskedasticity.

combination of model and missing data percentage. For example, the Kalman filter, mean, and regression imputations tend to perform relatively well across different models and missing data percentages. However, the effectiveness of imputation methods can vary based on the specific characteristics of the data and the modeling approach.

Table 4 shows the analysis results for the case where the number of samples is 600. Similar to the previous analysis, the Transformer model generally outperforms ARIMA and GARCH models across different missing data percentages (0%, 10%, 25%, 40%). This consistency suggests that the superiority of the Transformer model in predicting time series data is robust and not heavily influenced by missing data. The performance of imputation methods varies across different models and missing data percentages. For instance, Kalman filter, mean imputation, and NOCB imputations show relatively stable performance across various scenarios, indicating their effectiveness in handling missing data in time series analysis. However, some methods like LOCF, linear interpolation, seasonal decomposition, moving average exhibit fluctuating performance depending on the combination of model and missing data percentage. As observed in the previous analysis, higher percentages of missing data lead to higher RMSE values across all models and imputation methods. This consistent trend emphasizes the detrimental effect of missing data on the accuracy of time series predictions.

Table 5 shows the analysis results for the case where the number of samples is 1000. The RMSE values for the Transformer model are consistently lower compared to ARIMA and GARCH models, indicating better predictive performance. For this scenario, some imputation methods, such as mean, and regression imputations show relatively stable performance across different scenarios. As observed in previous analyses, higher percentages of missing data lead to higher RMSE values across all models and imputation methods. This trend underscores the importance of handling missing data effectively in time series analysis to maintain prediction accuracy.

To examine the estimation model, Table 6 provides useful summary information. The Transformer model has been combined to present the lowest RMSE value 5 times with regression imputation, 3 times with NOCB imputation and 1 time with Kalman filter. Therefore, regression imputation and NOCB imputation are compatible with the Transformer model. The ARIMA model has been combined to present the lowest RMSE value 7 times with mean imputation, 2 times with Kalman filter. Therefore, mean imputation and Kalman filter are compatible with the ARIMA model. The GARCH model has been combined to present the lowest RMSE value 5 times with Kalman filter imputation, 3 times with mean imputation and 1 time with linear interpolation. Therefore, Kalman filter and mean imputation are compatible with the ARIMA model.

To examine sample sizes, Table 7 provides useful summary information. LOCF was not applied to any of the datasets. NOCB was applied 2 times to the dataset with size 600 and once to the dataset with size 1000 but was not applied to the dataset with size 200. Mean imputation was applied 3 times to the dataset with size 200, 3 times to the dataset with size 600, and 4 times to the dataset with size 1000, indicating consistent usage across all dataset sizes. Linear interpolation was applied once to the dataset with size 1000 and not to the datasets with sizes 200 or 600. Seasonal decomposition and moving average were not applied to any of the datasets. Regression imputation was applied 2 times to the dataset with size 200, once to the dataset with size 600, and 2 times to the dataset with size 1000, showing moderate usage. Kalman filter was applied 4 times to the dataset with size 200, 3 times to the dataset with size 600, and once to the dataset with size 1000, indicating frequent usage, especially for smaller datasets. In general, mean and Kalman filter imputation methods had high usage, NOCB and regression imputation had moderate usage, linear interpolation had

Table 6. Lowest RMSE (Best Combination) Counts according to the Estimation Methods

	TraMod	ARIMA	GARCH	Comments
LOCF				No counts provided
NOCB	3			NOCB has 3 counts for TraMod
Mean		7	3	Mean has 7 counts for ARIMA and 3 for GARCH
Linear			1	Linear has 1 count for GARCH
Seasonal				No counts provided
MovAve				No counts provided
Regression	5			Regression has 5 counts for TraMod
KalmFil	1	2	5	Kalman Filter has 1 count for TraMod, 2 for ARIMA, and 5 for GARCH

* LOCF: Last Observation Carried Forward, NOCB: Next Observation Carried Backward, Mean: Mean Imputation, Linear: Linear Interpolation, Seasonal: Seasonal Decomposition, MovAve: Moving Average Imputation, Regression: Regression Imputation, KalmFil: Kalman Filter Imputation; TraMod: Transformer Model, ARIMA: AutoRegressive Integrated Moving Average, GARCH: Generalized AutoRegressive Conditional Heteroskedasticity.

Table 7. Lowest RMSE (Best Combination) Counts according to the Sampe Sizes

	200	600	1000	Comments
LOCF				Not applied
NOCB		2	1	Moderately used, preferred in larger datasets
Mean	3	3	4	Consistently used across all dataset sizes
Linear			1	Limited use, only in the largest dataset
Seasonal				Not applied
MovAve				Not applied
Regression	2	1	2	Moderately used, shows balanced usage across sizes
KalmFil	4	3	1	Frequently used, especially in smaller datasets

* LOCF: Last Observation Carried Forward, NOCB: Next Observation Carried Backward, Mean: Mean Imputation, Linear: Linear Interpolation, Seasonal: Seasonal Decomposition, MovAve: Moving Average Imputation, Regression: Regression Imputation, KalmFil: Kalman Filter Imputation; 200, 600, and 1000 represent sample sizes.

low usage, and LOCF, seasonal decomposition, and moving average were not used. These observations can help in determining the most suitable imputation methods based on dataset size and missing data patterns. The frequency values in Table 6 and Table 7 are presented comparatively in Figure 1.

The combination of two plots in Figure 1 offers a comprehensive overview of how imputation methods are distributed across both model types and dataset sizes. The consistent use of visual elements such as transparency for zero counts, color differentiation, and well-placed labels ensures clarity and readability, facilitating quick comparison across different categories. For instance, LOCF, Moving Average, and Seasonal Decomposition are absent in both figures, confirming they were not used for any dataset. On the other hand, it is seen that Kalman filtering and Mean methods dominate the process.

Considering the values of RMSE values in the table and the fact that there are many scenarios and combinations, graphs of imputation techniques and estimation methods for sample sizes of 200, 600 and 1000 are presented in Figures 2-4 for clarity.

The Transformer model is superior to the ARIMA and GARCH models in complete data (original data) and missing data completion scenarios at all imputation rates. Moreover, the transformer model was able to provide a lower RMSE value than the original data in case of 10% imputation. This indicates the success of imputation techniques in completing missing data at low imputation rates.

Although the GARCH model had superior results compared to the ARIMA model in the absence of imputation, in the case of imputation, the ARIMA model performed better than the GARCH model for all imputation rates. However, contrary to this general trend, the GARCH model outperformed ARIMA at all imputation rates when imputed with Kalman Filtering under the n = 600 scenario. The GARCH model again performed better than ARIMA when imputed with NOCB, Mean, Linear, Seasonal Decomposition, and Kalman filter imputation methods under the n = 1000 scenario.

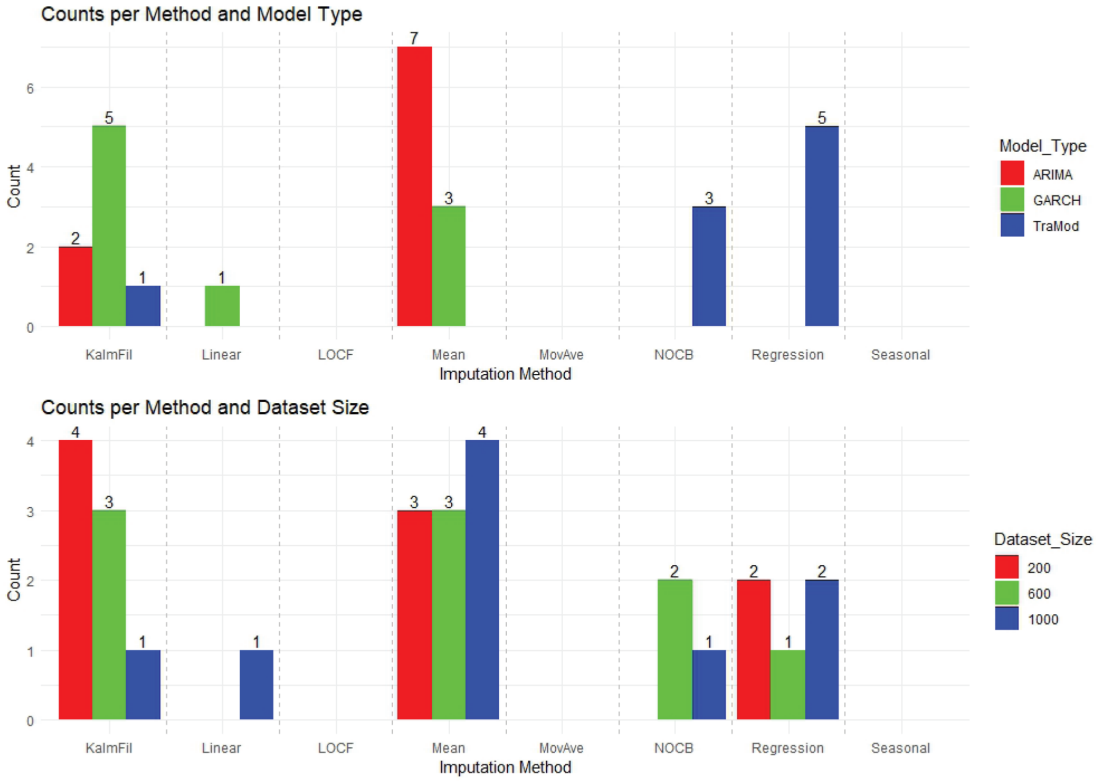

Figure 1. Counts for model types and datasets.

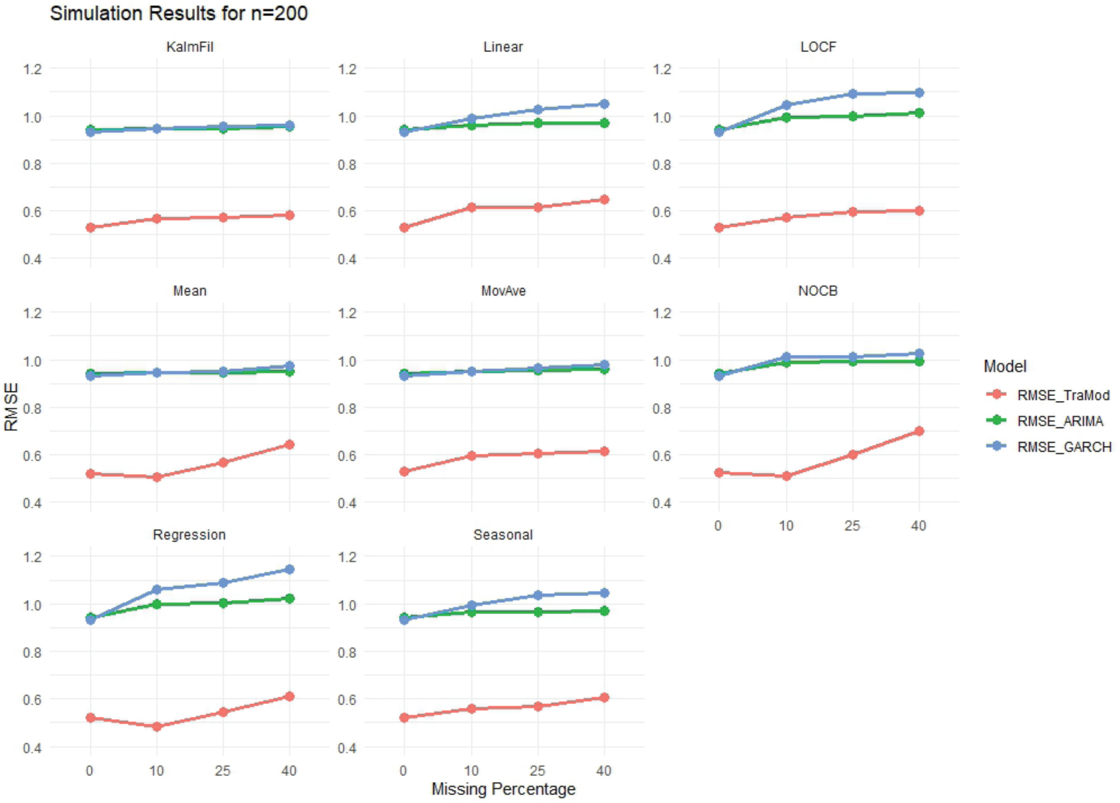

Figure 2. RMSE results for n=200.

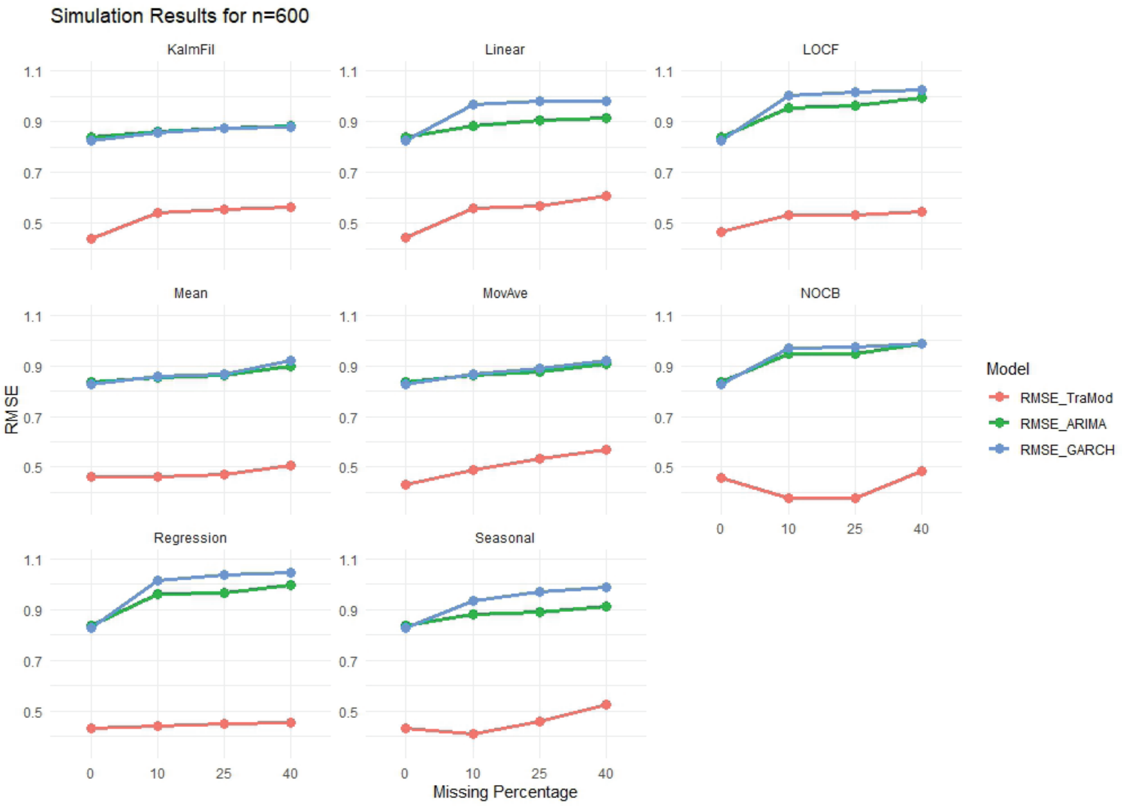

Figure 3. RMSE results for n=600.

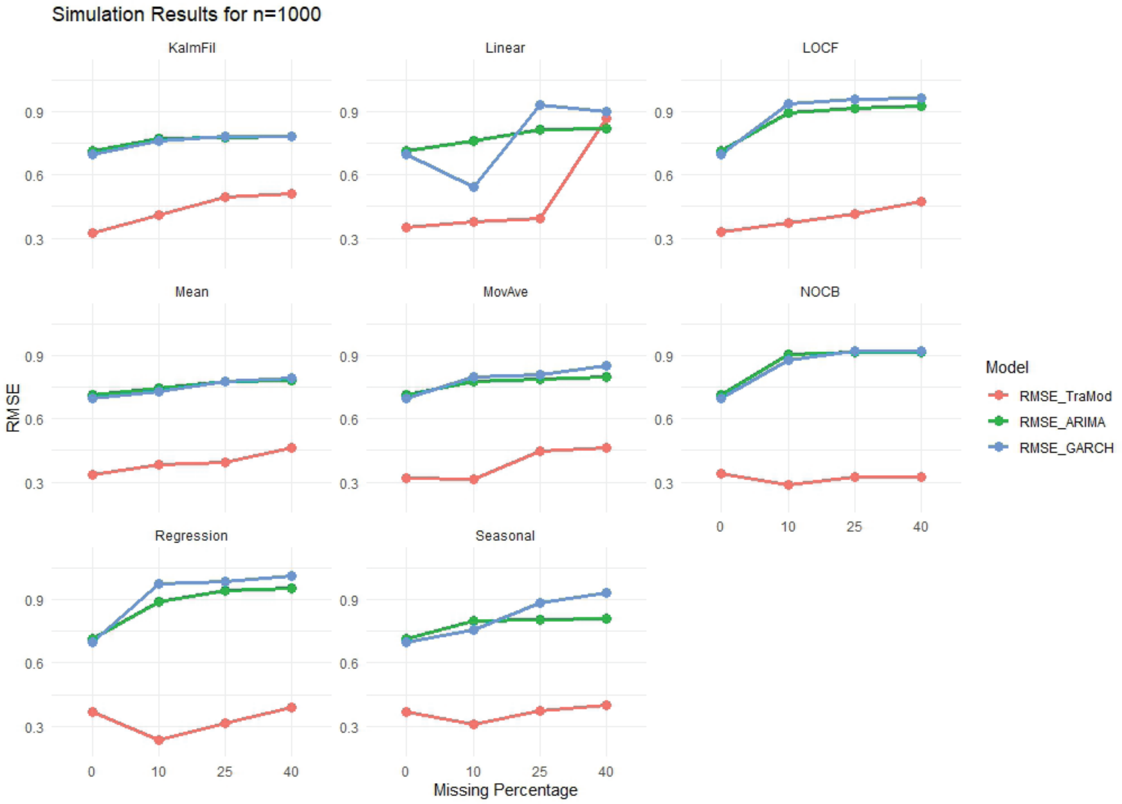

Figure 4. RMSE results for n=1000.

Table 8. RMSE Results for AAPL

n **= 600**	**0%**			**10%**			**25%**			**40%**		
	TraMod	**ARIMA**	**GARCH**	**TraMod**	**ARIMA**	**GARCH**	**TraMod**	**ARIMA**	**GARCH**	**TraMod**	**ARIMA**	**GARCH**
LOCF	2.7378	2.7432	12.8822	2.8462	2.8486	12.8954	3.0179	3.0193	12.9557	3.2425	3.2465	12.9787
NOCB	2.7378	2.7432	12.8822	2.8327	2.8349	12.8901	2.9586	2.9694	12.9049	3.6613	3.7292	12.9798
Mean	2.7378	2.7432	12.8822	2.7783	2.7906	12.9274	3.0725	3.0753	12.9826	3.9062	3.9561	12.9901
Linear	2.7378	2.7432	12.8822	2.7818	2.7936	12.9147	2.8193	2.8219	12.9366	2.9175	2.9812	12.9518
Seasonal	2.7378	2.7432	12.8822	2.7822	2.7963	12.9710	2.8219	2.8891	12.9846	2.8378	2.9012	12.9977
MovAve	2.7378	2.7432	12.8822	2.7957	2.7973	12.9122	2.8416	2.8529	12.9309	2.8870	2.9666	12.9494
Regression	2.7378	2.7432	12.8822	2.7928	2.8028	12.9132	2.8699	2.9871	12.9296	2.8862	3.0780	12.9809
KalmFil	2.7378	2.7432	12.8822	2.8041	2.8818	12.9060	2.8144	2.8203	12.9371	2.8329	2.8929	12.9490

* LOCF: Last Observation Carried Forward, NOCB: Next Observation Carried Backward, Mean: Mean Imputation, Linear: Linear Interpolation, Seasonal: Seasonal Decomposition, MovAve: Moving Average Imputation, Regression: Regression Imputation, KalmFil: Kalman Filter Imputation; TraMod: Transformer Model, ARIMA: AutoRegressive Integrated Moving Average, GARCH: Generalized AutoRegressive Conditional Heteroskedasticity

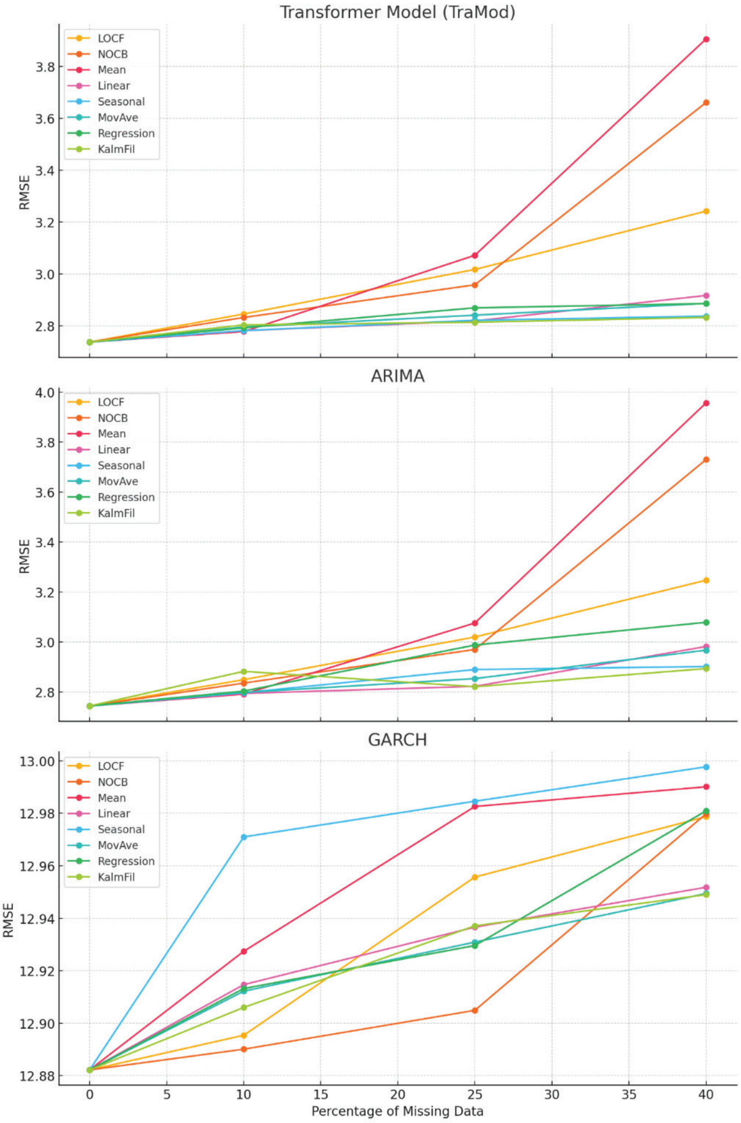

Figure 5. RMSE results for AAPL.

In the application, the Transformer model consistently outperforms the ARIMA and GARCH models across different imputation methods and missing data levels, highlighting its effectiveness in dealing with incomplete data.

For low levels of missing data (10%), mean imputation performs best with both the Transformer model and ARIMA, suggesting that conventional imputation methods are effective when data completeness is relatively high. However, as the level of missing data increases to medium (25%) and high (40%) levels, Kalman Filtering emerges as the superior imputation method, providing the most accurate results across all models. Specifically, at high levels of missing data, Kalman Filtering shows the best performance with the GARCH model. On the other hand, for lower and medium missing data levels, NOCB is particularly effective with the GARCH model, outperforming other imputation methods.

The results in Table 8 and their corresponding analyses are more easily interpreted with the graphs in Figure 5. Figure 5 presents RMSE results for the estimation models and imputation techniques across different levels of missing data in the AAPL dataset. The missing data levels are represented as percentages (0%, 10%, 25%, 40%). The RMSE values are shown for three forecasting models: the Transformer model, ARIMA, and GARCH. These models are assessed using eight imputation methods: LOCF, NOCB, Mean, Linear Interpolation, Seasonal Decomposition, Moving Average, Regression, and Kalman Filtering.

In the case of imputing 10% and 25% missing data, the Transformer model and ARIMA produce close and consistent RMSE values. The differences between imputation methods become more pronounced with higher missing data levels. However, this pattern exhibits more volatility for the GARCH model. Additionally, the results from the Transformer model and ARIMA show close alignment. Moreover, as the level of missing data increases (from 10% to 40%), there is a noticeable increase in RMSE across all models and imputation methods. This trend is consistent, demonstrating that higher missing data levels generally degrade model performance.

CONCLUSION

The results show that the choice of imputation technique significantly influences the accuracy of predictions in time series analysis. Techniques such as mean imputation and Kalman filter imputation have shown reliability and effectiveness across different model types, while methods like LOCF, seasonal decomposition, and moving average were less utilized, potentially due to their unsuitability for the given data. The Transformer model shows promise as a predictive modeling technique, consistently outperforming traditional ARIMA and GARCH models in various scenarios. The Transformer's self-attention mechanism allows it to capture long-term dependencies more effectively than ARIMA and GARCH models. Results emphasize the importance of careful consideration when dealing with missing data and selecting appropriate imputation methods to enhance the accuracy of time series analysis.

Higher percentages of missing data imputation generally lead to higher RMSE values across all models and imputation methods. This indicates that imputation rate can significantly affect the accuracy of time series predictions, regardless of the modeling approach used. The decrease in RMSE values as the number of samples increases is an important indicator of the consistency of the analyses, aligning theoretical expectations with empirical observations. Specifically, at low levels of missing data (10%), mean imputation is highly effective with both the Transformer model and ARIMA, suggesting that traditional imputation methods work well when data is mostly complete. However, as the proportion of missing data increases (25% and 40%), Kalman filtering becomes the superior method, yielding the most accurate predictions across all models. The GARCH model shows more volatility in performance compared to Transformer and ARIMA, with Kalman filtering particularly beneficial at high missing data levels. Interestingly, NOCB imputation performs well with the GARCH model at lower and medium missing data levels.

Despite the varied application of techniques, this study reflects an exploratory approach to determine the best imputation method for different model types and datasets. These insights can guide future imputation strategy choices, emphasizing techniques that proved useful and exploring underutilized methods to potentially enhance model performance.

The Transformer model's superiority is evident across all scenarios, and Kalman filtering emerges as the most reliable imputation method as missing data levels increase. One potential limitation of this study is the computational complexity and scalability of the Transformer model. Transformer's self-attention mechanism, while powerful, can be computationally intensive, making it less feasible for very large datasets or in resource-constrained environments. Future research should identify challenges with time counters and real root values; and investigate optimizations and alternatives for the solution. Additionally, it would be beneficial to further explore and evaluate the performance of underutilized imputation methods, such as seasonal decomposition and moving average, to determine their potential effectiveness in different contexts. Investigating the impact of varying sample sizes on the robustness of imputation methods and predictive models can also provide deeper insights.

ACKNOWLEDGEMENTS

The previous abstract version of the study was presented at the "9th International Conference on Business Management and Economics" as an oral presentation with the title "Imputation Methods Effect on the Goodness of Fit of the Statistical Model" [43]. This study was supported by

Eskişehir Technical University Scientific Research Project Commission under grant no: 24ADP185.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

REFERENCES

[1] Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. 2nd ed. Melbourne: OTexts; 2018. [CrossRef]

[2] Engle RF. GARCH 101: The use of ARCH/GARCH models in applied econometrics. J Econ Perspect 2001;15:157–168. [CrossRef]

[3] Zhao L, Wen X, Wang Y, Shao Y. A novel hybrid model of ARIMA-MCC and CKDE-GARCH for urban short-term traffic flow prediction. IET Intell Transp Syst 2022;16:206–217. [CrossRef]

[4] Devianto D, Yollanda M, Maiyastri M, Yanuar F. The soft computing FFNN method for adjusting heteroscedasticity on the time series model of currency exchange rate. Front Appl Math Stat 2023;9:1045218. [CrossRef]

[5] Chandola A, Pandey RM, Agarwal R, Rathour L, Mishra VN. On some properties and applications of the generalized m-parameter Mittag-Leffler function. Adv Math Model Appl 2022;7:130–145.

[6] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. 2017:5998–6008.

[7] Efimova O, Serletis A. Energy markets volatility modeling using GARCH. Energy Econ 2014;43:264–273. [CrossRef]

[8] Zhou B, He D, Sun Z. Traffic modeling and prediction using ARIMA/GARCH model. In: Modeling and Simulation Tools for Emerging Telecommunication Networks. 2006:101–121. [CrossRef]

[9] Li J, Yin J, Zhang R. Analysis and forecast of USD/EUR exchange rate based on ARIMA and GARCH models. In: Li X, Yuan C, Kent J, (editors). Proceedings of the 7th International Conference on Economic Management and Green Development (ICEMGD 2023). Midtown Manhattan, New York City: Springer; 2024. p. 566–575. [CrossRef]

[10] Yaziza SR, Azizanb NA, Zakariaa R, Ahmadc M. The performance of hybrid ARIMA-GARCH modeling in forecasting gold price. In proceedings of the 20th International Congress on Modelling and Simulation. 2013. p. 1201–1207.

[11] Adebiyi AA, Adewumi AO, Ayo CK. Comparison of ARIMA and artificial neural networks models for stock price prediction. J Appl Math 2014;2014:614342. [CrossRef]

[12] Valipour M, Banihabib ME, Behbahani SMR. Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. J Hydrol 2013;476:433–441. [CrossRef]

[13] Wang Y, Shen Z, Jiang Y. Comparison of autoregressive integrated moving average model and generalized regression neural network model for prediction of haemorrhagic fever with renal syndrome in China: a time-series study. BMJ Open 2019;9:e025773. [CrossRef]

[14] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. In NIPS 2014 Workshop on Deep Learning, December 2014. 2014.

[15] Dauphin YN, Fan A, Auli M, Grangier D, Precup D, Teh YW. Language modeling with gated convolutional networks. In: Proceedings of the 34th International Conference on Machine Learning 2017;70:933–941.

[16] Wen Q, Zhou T, Zhang C, Chen W, Ma Z, Yan J, Sun L. Transformers in time series: a survey. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence 2023. p. 6778–6786. [CrossRef]

[17] Waljee A, Mukherjee A, Singal A, Zhang Y, Warren J, Balis U, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open 2013;3:e002847. [CrossRef]

[18] Mohamed C, Sedory S, Singh S. Improved mean methods of imputation. Statistics Optim Inf Comput 2018;6:526–535. [CrossRef]

[19] Dokumentov A, Hyndman R. STR: Seasonal-trend decomposition using regression. INFORMS J Data Sci 2022;1:50–62. [CrossRef]

[20] Qin Y, Zhang S, Zhu X, Zhang J, Zhang C. Semi-parametric optimization for missing data imputation. Appl Intell 2007;27:79–88. [CrossRef]

[21] Amini A, Thevenaz P, Ward J, Unser M. On the linearity of Bayesian interpolators for non-Gaussian

continuous-time AR(1) processes. IEEE Trans Inf Theory 2013;59:5063−5074. [CrossRef]

[22] Raubitzek S, Neubauer T, Friedrich J, Rauber A. Interpolating strange attractors via fractional Brownian bridges. Entropy. 2022;24:718. [CrossRef]

[23] Sharma MK, Chaudhary S, Rathour L, Mishra VN. Modified genetic algorithm with novel crossover and mutation operator in traveling salesman problem. Sigma J Eng Nat Sci 2024;42:1876−1883. [CrossRef]

[24] Negero NT, Duressa GF, Rathour L, Mishra VN. A novel fitted numerical scheme for singularly perturbed delay parabolic problems with two small parameters. Partial Differ Equ Appl Math 2023;8:100546. [CrossRef]

[25] Hogeme MS, Woldaregay MM, Rathour L, Mishra VN. A stable numerical method for singularly perturbed Fredholm integro-differential equation using exponentially fitted difference method. J Comput Appl Math 2024;441:115709. [CrossRef]

[26] Mishra LN, Raiz M, Rathour L, Mishra VN. Tauberian theorems for weighted means of double sequences in intuitionistic fuzzy normed spaces. Yugoslav J Oper Res 2022;32:377−388. [CrossRef]

[27] Rathour L, Singh V, Yadav H, Sharma MK, Mishra VN. A dual hesitant fuzzy set theoretic approach in fuzzy reliability analysis of a fuzzy system. Inf Sci Lett 2024;13:433−440. [CrossRef]

[28] Atmaca K, Yenilmez I. RNNs and transformer model in case of incomplete time series. In: V. International Applied Statistics Congress (UYIK - 2024); Istanbul, Turkiye; May 21−23, 2024.

[29] Yenilmez I, Atmaca K. Performance of deep learning models on imputed time series data: a simulation study and application to leading airline companies' stock price. Int J Adv Eng Pure Sci 2025;37(Suppl):30−39. [CrossRef]

[30] Sharma MK, Sadhna, Bhargava AK, Kumar S, Rathour L, Mishra LN, Pandey S. A fermatean fuzzy ranking function in optimization of intuitionistic fuzzy transportation problems. Adv Math Models Appl 2022;7:191−204.

[31] Sharma MK, Dhiman N, Kumar S, Rathour L, Mishra VN. Neutrosophic Monte Carlo simulation approach for decision making in medical diagnostic process under uncertain environment. Int J Neutrosophic Sci 2023;22:8−16. [CrossRef]

[32] Soares G, Chavarette F, Gonçalves A, Faria H, Outa R, Mishra V. Optimizing the transition: replacing conventional lubricants with biological alternatives through artificial intelligence. J Appl Comput Mech. 2024;1−9.

[33] Hassan J. ARIMA and regression models for prediction of daily and monthly clearness index. Renew Energy 2014;68:421−427. [CrossRef]

[34] Yang B, Bao W, Chen Y. Time series prediction based on complex-valued S-system model. Complexity. 2020;2020:1−13. [CrossRef]

[35] Ali M, Yusof KM, Wilson B, Ziegelmueller C. Traffic speed prediction using GARCH-GRU hybrid model. IET Intell Transp Syst 2023;17:2300−2312. [CrossRef]

[36] Streiner DL. The case of the missing data: methods of dealing with dropouts and other research vagaries. Can J Psychiatry 2002;47:68−75. [CrossRef]

[37] Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2nd ed. New York: Wiley Interscience; 2002. [CrossRef]

[38] Bloomfield P. Fourier Analysis of Time Series: An Introduction. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc.; 2000. [CrossRef]

[39] Cleveland RB, Cleveland WS, McRae JE, Terpenning I. STL: A seasonal-trend decomposition procedure based on LOESS. J Official Stat 1990;6:3−33.

[40] Brockwell PJ, Davis RA. Introduction to Time Series and Forecasting. New York, NY: Springer; 2016. [CrossRef]

[41] Allison PD. Missing Data. Thousand Oaks, CA: Sage Publications; 2001.

[42] Kalman RE. A new approach to linear filtering and prediction problems. J Basic Eng 1960;82:35−45. [CrossRef]

[43] Yenilmez I. Imputation methods effect on the goodness of fit of the statistical model. In: Proceedings of the 9th International Conference on Business, Management and Economics (9th ICBMECONF). March 1-3, 2024; Vienna, Austria.