Understanding and Overcoming Missingness in Satellite-Derived Tank Volume Data

J. Baker, SN: 19074278, COMP0047

ABSTRACT

This study investigates methods for imputing missing values in satellite-derived oil tank volume data, focusing on the PADD-3 region of the U.S. over a 50-month period. Given the dataset's high level of incompleteness and complex spatial-temporal structure, we first classified the missingness mechanism, ultimately identifying it as Missing At Random (MAR) through a combination of statistical tests and visual analysis. Building on this, we evaluated a range of imputation techniques—including MICE, K-Nearest Neighbours, Random Forest, Kriging, and Graph Laplacian Interpolation—against baseline models such as mean imputation and linear interpolation. Validation was performed using an MCAR-style mask, with model performance assessed through bias, RMSE, MAE, and R^2 metrics. Contrary to expectations, linear interpolation outperformed all advanced methods, suggesting that high data noise and sparsity may hinder more complex models. This work underscores the importance of aligning imputation strategies with the underlying data structure and validates the use of simpler models when data quality constrains complexity.

I. INTRODUCTION

As discussed in our group presentation, our two main objectives were to address large data gaps and explore the data's unique characteristics. These goals reflect both the primary challenge posed by our data and the unique opportunity it offers. Satellite-derived oil storage data is particularly valuable, as it can help anticipate changes in global oil inventories and influence futures pricing by revealing regional surpluses or deficits. However, its usefulness is significantly diminished when the data suffers from incompleteness, especially in key metrics such as tank fill percentage (`fill_pct`).

Our data is derived from satellite imagery collected over 50 months across the PADD-3 region of the United States, specifically Louisiana. Six satellites captured measurements from 693 unique oil storage tanks. The fill level of each tank was calculated from satellite imagery using shadow analysis. As the volume of oil inside the tank changes, the internal floating roof rises or falls. By measuring the shadow cast, an estimate of the tank's fill level (`fill_pct`) can be derived. Tanks can be sorted in multiple ways; each tank has unique geospatial data and is grouped into one of 27 unique farms (tank farm). These farms are divided into refineries and terminals (farm type) based on operational function. Terminals are required only for storage, while refineries produce petroleum products. Due to the sparseness of daily and weekly data, we grouped our tank data by month and took averages if multiple values were available.

During the group portion of our work, we explored a range of imputation methods, including K-Nearest Neighbours and Exponential Moving Averages, and we reclassified any

undefined tanks to tank farms using a nearest neighbours approach. We modelled relationships between data subsets such as seasonal and monthly changes in mean fill and regional shifts in tank volatility. We also investigated the correlation structure between tank farms and between individual tanks using matrices, Maximal Spanning Trees and Triangulated Maximally Filtered Graphs. While only a subset of this work was included in our group presentation, further analysis was conducted to deepen our understanding of the data. These collective efforts supported our dual goals of addressing missingness and uncovering the relationship between different observed data types.

Building on this foundation, the individual component of the project focusses on evaluating how the unique features of the dataset, such as spatial layout and inter-farm relationships, could influence the performance of different imputation techniques. Specifically, this study explores how to choose an imputation model for missing data, how to classify the missingness of data to inform this choice, and critically how to statistically demonstrate the superiority of one imputation method over another. The objective was not only to assess a diverse range of imputation methods but also to establish a robust and unbiased framework for their comparison and to determine how we can objectively validate imputation performance in the presence of uncertainty and incomplete data.

II. AN INTRODUCTION TO MISSINGNESS

A. Types of Missingness

To begin, we define the different types of missingness as classified by Rubin [1]

- **MCAR (Missing Completely At Random)**: The probability of missingness is independent of both observed and unobserved data. In this case, missing values are randomly distributed across the dataset and have no systematic relationship with any variables.
- **MAR (Missing At Random)**: The probability of missingness depends on observed data, but not on the missing data itself. For example, missingness might depend on time or location, but not on the missing value.
- **MNAR (Missing Not At Random)**: The probability of missingness depends on the missing data itself, even after accounting for observed data. Missingness is driven by the actual (unobserved) value, such as when extreme values in a variable `fill_pct` are more likely to be omitted or when measurement anomalies result in values being erroneously marked as missing.

These types of missingness are formally described as follows:

- **MCAR (Missing Completely At Random)**: $P(M \mid Y_{obs}, Y_{mis}) = P(M)$
- **MAR (Missing At Random)**: $P(M \mid Y_{obs}, Y_{mis}) = P(M \mid Y_{obs})$
- **MNAR (Missing Not At Random)**: $P(M \mid Y_{obs}, Y_{mis}) \neq P(M \mid Y_{obs})$

Where M is the missingness indicator (1 if a value is missing, 0 otherwise), Y_{obs} are the observed values and Y_{mis} the missing values.

Understanding the mechanism behind missing data is critical, as it directly influences how imputation affects the structure and reliability of the dataset. When data is MCAR we may ignore many of the complexities that arise because data is missing, apart from the obvious loss of information. Mean imputation, one of the simplest possible methods to fill values, produces unbiased estimators of means under MCAR, allowing easy and powerful imputations for MCAR data. Likewise, listwise deletion (complete-case analysis) under MCAR yields unbiased estimates of means, variances, and regression coefficients. It also yields standard errors and significance levels that are correct for the reduced subset of the data, although they are often larger relative to what they would be with the full dataset due to a decreased sample size [2]. This makes listwise deletion a powerful and easy solution for MCAR data. However, it can be inefficient and wasteful, especially in sparse datasets where many cases are discarded.

For MAR data, listwise deletion may still produce unbiased regression estimates if the observed values that predict missingness are included in the model [3]. However, for descriptive statistics such as means and variances, listwise deletion may yield biased results under MAR. Therefore, more sophisticated methods such as multiple imputation or random forest imputation are typically recommended, as they can make better use of the available information without introducing systematic bias.

In contrast, MNAR data presents a more complex challenge. Standard techniques such as listwise deletion or imputation cannot be assumed to provide unbiased estimates without explicitly modelling the missing data mechanism. MNAR situations often require selection models, pattern-mixture models, or sensitivity analyses to assess the robustness of conclusions drawn from incomplete data [4].

Understanding the underlying mechanism (MCAR, MAR, or MNAR), is therefore crucial when deciding how to handle missing data. Misclassification of the mechanism can result in biased estimates, spurious findings, and unreliable conclusions.

B. MCAR

We begin by exploring whether our data is Missing Completely At Random (MCAR). To do this, we apply Little's MCAR test [5] which is designed to detect systematic differences between observed and missing values. While Little's test assumes multivariate normality, it is known to be relatively robust to violations of this assumption, particularly with large sample sizes. However, it critically depends on the assumption that the observations (rows) are independent.

This presents a challenge for our dataset, which consists of monthly time series data that are likely not independent. To evaluate this, we plot the lagged autocorrelation function. Figure 1 shows there is significant correlation in the one- and two-month lags, exceeding the 95% confidence interval. These results provide evidence against the hypothesis of independence of the time series, thereby violating a key assumption of Little's MCAR test.

As a result, we are therefore unable to apply this test and must consider alternative methods. We begin with a visual inspection. Figure 2 shows the data over 50 months, grouped by tank farm. We observe that the missing data is not randomly distributed; there are clear patterns of missingness with distinct gaps appearing across months, tanks, and tank types. Although not conclusive, this visual evidence strongly suggests that the data is not MCAR.

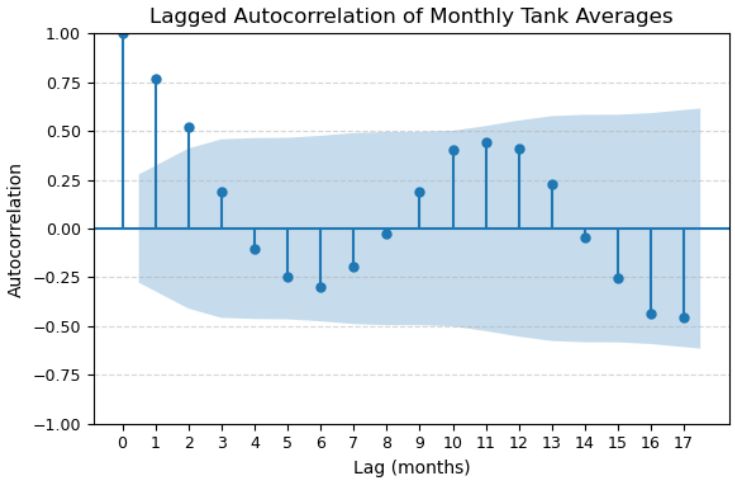

FIG. 1: Lagged autocorrelation plot. The vertical axis shows autocorrelation values, while the horizontal axis indicates lag in months. The shaded area represents the 95% confidence interval.

FIG. 2: Heatmap showing the pattern of missing data across 50 months and unique tanks, grouped by tank farm. Refineries occupy the left 70% of the horizontal axis, while terminals are located in the right 30%. Missing data is shown in yellow and observed data in blue.

C. MAR

Having explored the possibility of MCAR, we now investigate whether the data is Missing At Random (MAR). If the data is MAR, then its missingness is related to other observed data. In contrast, if the data is MNAR, the missingness is related to the unobserved values of `fill_pct` itself.

To begin, we examine the relationship between missingness and static tank characteristics such as farm type, tank farm, and geographical location, as well as temporal patterns by month and season.

We first calculate the percentage of missing values for each tank and analyse it by farm

type. Refineries comprise 70.27% of the dataset. From our group data analysis, we observed that 45.04% of refinery data is missing, whereas 56.44% of terminal data is missing on average. Figure 3, shows the disparities in missingness between farm types in greater detail. To formally test the hypothesis that missingness is related to farm type. We consider

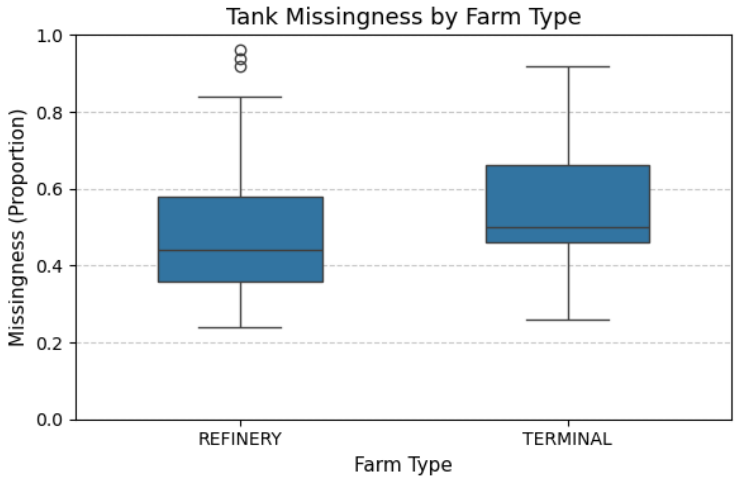

FIG. 3: Box plot showing the distribution of missingness (proportion of missing values) by farm type. The box represents the interquartile range (IQR), whiskers extend to 1.5×IQR, and outliers are shown as individual points.

two potential statistical methods; the Student's t-test [6], which requires assumptions of normality and equal variances, and the non-parametric Mann-Whitney U test [7]. We test for normality using the Shapiro-Wilko test [8], the null hypothesis being that the data comes from a normal distribution. Refinery data returns a p-value of 1.03e-14 and terminal data returns a p-value of 2.44e-10, indicating strong evidence to reject the null hypothesis in both cases. We also use Levene's test [9] to compare equality of variances. Levene's test assumes the null hypothesis that all groups have equal population variances. It is generally robust against non-normal distribution, meaning the results remain valid even when the data are not normally distributed. Levene's test returns a p-value of 0.827, indicating that there is no evidence to reject the null hypothesis that these variances are equal. Given these two test results, we must use the Mann-Whitney U test, a non-parametric test that will compare the rank sums of each distribution. The null hypothesis is that the distributions of both groups are identical. The Mann-Whitney U test returns a p-value of 2.488e-19 indicating very strong evidence to reject the null hypothesis. This shows that missingness differs significantly between the two tank farm types.

Next, we examine the relationship between missingness and specific tank farms. The dataset includes 27 different tank farms, which are classified as refineries or terminals. To evaluate whether missingness is influenced solely by farm type or is specific to individual tank farms, we first generate a box plot. Figure 4 indicates that missingness is not uniformly distributed between tank farms. Tank farms such as 8816 show a median missingness of 0.3% while others such as 6655 show a median missingness of 0.7%. Combine with the wide interquartile ranges and outliers, this reinforces the hypothesis that missingness is related to tank farm identity, suggesting MAR. We implement a Kruskal-Wallis H test [10], a non-parametric method to compare whether any group has a systematically different central tendency, meaning its values tend to be higher or lower than the rest of the group. This test

does not assume normality or equal variances, as it compares the rank sums across groups, making it suitable for immediate use. The null hypothesis is that all groups come from the same distribution. Our test returns a p-value of 3.422e-95 indicating that we can reject the null hypothesis. This suggests that at least one group differs in distribution, but not which one. That information can be drawn from our plot.

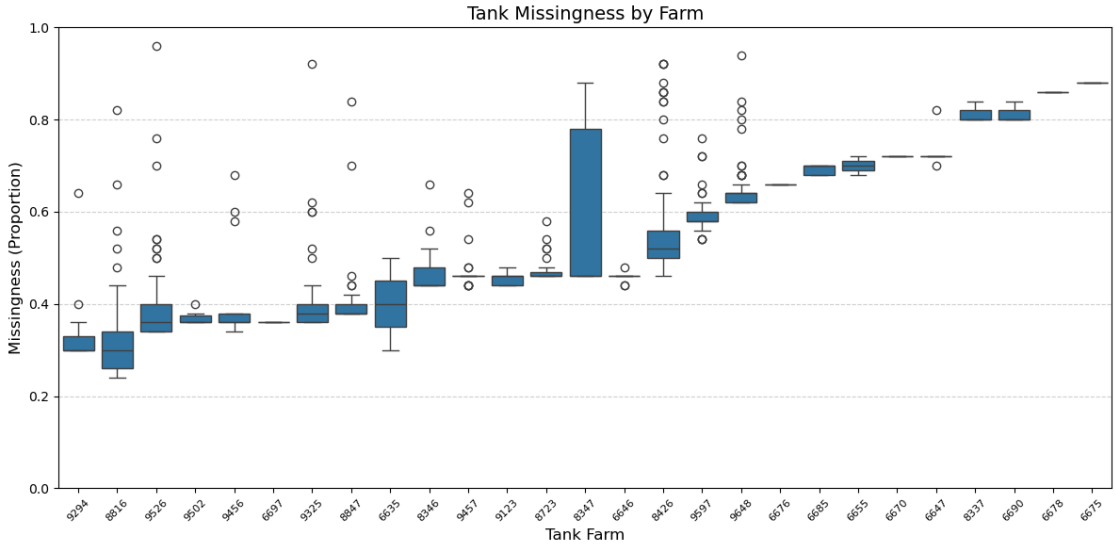

FIG. 4: Box plot showing the distribution of missingness (proportion of missing values) by tank farm. Farms are ordered by median missingness. The box represents the interquartile range (IQR), whiskers extend to 1.5×IQR, and outliers are shown as individual points.

We also examine the relationship between geographical location and missingness. Correlation analysis reveals that latitude has a correlation of $r = 0.31$ and longitude a correlation of $r = -0.6$ with missingness (Figure 5). As tank farms are often geographically clustered, these correlations likely mirror the findings of the tank farm analysis and do not provide independent evidence of MAR. Notably, the correlation matrix in Figure 5 does not show a relationship between `max_vol` (maximum tank volume) and missingness.

Finally, we assess whether missingness varies over time. Figure 2 shows clear patterns in missingness by row (month), indicating possible time dependence in missingness. To explore this further, we calculate the percentage of missing `fill_pct` values per month and annotate the data points by season to explore possible seasonal effects. Although there is no definitive indication of seasonal dependency, Figure 6 shows a clear concentration of missing data during the first 15–18 months of the time series, similar to our missing data heatmap. To evaluate the relationship between missingness and season, we implement a Kruskal-Wallis H test on the grouped monthly data. With a p-value of 0.117, there is not enough evidence to reject the null hypothesis that missingness does not vary by season. However, the early concentration of missing values points to broader temporal trends, which further supports the MAR assumption.

D. MNAR

Before drawing conclusions on the nature of missingness, we must consider the possibility that the data is Missing Not At Random (MNAR), that is, missingness depends directly on

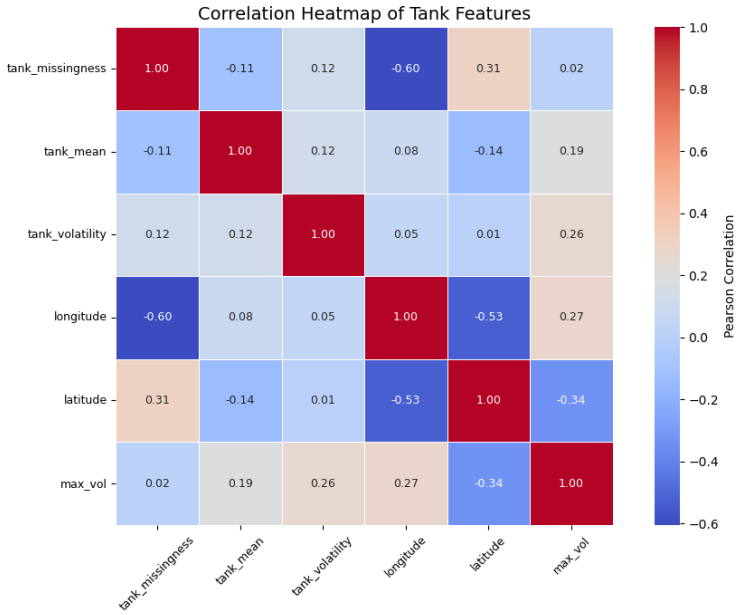

FIG. 5: Pearson correlation matrix showing relationships between missingness and static tank features.

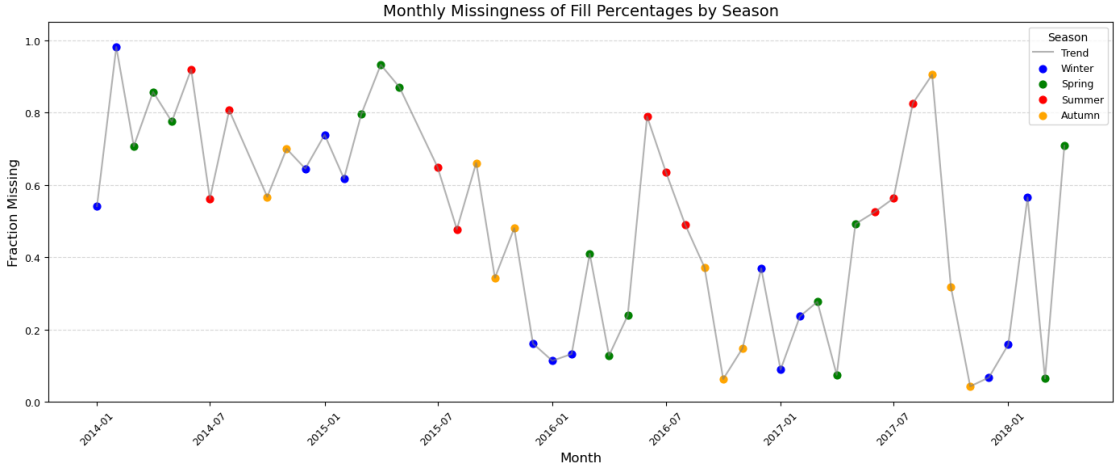

FIG. 6: Line and scatter plot showing the monthly missingness of `fill_pct` values, coloured by season. Higher values indicate a greater fraction of missing data.

the unobserved values of the target variable `fill_pct`. MNAR is the most challenging form of missingness to identify, as it cannot be definitively confirmed using observed data alone.

To explore MNAR, we begin by comparing each tank's average `fill_pct` to its missingness rate. This allows us to assess whether average fill levels have a relation to missing data. In Figure 7, individual data points show no strong trend, the rolling average suggests a slight increase in missingness among tanks with lower average fill levels. This provides weak evidence in support of MNAR. The correlation between tank mean and missingness is $r = -0.11$, as reported in Figure 5.

Next, we investigate whether volatility in tank measurements is related to missingness. Specifically, we define volatility as the standard deviation of `fill_pct` for each tank. In Figure 8 the direct relationship remains inconclusive, the rolling average displays a subtle 'smile'

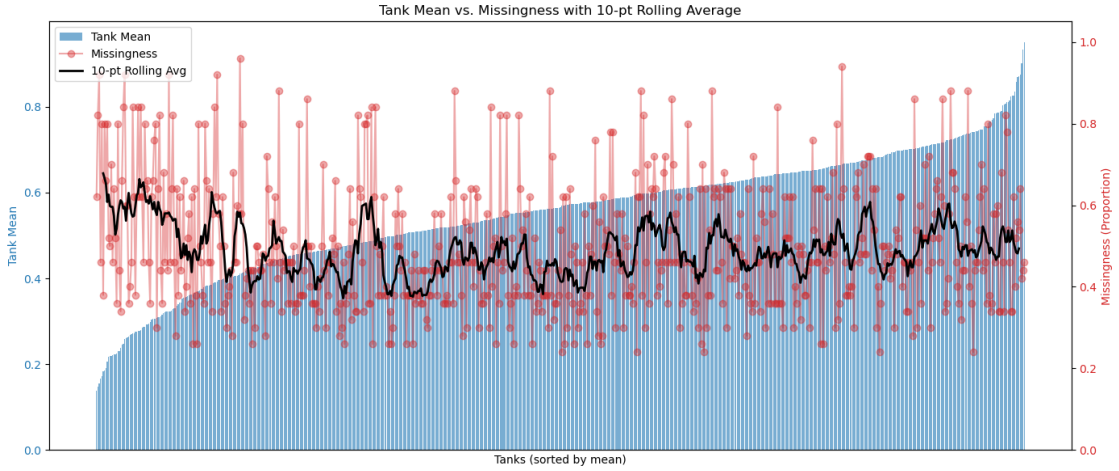

FIG. 7: Dual-axis plot showing tank mean (blue) and corresponding missingness (red), sorted by tank mean. With a rolling 10-point average of missingness (black).

pattern, with missingness increasing at both high and low ends of the volatility spectrum. This pattern offers tentative support for MNAR. However, the correlation coefficient is only $r = 0.12$ (Figure 5), further emphasising that this association, while suggestive, is not definitive. In conclusion, although weak correlations and visual trends hint at a possible MNAR component, the evidence is insufficient to confirm this. Compared to the stronger statistical and visual signals supporting MAR, the case for MNAR remains largely speculative.

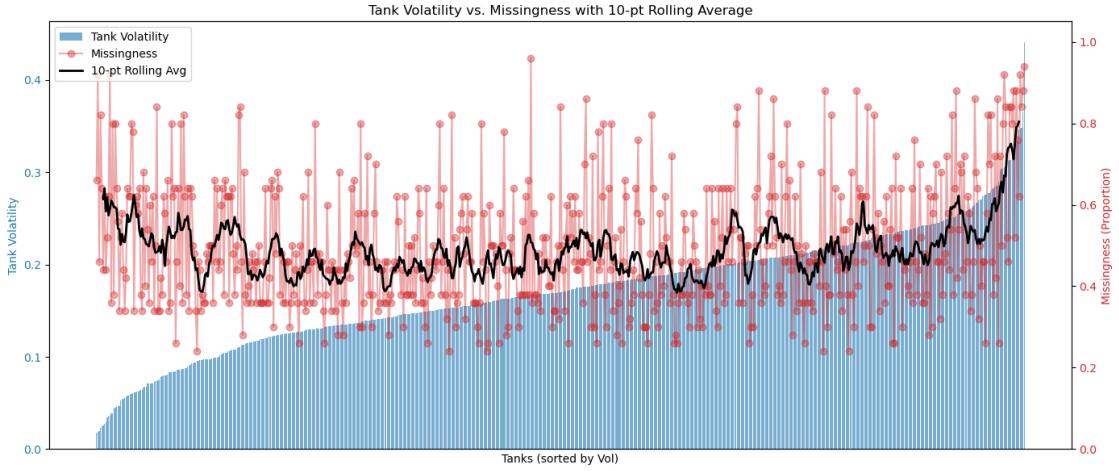

FIG. 8: Dual-axis plot showing tank volatility (blue) and corresponding missingness (red), sorted by tank volatility. With a rolling 10-point average of missingness (black).

E. Conclusion on Missingness Mechanism

To conclude this section, we explored the possibility of MCAR, MAR, and MNAR in our dataset. When examining MCAR, we found little evidence to support the hypothesis that the missingness was unrelated to the observed data. As such, we rejected the MCAR hypothesis, concluding that the missingness is more likely MAR or MNAR.

In contrast, our investigation into MAR revealed strong evidence that missingness is associated with observed data. Significant differences in missingness were found across farm types and individual tank farms. These patterns were statistically confirmed by Mann-Whitney U and Kruskal-Wallis H tests. Additional support came from geographical correlations and temporal clustering, all consistent with the MAR assumption.

Finally, in our examination of MNAR, we discovered limited evidence suggesting a relationship between missingness and unobserved values using the mean and volatility of `fill_pct`. As such, the case for MNAR remains speculative and lacks the robust support observed for MAR.

Based on this evidence, we assume that the data is MAR, with missingness related to observed features in the dataset. This assumption allows us to employ imputation models designed for data that is MAR, such as Multiple Imputation by Chained Equations (MICE), K-Nearest Neighbours (KNN), and Random Forest Imputation. These methods leverage relationships in observed data to yield approximately unbiased unbiased and efficient estimates under MAR as show by Rubin [11], Beretta and Santaniello [12] and Shah et al. [13] respectively. Conversely, we avoid techniques like listwise deletion and pairwise deletion, which rely on the MCAR assumption and are likely to introduce bias to our dataset.

III. DATA AND MODEL VALIDATION

A. Data

Before we begin our imputation analysis, we need to carefully prepare our dataset. To properly evaluate the quality of the imputation methods, we must compare the imputed values to their true known counterparts. However, since there is no complete version of our dataset exists for such a comparison, we must create a simulation by masking known values. This allows us to compare imputed values with their true known counterparts and assess imputation quality.

Our first step was to attempt to construct a fully complete subset of data by removing all rows or columns that contain missing values. A quick check using

```
(data_monthly_fill.isna().sum() == 0).sum()
(data_monthly_fill.isna().sum(axis = 1) == 0).sum()
```

revealed that every row and column in the dataset contains at least one missing value. Thus, a complete subset cannot be created through this method.

Instead, we adopt a new approach. We take our monthly time-series data, organised by tank, and iteratively remove the row or column with the most missing data until the resulting dataset contains no missingness. This method yields a complete dataset, after removing 84.77% of the data, leaving only 14 months and 377 tanks. While technically complete, this stripped-down dataset lacks essential characteristics, such as temporal continuity and tank farm diversity. As a result, any model validated on this reduced dataset would likely not generalise to the broader problem space, undermining the purpose of the exercise.

An alternative strategy focusses on preserving meaningful structure in the data. The original dataset has a missingness rate of 49.2%, classifying it as having high missingness bordering on very high by standard definitions. Firstly, we eliminate the first 15 months of the dataset entirely. As discussed in our presentation, this early portion contains a large

amount of missing data, harming time series modelling. These months also disproportionately contribute to tank removals, so their exclusion enhances the quality of the dataset. After this cleaning process, we settle on a missingness threshold of 30%. We again remove rows and columns with the highest missingness, but this time with the additional goal of preserving data variety. Specifically, we want a balanced representation of refineries and terminals, multiple tank farms, and a continuity in our monthly data. This results in the removal of 15.77% of the data (not including the early months), yielding a final dataset with 29.97% overall missingness and 31 time points, with only two time series breaks (2016-06 and 2017-08 - 2017-09). We also keep 659 tanks out of 693, retaining structure in tank farm and tank type. This configuration strikes a suitable balance between the quality of the features and the retention of data. This strategy allows us to construct a reduced dataset that is both representative and usable. We also apply a 5% masking of the validation data, simulating a Missing Completely At Random (MCAR) scenario. These values will serve as our test set for evaluating imputation performance. Since this adds to the total missingness, it is crucial to strike a balance between reducing overall missingness and preserving enough data for validation. This results in approximately 33.47% of the data being imputed, approaching the upper bounds for reliable performance by most imputation methods. The rate of data loss rises quickly but linearly as the missingness threshold increases, as illustrated in Figure 9.

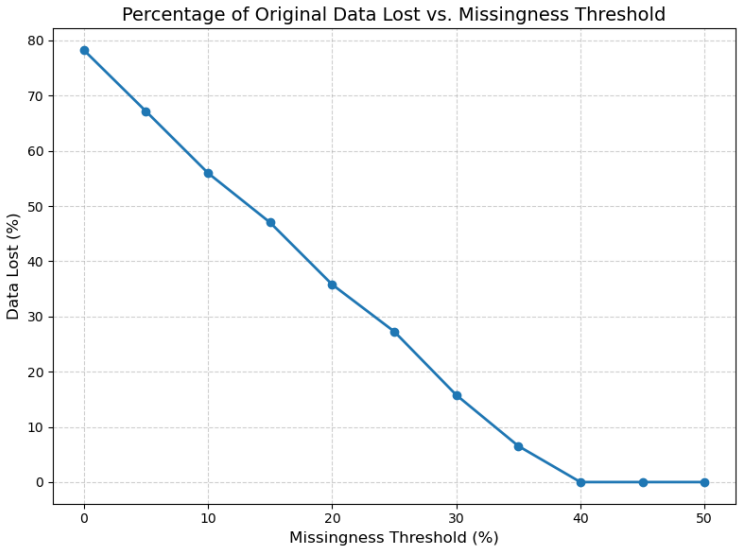

FIG. 9: Percentage of original data lost vs. missingness threshold after data cleaning.

B. Model Accuracy

To assess the performance of our imputation models, we use a set of complementary metrics that capture bias, precision, and explanatory power. These metrics help identify not only how accurate a model is, but also the nature of its errors.

Absolute bias (bias) and percentage bias measure the average systematic error between the predictions of each model and the actual observed values. The bias is computed as:

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i), \quad \text{Percentage Bias} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{y}_i - y_i}{y_i}\right) \times 100 \tag{1}$$

A positive bias indicates systematic overprediction, while a negative bias indicates underprediction. Ideally, bias should be close to zero, suggesting that there is no consistent deviation in either direction. Percentage bias expresses this error relative to the actual values y_i, providing a scale-invariant comparison between models. This allows for easier interpretation when comparing models with different output ranges. The absolute bias gives a direct measure of distance in the original units and both will be used here.

Root Mean Squared Error (RMSE) measures the average magnitude of the error, emphasising larger deviations:

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

RMSE is particularly useful when large prediction errors are undesirable or costly, as in our case when forecasting `fill_pct`. However, its sensitivity to outliers makes it less robust in the presence of noise. Therefore, we use it alongside Mean Absolute Error (MAE) for a more balanced evaluation.

Mean Absolute Error (MAE) measures the average absolute difference between predicted and observed values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{3}$$

Unlike RMSE, MAE gives equal weight to all errors, making it more robust to outliers. MAE is valuable in real-world data applications where irregularities are common. Both RMSE and MAE have the advantage of being interpretable in the same units as the target variable. For example, a MAE or RMSE of 0.1 corresponds to an average prediction error of 0.1.

The R^2 score, or coefficient of determination, quantifies the proportion of variance in the observed data that is explained by the model. It is calculated as one minus the residual sum of squares divided by the total tum of squares:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

An R^2 of 1 indicates perfect predictive performance, while a value of 0 means the model performs no better than simply predicting the mean of the test sample. Negative values, though rare, imply that the model performs worse than mean imputation. Due to its sensitivity to outliers and limited interpretability in high-missingness contexts, R^2 should be interpreted in conjunction with RMSE and MAE for a comprehensive evaluation of the model.

IV. IMPUTATION METHODOLOGY

Baseline Models

A. Mean Imputation

Mean imputation is one of the simplest methods for handling missing data. For each unique tank ID with missing `fill_pct` values, we replace the missing entries with the mean of the observed values for that specific tank.

Mean imputation has several critical drawbacks. As this method inserts a constant value in place of missing entries, it reduces the natural variance of the data and can distort covariance relationships, especially in datasets that are MAR. As discussed previously, this makes mean imputation biased under MAR assumptions. Additionally, it may severely impact downstream modelling tasks by masking important variation or introducing artificial patterns. Nonetheless, we include mean imputation as a baseline method for comparative purposes. Its performance serves as a useful reference point for evaluating more advanced techniques.

B. Linear Interpolation

Linear interpolation imputes missing values by assuming a linear progression between two known data points. In our case, we apply this method to `fill_pct` values across time for each tank individually. Where values are missing at the start or end of the time series, one uses forward-fill or backward-fill, respectively, to propagate the nearest known value and complete the sequence.

Linear interpolation is simple and intuitive, however it is limited in several ways. It cannot capture nonlinear dynamics, instead assuming a constant rate of change between observations. This results in potentially unrealistic transitions, especially in cases of rapid or irregular fluctuations. Furthermore, its reliance on boundary filling introduces additional bias, by extending static values without modelling. Crucially, the method does not incorporate information from other variables, making it a purely univariate technique that overlooks potentially informative correlations in the dataset.

Fit Models

C. Multiple Imputation with Chained Equations (MICE)

Multiple Imputation by Chained Equations (MICE) is a widely used framework for imputing missing data. It operates under the assumption that the data is Missing At Random (MAR), and consists of two core components, the concept of multiple imputation and the mechanism of chained equations. Multiple imputation involves generating several plausible versions of the complete dataset instead of just one. Each version is imputed independently and statistical analysis is performed on each completed dataset. The final results are obtained by pooling these estimates, which helps capture the uncertainty introduced by the missing data. Pooling is typically done using Rubin's rules [11]. If $\hat{Q}_m$ is the estimate from

the m-th imputed dataset, and U_m is the corresponding variance, the overall estimate $\bar{Q}$ and total variance T are computed as:

$$\bar{Q} = \frac{1}{M}\sum_{m=1}^{M}\hat{Q}_m, \quad B = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{Q}_m - \bar{Q})^2, \quad \bar{U} = \frac{1}{M}\sum_{m=1}^{M}U_m, \tag{5}$$

$$T = \bar{U} + \left(1 + \frac{1}{M}\right)B, \tag{6}$$

where B is the variance between the imputations and $\bar{U}$ is the average variance within the imputations.

MICE uses a chained equation approach, where each variable with missing values is imputed in turn by modelling it conditionally on the other variables. The process begins by initialising the missing values using a simple method, such as mean imputation. The algorithm then iteratively cycles through each variable that contains missing values. For each such variable, a regression model is fitted using the other observed variables as predictors. The missing entries are then replaced with samples drawn from the predictive distribution of that model. This sequence is repeated across all incomplete variables for multiple iterations until the imputed values stabilise and the process converges. In our implementation of MICE, we use linear regression to impute `fill_pct`, which is a continuous variable. The standard linear regression model assumes a linear relationship between the predictors and the response variable:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \tag{7}$$

Here, $\mathbf{X}$ is the matrix of observed predictor variables, $\mathbf{y}$ is the response vector, and $\boldsymbol{\varepsilon}$ represents the error term. The model coefficients $\boldsymbol{\beta}$ are estimated by minimising the residual sum of squares (RSS).

D. K-Nearest Neighbours (KNN) Imputation

K-Nearest Neighbours (KNN) imputation estimates missing values by identifying the k most similar observations (rows) in the dataset based on the other observed features. Typically, similarity is measured using Euclidean distance, although other distance metrics can also be used depending on the nature of the data. Once the k nearest neighbours of a given instance are identified, the missing value is estimated by averaging the corresponding feature values from these neighbours. Formally, the imputed value $\hat{x}_{ij}$ for the i-th row and j-th feature is computed as:

$$\hat{x}_{ij} = \frac{1}{k}\sum_{x_n \in N_k(x_i)} x_{nj} \tag{8}$$

Where $\hat{x}_{ij}$ is the missing value to be imputed, $N_k(x_i)$ denotes the set of k nearest neighbours to row x_i (based on available features), and x_{nj} is the value of feature j in the n-th neighbouring row.

This method is particularly well-suited to our dataset, as our observed features, such as tank type, tank farm, and geographical location, exhibit strong local patterns. These patterns extend to the time series structure as well, and KNN is well suited to leverage

such localised similarity. The method is non-parametric and does not assume a particular distribution for the data, making it a flexible choice for our use case. However, KNN is computationally expensive, especially with large datasets, since it requires computing distances for every imputation. Additionally, it may perform poorly in the presence of noise or when missingness is highly clustered, as it can struggle to find meaningful neighbours in such cases.

E. Random Forest Imputation

Random Forest imputation estimates missing values by training a Random Forest model to predict each incomplete variable based on the other observed variables. As a model-based imputation method, it is capable of capturing nonlinear relationships and complex interactions between both numerical and categorical features, making it particularly well-suited for datasets like ours with heterogeneous structure and dependencies. One popular implementation of Random Forest imputation is MissForest, introduced by Stekhoven and Bühlmann [14]. This approach builds a Random Forest from observed data and iteratively predicts missing values. While MissForest is highly accurate in many settings, we choose not to use it due to its limitations when handling datasets with large clusters of missing values, an issue present in our case.

Instead, we implement a Random Forest Regressor within the `Scikit-learn` IterativeImputer framework [15]. This method uses the same chained equations approach as MICE, allowing it to refine imputation estimates over multiple iterations. Initially, missing values are filled with simple estimates, and then each feature with missing values is modelled as a function of the others using Random Forest regression. This process is repeated iteratively across all incomplete variables until the imputations converge. This implementation offers two key advantages. First, it benefits from the robustness and flexibility of Random Forests, particularly their ability to handle interactions and nonlinearity. Second, it avoids the convergence and performance issues associated with MissForest in high-missingness environments, making it a much more effective strategy for our dataset.

F. Kriging Imputation

Kriging [16] is a spatial interpolation technique that estimates the value at an unmeasured location by taking a weighted average of nearby known values, where the weights are derived from the spatial configuration of the data. Unlike simpler distance-based interpolation methods, kriging incorporates not only the distance between points but also the spatial correlation of the data. The kriging estimator for a value at an unknown location s_0 is given by:

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) \tag{9}$$

Where $Z(s_i)$ are the known values at spatial locations s_i and λ_i is the weight assigned to each observation. These weights are selected to minimise the variance of the prediction error while ensuring that the estimator is unbiased. This is achieved by solving a system of equations derived from the spatial autocorrelation of the data.

The spatial dependence is modelled using a variogram, which quantifies how the similarity between points decreases as the distance between them increases. The (semi)variogram is defined as:

$$\gamma(h) = \frac{1}{2}\text{Var}(Z(s) - Z(s+h)) \tag{10}$$

Where h represents the lag distance between spatial locations. In practice, we fit a theoretical variogram model to the empirical variogram derived from the data, specifically, we use a Gaussian model. This guides the calculation of weights λ_i and enables kriging to generate both point estimates and confidence intervals.

Unlike simpler interpolation methods such as Inverse Distance Weighting (IDW), kriging adapts to the underlying spatial structure of the dataset. By accounting for both distance and correlation, kriging provides more accurate and reliable predictions in cases where spatial dependence plays a significant role. This makes it particularly applicable to our setting, where infrastructure and environmental factors introduce spatial patterns in tank measurements.

G. Graph Laplacian Interpolation

Graph Laplacian Interpolation is a method used to impute missing values on the nodes of a graph. The central assumption is that the target variable (`fill_pct`) varies smoothly over the graph structure. In the context of our dataset, nodes represent tanks and edges represent a notion of similarity between tanks. We derive our edges from a sparse weighted correlation matrix constructed using a Triangulated Maximally Filtered Graph (TMFG) [17]. Formally, let the graph $G = (V, E)$, consist of a node set V and an edge set E. The adjacency matrix ($\mathbf{A}$) represents the weights A_{ij} is the correlation between nodes i and j if they are connected, and the degree matrix ($\mathbf{D}$) is a diagonal matrix with entries $D_{ii} = \sum_j A_{ij}$.

Using these, the (unnormalised) graph Laplacian is defined as:

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \tag{11}$$

Let $f : V \to \mathbb{R}$ map each node to its corresponding `fill_pct`. The goal is to find the function f that minimises the graph energy, an objective function that captures how smoothly values change over connected nodes:

$$\mathcal{E}(f) = \frac{1}{2}\sum_{i,j} A_{ij}(f(i) - f(j))^2 = f^T\mathbf{L}f \tag{12}$$

This energy is minimised when the values of f vary smoothly across the graph, meaning similar nodes (as determined by A_{ij}) have similar values. Since some values of f are already known, we partition the function and Laplacian matrix accordingly:

$$f = \begin{bmatrix} f_L \\ f_U \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} \mathbf{L}_{LL} & \mathbf{L}_{LU} \\ \mathbf{L}_{UL} & \mathbf{L}_{UU} \end{bmatrix} \tag{13}$$

Here, f_L represent values at labelled (observed) nodes and f_U represent values at unlabelled (missing) nodes. To minimise $\mathcal{E}(f)$ subject to the known values f_L, we solve the following linear system:

$$\mathbf{L}_{UU} f_U = -\mathbf{L}_{UL} f_L \tag{14}$$

This formulation corresponds to a harmonic extension and is derived from the discretised Dirichlet problem on graphs. The result is a sparse, closed-form linear system that produces the smoothest possible extension of f_L to the unknown nodes. To improve convergence, we initialise the missing values using mean imputation and iteratively update the solution using the graph Laplacian until it stabilises. This chained equation approach enhances robustness to missingness and allows progressive refinement of estimates. Graph Laplacian Interpolation is a powerful non-parametric method. It is highly adaptable, allowing the use of arbitrary relationships in defining graph weights, and produces deterministic solutions via linear algebra. However, its effectiveness depends on the quality and connectivity of the underlying graph structure.

V. RESULTS

To evaluate model performance, we assessed all imputation methods on a dataset with an overall missingness of 33.47%, which includes an artificial 5% MCAR mask. Each model was tested across five different random seeds as noted in each column. We did this to ensure a fair performance test for each model and to guarantee a balanced range of missing values. The results were averaged and accompanied by standard deviations to account for variability across the trials. Table I in Appendix A summarises performance across all models.

With the exception of linear interpolation, most methods produced comparable results, with average RMSE and MAE values falling within the ranges of approximately ∼(0.28-0.29) and ∼(0.23-0.24) respectively. However, these models also returned highly negative $\mathbf{R^2}$ scores ∼(-1). These models produce similar results to mean imputation with low errors, but high R^2 scores. This is very indicative of mean regression behaviour, likely due to excessive noise or missingness in the dataset. This outcome occurs when fitted models revert to predicting a mean value to minimise their loss function, instead of learning the datasets patterns. In contrast, linear interpolation has the lowest mean RSME and MAE at 0.186 and 0.133 respectively, and a near zero but positive R^2 Score of 0.00162 which vastly outperforms the other models' negative scores. It is important to note linear linear interpolation is a baseline model, included to test if our fitted models are capable of learning the trends in the data.

Graph Laplacian interpolation is our second best model. The TPFG built off correlations between tanks performs well here, with average RMSE and MAE values of 0.279 and 0.228, respectively, with low standard deviation indicating that these consistent performance across each seed. The model also produced one of the lowest percentage biases, although its high standard deviation indicates these results were inconsistent across seeds. The R^2 Score is the lowest across all models, indicating that although this method performs well on an average prediction, it struggles to model the true variability of the data. Graph Laplacian interpolation should not exhibit mean regression tendencies, as it preserves local structure through a closed-form solution to a smoothness-based linear system. However, introducing a chained equation framework may have introduced mean regression effects over iterative cycles.

KNN, Kriging and Random Forrest imputation perform similarly, with average errors of ∼(0.28-0.29) for RMSE and ∼0.23 for MAE. They also have low R^2 scores (<-1) and low biases on average, but exhibited high standard deviations, suggesting inconsistency in

underestimation or overestimation across different seeds. Their overall performance aligned with that of the mean imputation baseline and fell below the linear interpolation baseline.

MICE recorded the highest RMSE (0.300) and MAE (0.245) among all methods, making it our worst performing model. These values are contrary to its comparative low R^2 Score (-0.982) which, though still negative, outperformed that of other fitted models and our mean baseline, suggesting that MICE predicted values close to the mean of the test sample. MICE has low bias with the lowest bias standard deviation of all models. This result aligns with MICE's methodological design, which combines multiple regressions to approximate central tendencies.

The results of our fit models and the outstanding performance of linear interpolation could be due to multiple different factors. The most likely explanation is that the dataset contains too much noise and missingness for complex models to effectively learn meaningful patterns. In such scenarios, it is common for models to regress toward the mean in order to minimise loss, which would explain the performance similarity between our fitted models and simple mean imputation. An alternative, albeit less likely, explanation is that the data does not support complex, nonlinear patterns and a predominantly linear structure exists. This would justify the strong performance of linear interpolation. However, we do not consider this to be the most plausible interpretation. As we demonstrated in our earlier group presentation, clear relationships exist between `fill_pct` values and multiple features, including tank farms, tank types, and geographic location, as illustrated by the correlation matrix we showed. We also showed a relation between `fill_pct` and time, when we demonstrated temporal and seasonal trends in our data, by graphing quarterly trends and monthly averages.

These observations suggest that the underperformance of more sophisticated models is more likely a result of excessive missingness or high noise, rather than a true absence of informative structure. Further study would be required to confirm either theory definitively. To do so, future work could explore data augmentation using bootstrapped seasonal trends or Gaussian process interpolation, or de-noising strategies such as moving average smoothing or autoencoder-based filtering, to better assess the signal-to-noise ratio and validate these conclusions.

VI. DISCUSSION

This assumption guided our choice of imputation methods. We selected models well-suited to MAR data including MICE, KNN, and Random Forest Imputation due to their ability to leverage inter-variable relationships. These methods are designed to mitigate bias under MAR conditions. We also included Kriging and Graph Laplacian Interpolation to account for spatial and relational structure specific to our dataset. These selections reflect a strategy tailored to both theoretical alignment with MAR and the practical characteristics of the data.

The imputation results further validated the MAR assumption. Although our fitted models often exhibited high errors, which was likely due to the data's unique challenges such as high noise and large clusters of missingness. We found that our models returned consistently low percentage bias. This outcome is significant, as the goal of classifying the missingness mechanism was to inform model selection in a way that minimises bias. Among the models tested, MICE achieved the lowest overall bias when accounting for standard deviation, aligning with expectations for a model explicitly designed for MAR data. As

such, our decision to classify the missingness mechanism prior to imputation proved to be a critical and ultimately successful strategy.

Our study did have some limitations. We validated our methods using an MCAR-style mask, rather than testing against more structured missingness. Given our dataset is MAR, future work could employ a stratified MAR masking strategy, removing 5% at random from each subgroup of data (farm type, tank farm). This would ensure a more balanced set of data for imputation testing. While this approach would have strengthened our validation, we opted not to pursue it due to the dataset's already high level of natural missingness. In particular, a stratified mask would have disproportionately affected features with already high levels of missingness. By imposing uniform masking across subgroups, there was a risk of further obfuscating signals in the sparsest segments, potentially weakening the model's ability to learn meaningful patterns from these areas. Likewise, testing under a synthetic MNAR mask, where data with specific values would be more likely to be removed could further clarify model behaviour. However, our selected models are not designed to handle MNAR.

Further research could also evaluate the ability of each model to preserve the distributional characteristics of the data. For instance, one could impute across subsets, such as complete tank farms, and then compare the statistical distributions of `fill_pct` on imputed versus original data. To compare the distributions one would use visual tools such as Histograms or Q-Q graphs, or statistical tests such as a Kolmogorov–Smirnov (K-S) Test, which computes the largest absolute value of the difference between the cumulative distribution functions.

This analysis highlights the importance of rigorously diagnosing the missingness mechanism before applying imputation methods. When assumptions are well-founded, as with MAR in our case, suitable models can produce low-bias interpolations even under difficult data conditions. However, our study also illustrates the importance of validating imputation strategies within the context of the data. MICE, often considered one of the most powerful and reliable imputation models, was outperformed by linear interpolation, a much simpler baseline, due to the specific characteristics of our dataset. This underscores the risk of applying sophisticated models without properly aligning them with the structure and challenges of the data. Ultimately, this study demonstrates that imputation must be approached deliberately, guided by a clear understanding of missingness and data structure. Applying models without such context risks introducing bias and undermining the validity of results. Rigorous diagnosis, not blind application, is essential to building reliable and effective imputation strategies.

---

[1] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581.

[2] Paul D. Allison. *Missing Data*, volume 136 of *Quantitative Applications in the Social Sciences*. Sage Publications, Thousand Oaks, CA, 2002.

[3] Roderick J.A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, 1992.

[4] Roderick J.A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Hoboken, NJ, 2nd edition, 2002.

[5] Roderick J.A. Little. A test of missing completely at random for multivariate data with

missing values. *Journal of the American Statistical Association*, 83(404):1198–1202, 1988. doi:10.1080/01621459.1988.10478722.

[6] Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908. doi:10.2307/2331554.

[7] Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. doi:10.1214/aoms/1177730491.

[8] Samuel S. Shapiro and Martin B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965. doi:10.2307/2333709.

[9] Howard Levene. Robust tests for equality of variances. In Ingram Olkin, S. Ghurye, W. Hoeffding, W.G. Madow, and H.B. Mann, editors, *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pages 278–292. Stanford University Press, Palo Alto, CA, 1960.

[10] William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952. doi:10.1080/01621459.1952.10483441.

[11] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987.

[12] L. Beretta and A. Santaniello. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making*, 16(Suppl 3):74, 2016. doi:10.1186/s12911-016-0318-z.

[13] Jyoti S Shah, Ryen Lau, and Peter J Haug. Random forests and missing data for predicting clinical outcomes: a review. *Journal of Biomedical Informatics*, 50:43–52, 2014. doi:10.1016/j.jbi.2014.01.005.

[14] Daniel J. Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012. doi:10.1093/bioinformatics/btr597.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[16] D.G. Krige. *A statistical approach to some mine valuations and allied problems at the Witwatersrand*. PhD thesis, University of Witwatersrand, Johannesburg, South Africa, 1951.

[17] Antonio Briola and Tomaso Aste. Dependency structures in cryptocurrency market from high to low frequency. *arXiv preprint arXiv:2206.03386*, 2022.

Appendix A: Results

TABLE I: Performance of imputation models across five random missingness seeds. Final column reports mean ± standard deviation. Each seed is noted in the column for reproducibility

	Seed:	**42**	**9342025**	**8548609**	**4075194**	**2034376**	
Model	**Metric**						**Mean ± Std**
Mean	Bias	-0.0126	0.0101	-0.00476	-0.0101	-0.0112	-0.00571 ± 0.00932
	% Bias	-2.23	1.81	-0.84	-1.79	-2.00	-1.01 ± 1.66
	RMSE	0.281	0.289	0.295	0.302	0.292	0.292 ± 0.00773
	MAE	0.229	0.238	0.241	0.250	0.233	0.238 ± 0.00804
	R^2 Score	-1.05	-1.44	-1.24	-1.28	-1.09	-1.22 ± 0.157
Linear	Bias	-0.00786	-0.0035	-0.0167	-0.00478	0.000472	-0.00647 ± 0.00645
	% Bias	-1.39	-0.628	-2.95	-0.85	0.0845	-1.15 ± 1.14
	RMSE	0.187	0.194	0.179	0.182	0.188	0.186 ± 0.00579
	MAE	0.137	0.138	0.127	0.132	0.133	0.133 ± 0.00439
	R^2 Score	-0.089	-0.119	0.0992	0.0673	0.0496	0.00162 ± 0.0986
MICE	Bias	-0.00334	0.00463	0.00497	0.000945	0.00501	0.00244 ± 0.00366
	% Bias	-0.592	0.83	0.876	0.168	0.897	0.436 ± 0.65
	RMSE	0.295	0.298	0.303	0.310	0.294	0.300 ± 0.0066
	MAE	0.242	0.244	0.248	0.256	0.237	0.245 ± 0.00713
	R^2 Score	-0.903	-1.16	-0.892	-1.13	-0.823	-0.982 ± 0.153
KNN	Bias	-0.00711	0.00947	0.000364	-0.0061	-0.00565	-0.00181 ± 0.00695
	% Bias	-1.26	1.7	0.0642	-1.08	-1.01	-0.317 ± 1.24
	RMSE	0.273	0.276	0.288	0.289	0.286	0.282 ± 0.00737
	MAE	0.224	0.227	0.236	0.238	0.23	0.231 ± 0.00592
	R^2 Score	-1.26	-1.71	-1.31	-1.47	-1.17	-1.38 ± 0.212
RF	Bias	-0.0133	0.00615	-0.00962	-0.00257	0.00261	-0.00335 ± 0.00813
	% Bias	-2.35	1.1	-1.7	-0.457	0.467	-0.588 ± 1.44
	RMSE	0.275	0.278	0.284	0.29	0.282	0.282 ± 0.00576
	MAE	0.225	0.227	0.233	0.238	0.226	0.23 ± 0.00554
	R^2 Score	-1.14	-1.46	-1.33	-1.32	-1.12	-1.27 ± 0.143
Krige	Bias	-0.00786	0.013	-0.00504	-0.00636	-0.0124	-0.00373 ± 0.00976
	% Bias	-1.39	2.34	-0.889	-1.13	-2.21	-0.656 ± 1.75
	RMSE	0.282	0.291	0.294	0.304	0.292	0.293 ± 0.00786
	MAE	0.229	0.239	0.239	0.251	0.233	0.238 ± 0.00832
	R^2 Score	-0.985	-1.42	-1.1	-1.24	-1.05	-1.16 ± 0.173
Graph	Bias	-0.00485	0.0114	0.00311	-0.00462	-0.00632	-0.000256 ± 0.00749
	% Bias	-0.858	2.04	0.548	-0.821	-1.13	-0.0442 ± 1.34
	RMSE	0.271	0.275	0.281	0.289	0.278	0.279 ± 0.0068
	MAE	0.222	0.225	0.231	0.237	0.223	0.228 ± 0.00631
	R^2 Score	-1.21	-1.67	-1.41	-1.43	-1.22	-1.39 ± 0.188