Managing Missing Data and Predictions in Short Time Series

Francisco António[1] and Luís Cavique[2(✉)]

[1] Faculty of Engineering, Lúrio University, Pemba, Mozambique
francisco.araujo@unilurio.ac.mz
[2] Universidade Aberta and Lasige-FCUL, Lisbon, Portugal
luis.cavique@uab.pt

Abstract. Sales forecasting in the presence of Missing Data poses significant challenges, particularly for short time series where limited observations amplify the impact of incomplete records. This study analyzes a real-world transactional dataset (2021–2024) to predict quantities and prices for 2025. We classify missingness patterns and mechanisms (MCAR, MAR, MNAR) to inform the selection of imputation strategies. We evaluate techniques including MICE, Mean, KNN, and Linear Regression under simulated missingness rates, with KNN emerging as the most robust for the MAR mechanism. Regarding very short-term series predictions, the naive forecast Max2 (maximum of the last two observed values) outperformed moving averages. The results highlight the importance of mechanism-aware imputation and domain-tailored forecasting in sparse datasets. This work presents a practical framework for businesses to effectively utilize incomplete sales data.

Keywords: missing data · time series forecasting · imputation techniques · sales prediction · short time series

1 Introduction

Missing or incomplete data frequently hinders accurate forecasting in short time series contexts. These limitations impact the quality and reliability of models, particularly in sales environments where predictions are based on sparse or inconsistent historical records. This paper explores the mechanisms underlying Missing Data (MD), also reported as Missing Values (MVs), assesses their impact on predictive modeling, and evaluates imputation strategies in terms of accuracy and robustness.

Sales forecasting is crucial for optimizing inventory, allocating resources effectively, and informing strategic planning. However, real-world sales data often suffers from incompleteness due to operational gaps, entry errors, or systemic collection issues. Short time series—familiar in niche markets or for seasonal products—are especially vulnerable, as even minor missing data can distort trends and degrade model accuracy. While existing literature offers solutions for large datasets (e.g., ARIMA, LSTM), few address the unique constraints of sparse, incomplete time series.

J. Valente de Oliveira et al. (Eds.): EPIA 2025, LNAI 16121, pp. 282–294, 2026.
https://doi.org/10.1007/978-3-032-05176-9_22

This study bridges this gap by *(i) Diagnosing Missingness*, applying Rubin's framework (MCAR, MAR, MNAR) and Little's test to a real sales dataset (1,402 records, 2021–2024), *(ii) Evaluating Imputation*, comparing Multiple Imputation by Chained Equations (MICE), k-Nearest Neighbors (KNN), Mean and Linear Regression (LReg) under varying missingness rates (1–12%) using RMSE and MAE, and *(iii) Forecasting with Sparsity*, proposing a heuristic Max2 model for short series and validating it against moving averages. Given a real-world transactional dataset (2021–2024) with metadata (MD), the goal is to predict sales for 2025.

Following a review of MD types—patterns, mechanisms, and treatment methods—we apply imputation techniques and predictive models on a real-world sales dataset. The study leverages the Knowledge Discovery in Databases (KDD) methodology to guide data preparation and modeling efforts. Our main contributions include an empirical comparison of multiple imputation methods, and the application of a novel heuristic forecasting technique tailored to short time series with missing data (MD).

The remainder of the paper is structured as follows. Section 2 reviews related work. Section 3 presents the proposed model for managing MD and predicting short-term time series. Sections 4, 5 and 6 outline the methodology, which includes data cleaning, data prediction, and data summarization. Finally, Sect. 7 concludes.

2 Related Work

This section explores key aspects of MD relevant to short-time series forecasting. The discussion is structured around three central questions:

- Patterns: What data is missing?
- Mechanisms: Why is the data missing?
- Techniques: How can missing data be handled?

These elements provide a conceptual framework for understanding and addressing MD in sales and other time series contexts.

2.1 Patterns: What Data is Missing?

According to Newman [1] missing data is a statistical issue that arises when parts of a data matrix are incomplete, typically because respondents fail to provide information for one or more variables. Although the terms are often used interchangeably, MD patterns specifically refer to the structural arrangement of observed and missing data across a dataset. These patterns are essential for diagnosing data quality and guiding appropriate imputation strategies [2].

Newman [1] identifies three distinct levels of missingness: item-level, where MD occur when an individual skips one or more questions within a set of items (e.g., due to confusion or irrelevance); construct-level, where all items related to a specific construct are left blank, omitting entire sections of a survey or dataset; and person-level, where a complete absence of data for an individual results in no variables being filled in at all.

Recognizing these patterns helps determine whether missingness is isolated or systemic, which is crucial for selecting a treatment strategy.

2.2 Mechanisms: Why is the Data Missing?

Rubin's (1976) framework classifies MD mechanisms into three categories [3–5]. Missing Completely at Random (MCAR) is the probability of missingness independent of both observed and unobserved data. In this ideal case, the missingness is purely random. Missing at Random (MAR) is systematically related to other observed variables but not to the MD itself. Missing Not at Random (MNAR). The likelihood of missingness is related to the unobserved value, making the data non-ignorable and more challenging to handle. In practical applications, determining whether data is MNAR is particularly challenging, as it requires knowledge of the values that are not observed [1].

2.3 Techniques: How to Handle MD?

Handling MD involves both preventive strategies (e.g., data validation) and corrective techniques. Three broad approaches are commonly applied: deletion of rows with missing data is removed, which is only advisable when the missingness is Missing Completely at Random (MCAR) and the dataset is sufficiently large. Simple Imputation Methods, such as mean or median imputation, are computationally efficient but can distort variance and reduce model accuracy. Model-based imputation techniques, such as Multiple Imputation by MICE, KNN, Random Forest, and Deep Learning (e.g., GAIN, Autoencoders), provide more robust estimates by leveraging relationships among variables.

3 Proposed Model

This study employs the Knowledge Discovery in Databases (KDD) methodology, as it aligns with the project's objectives. KDD is a well-established framework in machine learning, widely applied in fields such as pattern recognition, statistics, databases, and artificial intelligence. It provides a structured process that culminates in actionable insights, often supported by data visualization [6]. The KDD process in this study is structured into three main steps: *(i) Data Cleaning, (ii) Data Prediction, and (iii) Data Summarization.*

3.1 Data Cleaning

The dataset used in this research originates from a real-world company and comprises approximately 1,402 transactional records spanning the years 2021 to 2024. As sales forecasting depends on historical purchase behavior, this transactional data forms a solid foundation for the prediction model.

Table 1 summarizes the dataset metadata. During the initial data cleaning, duplicate records were removed, which reduced the dataset to 1,370 entries. Records containing missing (*NA*), negative, or outlier values in the *Invoiced_Quant* variable were identified. After excluding all such invalid records, the dataset was reduced to 1,244 clean entries, ensuring data integrity for subsequent analysis.

Table 1. Dataset metadata

Attribute	Type	Description
Year	Int	from September of year X to August of year X + 1
Client	Varchar (5)	client name
Product	Varchar (1)	product name
Unitary Price	Float	invoiced price in monetary units
Budgeted Quantity	Int	budgeted products in units
Invoiced Quantity	Int	invoices for products in units

A pivot table was used to restructure the dataset from a long to a wide format, creating a sub-dataset that aligns better with time-series forecasting needs. The transformation grouped records by *Client*, *Product*, and *Unitary_Price*, while spreading the *Invoiced_Quant* values across four separate columns—one for each year (2021 to 2024). This restructuring resulted in a final sub-dataset of 735 records. The core numerical variables considered for summarization were *Unitary_Price* and *Invoiced_Quant* (see Table 2).

Table 2. Core attributes summary

	Attribute	
	Unitary_Price	*Invoiced_Quant*
Range	1.14 – 8.31	9 – 204,981
Mean	2.68	6,207.3
Median	2.74	1,571
NA	147	–

3.2 Data Prediction: Year 2025

To address MD in *Invoiced_Quant*, an imputation strategy based on the maximum quantity invoiced in the previous two years was employed. For predicting the *Unitary_Price* in 2025, a KNN model was applied. After reversing the transformation to restore the long format, the dataset expanded to 1,623 records.

Given the prevalence of MD, a detailed MD analysis was conducted. MD were classified according to Rubin's framework into: MCAR, MAR, and MNAR [3, 4, 7].

To test for MCAR, Little's test was applied [8]. Based on the classification, a complete dataset was used to simulate various MD scenarios, with missing rates ranging from 1% to 12%. These incomplete datasets were imputed using MICE with a Random Forest model, KNN, Mean, and LReg. This enabled the comparison of imputation techniques under different missingness conditions.

3.3 Data Summarization

After the data cleaning and transformation steps, a comprehensive summarization phase was conducted to explore the structure and key properties of the final dataset. This phase aims to uncover patterns, detect anomalies, and generate insights that can support model design and evaluation.

The cleaned dataset comprises 1,244 transactional records spanning the years 2021 to 2024, covering an average of 578 clients and 4 products.

4 Cleaning Phase

The dataset used in this study comprises 1402 observations, each corresponding to a unique combination of *Year*, *Client*, and *Product*. To ensure analytical integrity, the first step involves removing potential duplicate records based on these three identifiers. This duplication reduced the dataset slightly, ensuring that each row represents a distinct and meaningful observation.

4.1 Patterns

Using the *naniar* package from R, we visualized the pattern of missingness with the *vis_miss()* function. The resulting plot (see Fig. 1) clearly illustrates the proportion and distribution of MD across variables.

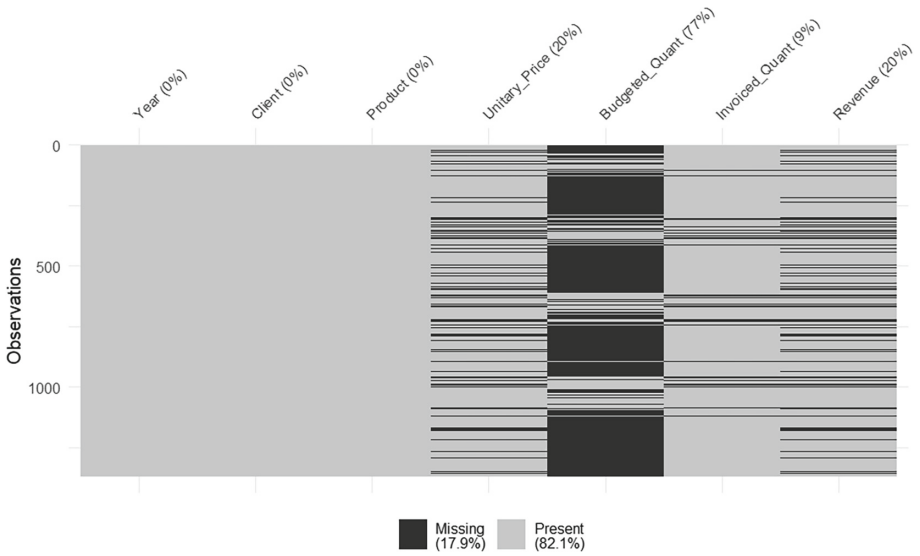


Fig. 1. Dataset MD patterns

This plot (Fig. 1) serves as a crucial diagnostic tool. It shows that while some key variables are complete (e.g., *Client*), others, especially those used in monetary calculations,

are highly incomplete. Zero values in financial and quantity fields often indicate either errors in data entry or special conditions such as unavailability. These were replaced by *NA* where appropriate:

These substitutions ensure that later computations and imputations do not treat invalid zeroes as valid values. Where *Revenue* was not recorded directly, it was recomputed as the product of *Unitary_Price* and *Invoiced_Quant*. However, if either component was missing, the result was also replaced by *NA*:

This strategy avoids misleading values and ensures that derived fields are coherent with their inputs. Since time series modeling often requires complete records for the target variable (*Invoiced_Quant*), observations with missing invoiced quantities were excluded: This reduced the dataset from 1370 to a smaller, yet more reliable, subset for forecasting purposes.

4.2 Mechanism

To discover the missing data mechanism present in our dataset, a Little's test [8] was applied to the numeric columns, except for the *Year*, *Client*, and *Product* attributes. The results indicate that the χ^2 statistic equals 308. This large value indicates a significant deviation from MCAR. The Degrees of Freedom (df) = 14 reflects the complexity of the MD patterns, rejecting the MCAR hypothesis. The Little's test also showed *Missing Patterns* = *7*, suggesting that there are seven distinct ways for MD to occur in this dataset.

Then, the dataset was analyzed to verify the cases where the missing *Unitary_Price* values were not significant. In this study, the Budgeted Quantity attribute was not considered due to the substantial amount of missing data (MD). Therefore, Invoiced_Quantity was regarded as the dependent variable for the missing *Unitary_Price* (MAR).

4.3 Techniques

Imputation models are a common technique for handling missing data. As noted by Liu and Wu [9], the selection of an appropriate forecasting model depends on the choice of evaluation metrics, although no universal standard exists. Among the most widely accepted metrics are the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE). These are computed as shown in Eqs. 1 and 2.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

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

In these formulas, $\boldsymbol{y_i}$ denotes the actual observed value. In contrast, $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ represents the predicted value generated by the model. Lower values of the evaluation metrics indicate higher model accuracy and better forecasting performance. To handle MD, we evaluated four algorithms: MICE, KNN, Mean, and LReg. The original dataset contained 1,244

records, with 12% MD in the *Unitary_Price* attribute. We first removed all missing data, resulting in a clean dataset of 1,100 records. Next, we introduced Missing at Random (MAR) MD into the *Unitary_Price* attribute, varying the missing rate from 1% up to 12%. The maximum missing rate (12%) was chosen to match the original percentage of MD in *Unitary_Price* after data cleaning. This approach allowed us to systematically assess the performance of each imputation method under controlled MD conditions (see Table 3 and Fig. 2).

Table 3. Imputations results using RMSE and MAE evaluation metrics

	RMSE				MAE			
%	MICE	KNN	Mean	L.Reg	MICE	KNN	Mean	L.Reg
2	1.19	**0.59**	0.97	0.98	0.99	**0.34**	0.70	0.71
4	1.18	**0.59**	0.78	0.87	0.90	**0.40**	0.65	0.71
6	1.25	**0.49**	0.84	0.92	0.92	**0.30**	0.64	0.75
8	1.01	**0.56**	0.81	0.74	0.81	**0.32**	0.65	0.60
10	1.27	**0.61**	0.94	0.79	0.93	**0.37**	0.74	0.66
12	1.16	**0.72**	0.94	0.83	0.84	**0.41**	0.68	0.61

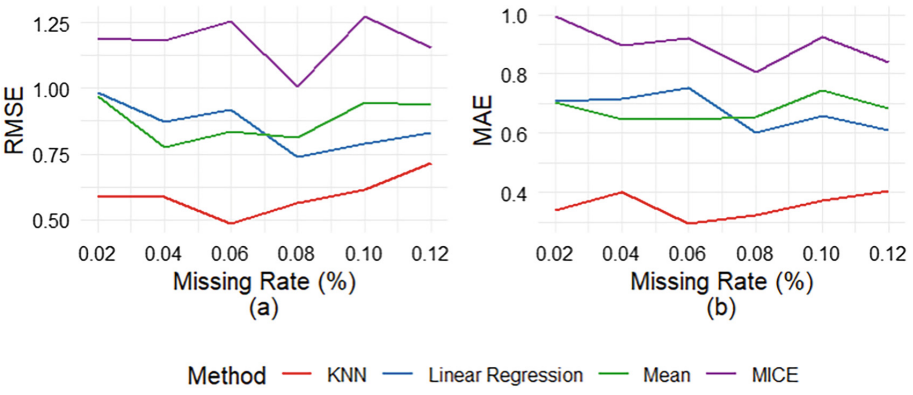


Fig. 2. RMSE and MAE – Imputation methods performance

The RMSE and MAE are consistent, indicating that the KNN algorithm performs best. Then the dataset (n = 1244) with 12% of *Unitary_Price* MD, was completed using the proposed algorithm, resulting in the non-MD dataset.

5 Predicting Phase

5.1 Forecasting Context and the Role of MD

Sales forecasting is the process of predicting future demand based on historical data [10]. Mentzer and Moon [11] define a sales forecast as "a projection into the future of expected demand, given a stated set of environmental conditions."

A variety of classical and modern forecasting methods have been proposed. Statistical models, such as ARIMA, SARIMA, and Exponential Smoothing, remain widely used for univariate time series [12]. Machine Learning techniques, including Random Forests, LSTM networks, and CNNs, have been successfully applied, especially for capturing nonlinear patterns [13–15]. Hybrid approaches combine decomposition techniques with deep learning.

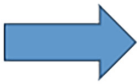
For example, Gao et al. [16] integrated Empirical Mode Decomposition with neural networks to enhance multivariate forecasting performance. Nevertheless, all these approaches are sensitive to MD. Data imputation is thus a critical step, especially in short time series, where even a small number of MD can substantially affect model performance. Traditional methods such as the Single Moving Average and KNN remain effective for straightforward sales environments, as demonstrated in the case of PT.CNC [10].

Despite the wide range of forecasting methods, very short time series often pose challenges where naive models can find feasible solutions.

5.2 Data Transformation and Quantities Prediction

The prediction of quantities for the year 2025 requires a data transformation, as shown in Fig. 3, where the tabular data gives rise to a set of small time series. As exemplified, there are several years during which customers have purchased any product. Training/learning phases from 2021 to 2023, and testing/prediction phase in 2024. After validating the model, it will be applied in 2025.

Year	Client	Product	Price	Quant
2022	1111	B	2.94	1322
2023	1111	B	3.36	1322
2021	1116	A	2.25	1948
2022	1116	A	2.55	28
2024	1116	A	2.22	460
2021	1116	B	2.49	30506
2021	1121	B	2.71	964
2022	1121	B	3.20	1763
2021	1121	C	1.60	2356
2022	1121	C	2.01	7069
2023	1121	C	2.02	6283
2024	1121	C	1.55	7854
2024	1126	B	3.15	823



		Year			
Product	Client	2021	2022	2023	2024
A	1116	1948	28		460
B	1111		1322	1322	
B	1116	30506			
B	1121	964	1763		
B	1126				823
C	1121	2356	7069	6283	7854

Fig. 3. Data transformation to predict invoice quantities

To evaluate the prediction for 2024, two metrics are used. The first is based on the Mean Absolute Error (MAE), and the second is related to the sum of predicted quantities. A final metric combines these two previous ones.

To find the percentage for MAE, the division of the average quantity (Avg) is used, given by MAE% = MAE / Avg * 100. The error of the sum of quantities is provided by the following:

$$SumQuantError\% = \frac{|ActualSumQuant - PredSumQuant|}{ActualSumQuant} * 100 \quad (3)$$

Finally, Mixed Error is given by a convex combination of MAE% and SumQuantError%. In this work, α is set to 0.05, as SumQuantError provides a more accurate representation of the error.

$$MixedError\% = \alpha.MAE\% + (1-\alpha).SumQuantError\% \quad (4)$$

Two approaches were used to predict the quantities for the year 2024: the classical moving average and a heuristic-based approach. Since the clients do not purchase the same product every year, there are MDs in the short-term series.

Naive models [17], such as using the last period's value as the forecast, are helpful in these conditions with short time series with MD. In this work, the naive model is defined as the maximum of the last two years, referred to as Max2.

The results of the mixed metric are presented in Table 4. The metric MAE% indicates significant errors for both techniques. On the other hand, the error related to the quantities is larger for the moving average technique. When combined in the mixed metric, the Max2 approach yields superior results and will be applied to the year 2025.

Table 4. Mixed metric for predicting techniques

Metrics	Prediction	
	Moving Average 2024	Max2 2024
MAE	2,260	2,422
MAE% = MAE/Avg	69.30%	74.27%
Sum Quant	1,775,018	2,512,861
Sum Quant Error %	25.94%	4.84%
Mixed Error %	47.62%	8.31%

In the next section, the summarizing phases, quantities are predicted using the heuristic Max2, and the price is obtained via the KNN method.

6 Summarizing Phase

The total invoiced quantity shows a steady year-over-year increase from 2021 to 2025.

Product C (cyan) consistently contributes the largest share of invoiced quantities across all years. Product B (green) sees a notable growth starting in 2022, becoming the second-largest contributor from that year forward. Product A (orange) and Product D (purple) contribute relatively small quantities, with slight increases over time. By 2025, the total amount invoiced is expected to surpass 2.5 million units, indicating strong growth (see Fig. 4).

By revenue, we mean the total income generated, calculated as the product of the predicted quantity and the corresponding price.

Like invoiced quantities, total revenue increases steadily from 2021 through 2025. Product B (green) overtakes Product C in revenue starting from 2022, suggesting it has a higher unit price or greater value impact. Product C (cyan) remains a major contributor to revenue but lags Product B slightly from 2023 onward. Product A and Product D contribute less to revenue, although their presence remains consistent. By 2025, revenue is expected to exceed 5 million, reflecting the successful combination of sales volume and pricing (see Figs. 5 and 6).

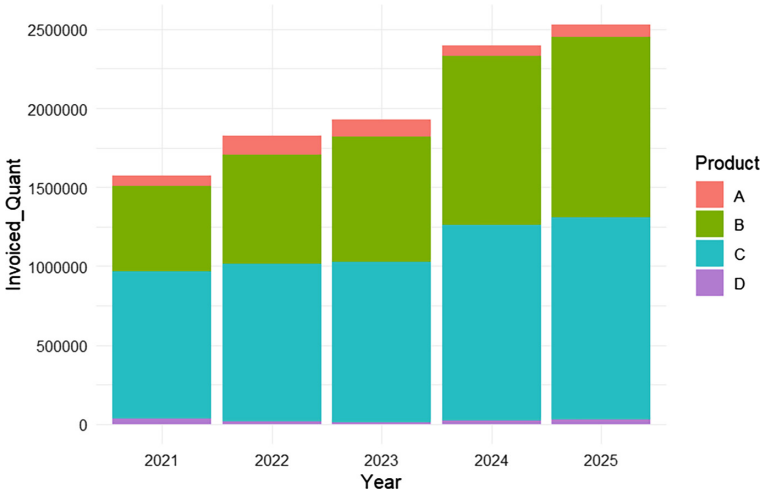


Fig. 4. Invoiced Quantity Prediction

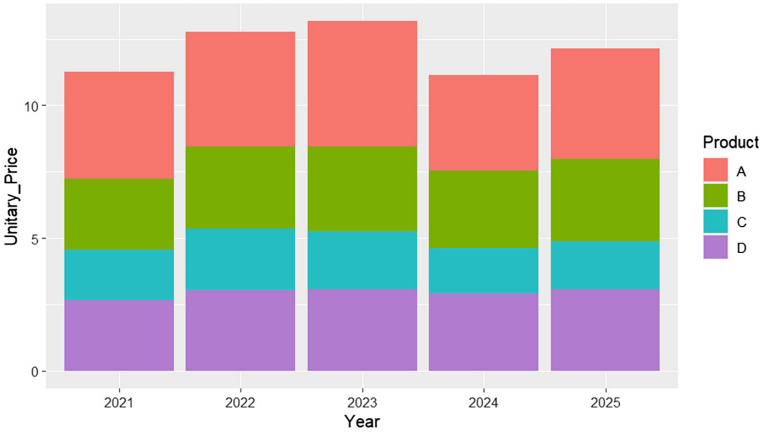


Fig. 5. Unitary Price Prediction

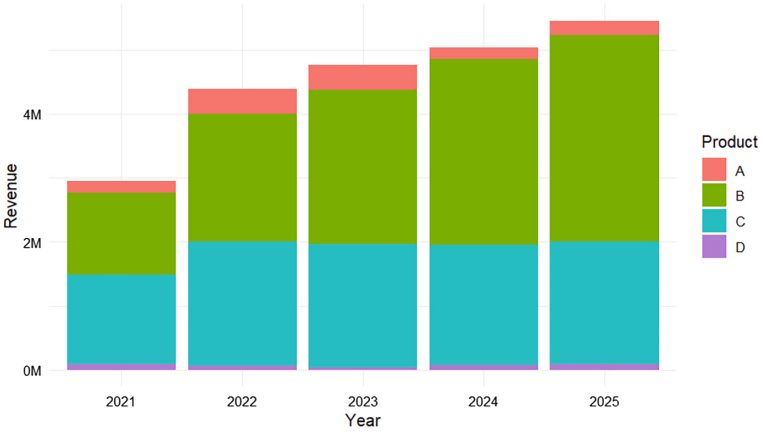


Fig. 6. Revenue Prediction

7 Conclusions

Given a real-world transactional dataset (2021–2024) comprising approximately 1,400 records with missing data (MD), the goal is to predict sales for 2025.

This study reinforces the importance of understanding and addressing MD when dealing with short time series in forecasting contexts. It demonstrates that not all imputation methods are equally effective across different types and rates of missingness.

The predicted price is obtained via KNN, and the quantities forecast use the Max2 naive approach.

In the price prediction, while Mean and Linear Regression perform well in specific scenarios, simpler methods, such as KNN, can still be competitive in environments with low missingness.

Despite the wide range of forecasting methods, very short time series often pose challenges where naive models can find feasible solutions. The Max2 naive approach

proposed for forecasting provides an efficient and interpretable alternative to more complex models in settings with frequent data gaps. Mixed error is given by a convex combination that balances a micro vision of MAE with a macro vision of the error given by the sum of the quantities.

The aggregated data for quantity, price, and revenue confirm a reliable and gradual evolution over the 2021–2025 period.

Future work may include expanding the evaluation to multivariate time series models and exploring more robust evaluation schemes, such as growing or rolling window validation, to better reflect real-world scenarios. Furthermore, evaluating recent deep learning-based imputation techniques—such as GAIN and Autoencoders—across multiple datasets could offer deeper insights into their generalizability and effectiveness in time series contexts.

Acknowledgments. This work was supported by the LASIGE Research Unit, reference UID/00408/2025 – LASIGE.

Competing of Interest. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Newman, D.A.: Missing data: five practical guidelines. Organ. Res. Methods **17**(4), 372–411 (2014). https://doi.org/10.1177/1094428114548590
2. Enders, C.K.: Applied missing data analysis. In: Applied Missing Data Analysis, pp. xv, 377. The Guilford Press, New York (2010)
3. Awan, S.E., Bennamoun, M., Sohel, F., Sanfilippo, F.M., Dwivedi, G.: Imputation of missing data with class imbalance using conditional generative adversarial networks. arXiv:arXiv:2012.00220 (2020)
4. Hallaji, E., Razavi-Far, R., Saif, M.: DLIN: deep ladder imputation network. IEEE Trans. Cybern. **52**(9), 8629–8641 (2022). https://doi.org/10.1109/TCYB.2021.3054878
5. Little, R.J.A.: A test of missing completely at random for multivariate data with missing values. J. Am. Stat. Assoc. **83**(404), 1198–1202 (1988). https://doi.org/10.2307/2290157
6. Singh, K., Booma, P.M., Eaganathan, U.: E-commerce system for sale prediction using machine learning technique. J. Phys.: Conf. Ser. (2020). https://doi.org/10.1088/1742-6596/1712/1/012042
7. Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976). https://doi.org/10.2307/2335739
8. Li, C.: Little's test of missing completely at random. Stand. Genomic Sci. **13**(4), 795–809 (2013). https://doi.org/10.1177/1536867X1301300407
9. Liu, Y., Wu, G.: Research on the prediction of short time series based on EMD-LSTM. J. Comput. Methods Sci. Eng. **23**(5), 2511–2524 (2023). https://doi.org/10.3233/JCM-226860
10. Prasetiamaolana, E., Syafrullah, M.: The use of single moving average and linear regression in spare part sales forecasting at PT. CNC, Int. J. Adv. Technol., Eng., Inf. Syst. **4**(1), 1 (2025). https://doi.org/10.55047/ijateis.v4i1.1587
11. Mentzer, J.T., Moon, M.A.: Sales Forecasting Management: A Demand Management Approach, 2nd ed. SAGE Publications, Inc. (2005). https://doi.org/10.4135/9781452204444
12. Eglite, L., Birzniece, I.: Retail sales forecasting using deep learning: systematic literature review. Complex Syst. Inf. Model. Quart. **30**, 53–62 (2022). https://doi.org/10.7250/csimq.2022-30.03

13. Mallik, R.S., Abhiram, R., Reddy, S.R., Jagadish, R.M.: A Comprehensive survey on sales forecasting models using machine learning algorithms. In: 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), pp. 1–6 (2022).https://doi.org/10.1109/ICERECT56837.2022.10060168
14. Ensafi, Y., Amin, S.H., Zhang, G., Shah, B.: Time-series forecasting of seasonal items sales using machine learning – a comparative analysis. Int. J. Inf. Manag. Data Insights **2**(1), 100058 (2022). https://doi.org/10.1016/j.jjimei.2022.100058
15. Johnson, M., Prakash, P.A., Saihareesh, V., Rajiv, A., Ananthi, S., Anandakumar, H.: Comparative analysis of future sales prediction using artificial intelligence. In: 2024 9th International Conference on Communication and Electronics Systems (ICCES), pp. 1175–1180 (2024). https://doi.org/10.1109/ICCES63552.2024.10859999
16. Gao, W., Li, C., Dong, S., Zhang, R.: Connector based short time series prediction. Sci. Rep. **15**(1), 7082 (2025). https://doi.org/10.1038/s41598-024-83122-y
17. Makridakis, S.G., Wheelwright, S.C., Hyndman, R.J.: Forecasting: Methods and Applications. John Wiley & Sons Inc., New York (1998)