City, University of London

MSc in Data Science

Project Report

2023/2024

Filling the Missing Data in Electricity Supply Area Dataset Using Machine Learning Methods

Ling-Yun Huang

Supervised by: Aidan, Slingsby

Submitted: 01.10.2024

Declaration

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work, I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed: Ling-Yun, Huang

Abstract

This dissertation investigates the effectiveness of various imputation methods for addressing missing values in the secondary electricity supply area dataset. Seven methods were tested: Mean/Median/Mode, k-Nearest Neighbours (KNN), Linear Regression (LR), Random Forest (RF), Multilayer Perceptrons (MLP), Inverse Distance Weighting (IDW), and Geographically Weighted Regression (GWR). This study found that MLP and RF outperformed simpler methods, significantly reducing error rates and capturing non-linear relationships in the data. However, none of the methods achieved perfect accuracy, particularly with high missing data percentages. The findings suggest that while advanced methods help improve results, there is still a need for better feature selection and more realistic simulation of missing data. Future studies should focus on selecting features that better capture important relationships and on simulating missing data patterns that mirror real-world situations. Overall, this research highlights the significance of selecting appropriate imputation methods for complex datasets and offers ways to improve prediction accuracy in the electricity supply area.

Keywords: Imputation Methods, Missing Data, Multilayer Perceptrons, Electricity Supply Area.

Contents

Chapter 1 Introduction and Objectives

1.1 Background

In many fields, having accurate and complete data is crucial for making precise predictions and better decisions. This is especially true in electricity supply areas. Reliable data is vital for ensuring smooth operations, effective planning, and efficient grid management in these systems (Department for Energy Security and Net Zero, 2022). However, missing data is a common challenge in real-world situations. It can arise from several reasons in electricity supply areas. Poor data collection methods can result from disorganised processes or caused by outdated technology. Sensor malfunctions and technical problems can also lead to missing data. Additionally, human errors during data entry or transmission can introduce gaps or mistakes in the records.

The complexity of modern electricity systems further complicates this issue. These systems handle vast amounts of data, increasing the likelihood of errors and data loss. As these systems become more advanced, ensuring that all data remains accurate and complete becomes even more challenging. Without reliable data, managing the electricity grid is difficult, as decisions may be based on incomplete information. Therefore, addressing the issue of missing data is crucial for effective operation and unbiased decision-making in the field.

In this project, we are working with Advanced Infrastructure Technology Ltd (AITL). The company focuses on using datasets and relevant tools to support decarbonisation. AITL provides the necessary information for planning energy transitions. Their goal is to make low-carbon power easier to manage and implement. By working with AITL, we aim to improve the accuracy and availability of data. This will help energy systems run more efficiently and support smoother operations. Ultimately, this collaboration will lead to better planning and a more sustainable future.

The simplest way to handle missing values in a dataset is to delete the entire row containing the missing data (Kaiser, 2014). This method is very easy and quick to apply. However, it can result in significant data loss, especially if the percentage of missing values is high. With fewer data points, it might reduce the power of analysis, affecting the robustness of the findings and introducing bias into the analysis. Other methods for handling missing values might offer a more balanced solution.

There are various methods for handling missing values, each with its strengths and limitations. Simple methods, like mean, median, or mode imputation, are easy to apply but do not consider complex patterns in the data. Feature-based methods, such as k-Nearest Neighbour imputation, Linear Regression, Random Forests, and Multilayer Perceptrons, use relationships between features and the target variable to impute the missing values. Spatial methods, such as Inverse Distance Weighting, use the location of data points to estimate missing values. Additionally, some methods combine both feature-based and spatial information to predict the missing values, such as Geometrically Weighted Regression.

To deal with the challenges of missing values in electricity supply areas, further research is needed. With a variety of imputation methods available, it is important to evaluate which one works best with the specific datasets used in this field. This research aims to identify the best method for imputing missing values and improving the accuracy of analyses. By finding the most suitable imputation methods, the research will help ensure more reliable data, supporting better operational efficiency and decision-making in electricity supply systems.

Research Questions:

- How do different imputation methods for electricity supply area datasets compare in terms of accuracy, efficiency, and suitability?
- What key factors influence the performance and optimisation of these methods to enhance their effectiveness?

1.2 Objectives

The main goal of this dissertation is to evaluate existing imputation methods for handling missing values in electricity supply area data. This study will focus on various methods, including k-Nearest Neighbours (KNN), Inverse Distance Weighting (IDW), and Multilayer Perceptrons (MLP). Each method has its strengths and limitations, and how well it works can vary based on the data type. This research aims to assess and compare these methods by analysing their performance on real electricity supply area datasets. Statistical metrics will be used to measure their accuracy and reliability.

Additionally, the research will investigate key factors that influence the performance of these imputation methods. For example, it will look at the number of neighbours in KNN, the weighting in IDW, and the number of neurons in MLP. Understanding these factors is crucial for optimising these imputation methods. The goal is to enhance the accuracy and reliability of the imputed data by fine-tuning these key elements.

By applying the most suitable imputation model to the actual dataset, the study aims to create a more reliable dataset for electricity supply areas. This improved data can then be used for more accurate analyses, which will support better decision-making and increase operational efficiency in electricity supply systems. Ultimately, the research will provide recommendations on the best imputation methods for similar data in the electricity sector.

In summary, the objectives of this dissertation are to evaluate and compare imputation methods for handling missing values in electricity supply area data. The study will optimise these methods by analysing key factors to enhance performance. Finally, it will provide recommendations for the best methods to manage missing values in electricity supply systems.

1.3 Beneficiaries

This research will benefit electricity network management by providing a comprehensive evaluation of imputation methods for missing data in the electricity supply area dataset. By identifying the most effective methods for filling in missing values, the study will help create more accurate and complete datasets. These improved datasets will enhance network management and planning. As a result, they will lead to better decisions and more efficient operations.

This research will also benefit the broader research community. It will provide insights into different imputation methods and their performance, especially in electricity supply data. This study will explore different imputation methods and key factors that impact their performance. It will provide useful insights for researchers with similar datasets. This will help them choose the best imputation methods, improving the accuracy and reliability of their analyses.

In summary, this research will support better management of electricity networks. It will also advance data analysis methods, particularly in areas dealing with missing data. Both electricity sector practitioners and researchers will benefit from these findings. This will lead to more accurate analyses and improve decision-making.

1.4 Work Plan

The work plan for this dissertation has six phases. Each phase depends on the one before it, ensuring a thorough and organised approach.

Phase 1 – Literature Review: This phase involves a thorough review of existing literature on data imputation methods, such as KNN, MLP, and IDW. The aim is to identify strengths, limitations, and gaps in current research. This will provide a foundation for the following phases.

Phase 2 – Data Preparation: In this phase, the focus is on understanding the dataset. It includes selecting a relevant dataset and creating the necessary features for later prediction. This phase is important for the subsequent analysis.

Phase 3 – Imputation Methods Application: This phase involves applying the selected imputation methods identified in Phase 1. Each method will be carefully considered, considering its key factors, limitations, and assumptions.

Phase 4 – Evaluation and Comparison: After applying different methods, evaluating their performance is crucial. This phase will focus on comparing the models, understanding how they perform, and identifying the key factors for each method. And with the results, this phase will also choose the best model for the next phase.

Phase 5 – Application of Actual Missing Values and Evaluation: In this phase, the chosen method will be applied to fill the real missing values in the dataset. The accuracy of the result will also be validated using another dataset. This ensures that the method works well in a real-world context.

Phase 6 – Reporting: The final phase involves documenting the entire research process. The results will provide a clear and concise summary of the findings.

1.5 Report Structure

This chapter provided the study's background, research questions, and objectives. It also outlined the beneficiaries, work plan, and report structure. There are five more chapters in the report.

Chapter 2 – Context: This chapter covers a literature review of existing imputation methods, discussing their strengths, limitations, and key factors. It also includes a comparison and evaluation of these methods, along with the metrics used for evaluation.

Chapter 3 – Methods: This chapter details the methods used in the study. This includes data collection, data preprocessing, and the imputation methods application. Along with the implementation of these methods, the model evaluation, and the application of the best model to actual missing values.

Chapter 4 – Results: This chapter presents the outcomes of the analysis. This includes the performance evaluation of each imputation method, the selection of the best model, and the validation results after applying the best model to real missing values.

Chapter 5 – Discussion: This chapter examines the results from Chapter 4 and discusses their relevance to the study’s objectives and research questions. It evaluates the suitability of the chosen imputation methods and the results of the imputed data.

Chapter 6 – Evaluation, Reflections and Conclusions: This chapter provides an overall evaluation of the research. It reflects on the study process and outcomes. Also summarises the key findings and suggests recommendations for future research.

1.6 Conclusion

In conclusion, this dissertation addresses the challenge of missing values in electricity supply area dataset. By investigating various imputation methods, this research aims to identify the best approach to enhance data accuracy. In the next chapter, we will review the relevant literature on existing imputation methods and discuss research that has compared these different techniques.

Chapter 2 Context

In this chapter, we will review the relevant literature to this study. We will begin by exploring various imputation methods for handling missing data. Then, we will examine studies that have compared these methods, mentioning the evaluation metrics they used, with a focus on the conclusions they made. This review will provide a foundation for understanding the approaches used in this research.

2.1 Imputation Methods

When dealing with missing values, there are many methods to choose from. In this section, we will explore the methods that will be used in the later analysis. We will focus on defining each method, identifying their key factors, and discussing their advantages and disadvantages. By understanding these aspects, we can better assess which methods are most suitable for our data and analysis needs.

2.1.1 Mean / Median / Mode Imputation

This method replaces missing values with the mean, median, or mode from the non-missing values in the training dataset. The choice between mean, median, or mode depends on the distribution of the data (Lodder, 2014). It is a quick and simple method for handling missing values. However, it is quite basic and may miss complex relationships in the data. As a result, it often performs worse than more advanced methods. Many studies (Jadhav et al., 2019; Kokla et al., 2019; Schmitt et al., 2015) use this approach as a baseline for comparing with more sophisticated imputation methods.

2.1.2 *k*-Nearest Neighbour Imputation

This method predicts missing values by identifying the *k* closest data points to the target data using distance metrics, such as Euclidean distance, from the training data (Beretta and Santaniello, 2016). By leveraging other features, *k*-Nearest Neighbour (KNN) can capture valuable information that simpler methods like mean or median imputation might miss. However, this also makes it sensitive to irrelevant features. If unrelated data is included, it can lead to less accurate imputation (Parr et al., 2008). Additionally, KNN is computationally intensive, especially when dealing with large datasets or data with many features.

When applying KNN, several factors need to be considered. First, standardising the features is crucial (Manimekalai and Kavitha, 2018). This ensures that all features have the same scale when calculating distances, preventing skewed results. Without standardization, features with larger scales could dominate the distance metric, leading to incorrect nearest neighbours. Another important consideration is the choice of *k* (Abidin et al., 2018). There is no single *k* value that works best in all situations. The optimal k depends on the specific dataset and requires careful analysis. A smaller *k* may capture more local patterns, while a larger *k* might generalise better but risk including irrelevant neighbours. Choosing the right *k* is essential for accurate imputation.

2.1.3 Linear Regression Imputation

Linear regression predicts missing values by modelling the relationship between the target variable and other features. This approach often provides more accurate estimates than simpler methods like mean or median imputation. Linear regression, however, assumes that relationships between features are linear, which may not always capture complex patterns in the data (Pigott, 2001). The method can also be sensitive to multicollinearity, where high correlations between features can make estimates less reliable. Techniques like Variance Inflation Factor (VIF) analysis can help identify and address this issue (Martin, 2022). Additionally, outliers can distort the regression model and affect imputation accuracy. While linear regression can be effective, it is important to be mindful of these factors to ensure robust and reliable imputation.

2.1.4 Random Forest Imputation

Random Forest employs a collection of decision trees to predict missing values. It builds multiple decision trees on different subsets of the data and averages their results, making predictions more robust and generalised (Breiman, 2001). This approach captures complex patterns and reduces the risk of overfitting. Compared to Linear Regression, Random Forest can handle non-linear patterns and complex interactions more effectively (Bastos et al., 2024). While Linear Regression can struggle with non-linearity and outliers, Random Forest provides more robust predictions by averaging from multiple trees. This helps minimise the impact of individual data points and reduces overfitting.

Although Random Forest has advantages like handling non-linear relationships and being robust to outliers, it also has drawbacks. It is computationally intensive due to the need to build multiple decision trees and is less interpretable (Dou et al., 2019). Additionally, it requires careful parameter tuning, such as choosing the forest size and the maximum depth of each tree (Catani et al., 2013). These factors can lead to high memory usage and longer processing times, which may be a limitation for large datasets.

2.1.5 Multilayer Perceptrons Imputation

The Multilayer Perceptrons (MLP) is an artificial neural network made up of fully connected layers (Kuligowski and Barros, 1998). It has an input layer, one or more hidden layers, and an output layer. The input layer has a number of neurones equal to the number of input features. The hidden layers contain neurones that apply weights and activation functions to process the data. In the output layer, predictions are generated; for regression tasks, like in this study, there is typically one neurone in this layer to predict the continuous target variable.

MLP has been a popular choice in many studies (Silva-Ramírez et al., 2011; Sundararajan and Sarwat, 2020; Tkachenko et al., 2019). Compared to Linear Regression, MLP captures complex or nonlinear patterns in data. However, to perform well, MLP needs careful tuning of key parameters (Guo et al., 2022). These include the number of hidden layers, neurones per layer, activation functions, and learning rates. MLP offers more flexibility, but it can be computationally demanding. Without proper regularisation like early stopping, it may overfit, especially with smaller datasets (Lawrence and Giles, 2000). While its predictive power is strong, using MLP requires attention to these factors for optimal performance.

2.1.6 Inverse Distance Weighting Imputation

Inverse Distance Weighting (IDW) is a spatial interpolation method that uses distances from known points to estimate missing values. In IDW, closer points have more influence on predictions, while distant points contribute less. This approach works well for geographical data where location is a key factor (Barudžija et al., 2024). Weights are assigned inversely proportional to distance, giving nearby points more weight. Compared to methods like Linear Regression or Multilayer Perceptrons, IDW is simpler and easier to apply. However, relying only on spatial information can make it less effective when the data contains complex patterns. IDW may miss relationships between other features that could improve accuracy.

The method's performance depends heavily on the number of neighbours used and how distances are weighted (Liu et al., 2021). A careful selection of both is needed to avoid biased predictions. If too few neighbours are used, the imputation may be too localised. Using too many neighbours can lead to overly general predictions (Baker et al., 2014). Thus, fine-tuning these settings is key to achieving the best results with IDW.

2.1.7 Geometrically Weighted Regression Imputation

Geographically Weighted Regression (GWR) is a spatial statistical method that combines spatial information with feature relationships. Unlike global models that use a single equation for the whole dataset, GWR estimates different coefficients for each location. Each data point in GWR has its own set of regression parameters (Matthews and Yang, 2012).

Compared to Linear Regression and IDW, GWR offers a more integrated approach (Wang et al., 2017). While Linear Regression provides a single global model, and IDW relies only on spatial information, GWR uses both spatial data and feature relationships for predictions. This can lead to higher accuracy if both types of information are valuable (Cullen and Guida, 2021). However, like LR, GWR may struggle with capturing non-linear or complex patterns in the data (Lin and Wen, 2011). The performance of GWR depends on selecting the right bandwidth and weight function (Khosravi and

Balyani, 2019). Tuning these parameters is crucial for balancing local accuracy and model stability. Overall, GWR provides more flexibility for spatial analysis but requires careful adjustment to achieve the best results.

2.2 Comparison of Methods

Numerous studies have compared different methods for handling missing values across various fields. Mean/Median/Mode is often used as a baseline due to its simplicity. Ideally, other methods should perform better than this basic approach (Hron et al., 2010; Jadhav et al., 2019; Kokla et al., 2019). More advanced methods are expected to capture complex patterns that simple imputation techniques cannot.

Additionally, several studies focus on specific method comparisons. For example, Osman et al. (2018) compared KNN and Neural Networks in water distribution system. Mamat and Mohd Razali (2023) examined KNN, Linear Regression, and Random Forest, showing how these methods differ in performance. Furthermore, MLP and Random Forest are often compared because of their complexity and ability to capture non-linear patterns effectively (Sundararajan and Sarwat, 2020; Tkachenko et al., 2019).

IDW, GWR, and Linear Regression are often compared due to their different methods of handling relationships (Wang et al., 2017). Both IDW and GWR use spatial information, but IDW relies only on spatial patterns. In contrast, GWR allows for different relationships based on location. Chen et al. (2012) found that GWR often performs better than IDW. Linear Regression is also frequently compared with GWR. While Linear Regression assumes a global linear relationship across the dataset, GWR captures local variations that Linear Regression might miss. Shi et al. (2022) showed that GWR performs better than Linear Regression when spatial patterns are important.

Generally, more advanced methods tend to perform better in handling missing values. For example, Silva-Ramírez et al. (2011) found that MLP demonstrated the best performance across different datasets compared to Linear Regression and mean/mode imputation. Sundararajan and Sarwat (2020) also showed that Random Forest performed best among the imputation methods tested including KNN and mean imputation. However, this is not always the case. Singh et al. (2017) showed that KNN and Random Forest performed similarly, with KNN being less complex and requiring less time to train.

Furthermore, it is important to note that no single method works best in every situation. Zhang et al. (2005) and Hagenauer and Helbich (2022) found that GWR performed better than MLP, while Emamgholizadeh et al. (2017) showed the opposite result. This highlights the need for careful method selection based on the specific dataset and task.

During the comparison with numerical targets, evaluation metrics are used to assess model performance. The most common metrics are Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These metrics are also commonly used in the studies cited above.

RMSE calculates the average squared differences between predicted and actual values. This makes it particularly sensitive to large residuals (Nevitt and Hancock, 2000). In contrast, MAE measures the average absolute difference. It treats all errors equally, providing a straightforward view of overall performance (Willmott and Matsuura, 2005). MAPE calculates the average absolute percentage difference. This offers a relative measure of prediction accuracy as a percentage, making it a desirable choice. However, MAPE can be less reliable when actual values are small or approach zero. In these cases, even a minor prediction error can lead to a disproportionately large percentage error. This can result in potentially misleading conclusions (de Myttenaere et al., 2016). Therefore, when using this metric, it is important to consider the distribution of the dataset (Chicco et al., 2021).

In addition to these global measures, it is important to explore non-global metrics that focus on specific areas within the dataset (Bondarenko and Raghunathan, 2016). For instance, evaluating the proportion of extreme values can provide insights into the accuracy of estimates in critical regions. Plotting residuals and examining their distribution through histograms or spatial mapping can reveal patterns that are not apparent from summary statistics alone. Such visualisations can help identify areas where estimates are less accurate or where imputation methods may struggle.

2.3 Conclusion

In this chapter, we reviewed various imputation methods for handling missing data. We discussed their definitions, advantages, limitations, and the key parameters that require careful assessment. We also discussed the importance of method selection based on the specific dataset and task. The comparative studies reviewed indicate that while advanced methods tend to perform better, no single approach guarantees to outperform others in every situation. Evaluation metrics are crucial for assessing model performance, however, careful consideration of their strengths and weaknesses is also necessary.

The next chapter will delve into the methods that will be applied in this project. This includes a discussion of the data sources, preprocessing steps, implementation details, and the specific imputation methods employed.

Chapter 3 Methods

This chapter outlines the methods used for imputing missing values in the secondary electricity supply area dataset. It includes data preparation, the application of various imputation methods, and an evaluation of model performance. Beginning with a description of the data gathering and preprocessing steps, followed by exploratory data analysis. We also discuss the implementation of different imputation techniques. Finally, the selected methods are applied to real missing values, and the results are validated using an external dataset.

3.1 Data Gathering

The data used in this study were sourced from Advanced Infrastructure Technology Ltd (AITL) Company and supplemented by publicly available datasets. The main dataset is secondary electricity supply area data, with six additional datasets providing extra information.

Main Dataset

- **Secondary Electricity Supply Area (ESA)**:
  This dataset is provided by AITL Company. It includes detailed information on Capacity Rating and Peak Load for each area. These areas cover five local authorities (Cherwell, Oxford, South Oxfordshire, Vale of White Horse, and West Oxfordshire) in the UK. Capacity Rating refers to the maximum load each area can handle, while Peak Load represents the highest level of electricity demand recorded in each area.

Additional Datasets

- **MSOA Non-Domestic Electricity Consumption (MSOA)**:
  This dataset was sourced from a government dataset (Department for Energy Security and Net Zero, 2024). It provides non-domestic electricity consumption data at the MSOA (Middle Super Output Area) level for 2022. It is useful for aggregating the non-domestic consumption in each area in the main dataset.

- **LSOA Domestic Electricity Consumption (LSOA)**:
  This government dataset (Department for Energy Security and Net Zero, 2024) provides domestic electricity consumption data at the LSOA (Lower Super Output Area) level for 2022. It is useful for aggregating domestic electricity consumption across the areas in the main dataset.

- **Postcode Domestic Electricity Consumption (Postcode):**
  This dataset provides domestic electricity consumption data similar to the LSOA dataset. However, it only includes postcodes with consumption over 100 kWh. While this allows for a more detailed

analysis of postcode level, it may miss some lower-consumption data, leading to potential gaps in the overall analysis.

- **Customer Vulnerability (CV)**:
  This dataset contains population information for larger geographic areas. It helps in understanding the demographics of the regions. This is useful for aggregating the population in the main dataset.

- **OS AddressBase Premium (OS)**:
  This dataset provides detailed information about building locations and types. It includes address-based data that specifies whether buildings are residential, commercial, or of other types. This information is crucial for analysing building numbers and density within each area. It also supports more accurate aggregation and analysis when combined with other datasets.

- **Primary Electricity Supply Area (PESA)**:
  This dataset represents a higher-level view of the electricity supply network compared to the secondary supply area dataset. It will be used to validate the imputed data once the missing values have been filled.

The selection of these support datasets was primarily guided by expert advice from AITL. Their experience in the energy sector helped identify key factors that influence the two main target variables, Capacity Rating and Peak Load. AITL emphasised the importance of focusing on features that are directly related to electricity consumption, such as demographic proportions and historical consumption patterns.

3.2 Data Preprocessing

Since the datasets have different boundaries, careful consideration is needed to define features that effectively predict our targets. The following explains how each feature is created.

- **Non-Domestic Electricity Consumption:**
  For each MSOA, we first use the OS dataset to calculate the number of commercial buildings, stored as “Commercial Building (MSOA)”. We then compute the non-domestic electricity consumption per building by dividing the total electricity consumption by the number of commercial buildings in each MSOA. This gives us the non-domestic electricity consumption per building “Non-Domestic Electricity Consumption (per Building)”.

$$
\begin{aligned}
&Non\text{–}Domestic\ Electricity\ Consumption\ (per\ Building) \\
&\quad = {Total\ Non\text{–}Domestic\ Electricity\ Consumption\ (MSOA)}\Big/{Commercial\ Building\ (MSOA)}
\end{aligned}
$$

Next, we calculate the number of commercial buildings for each ESA using the OS dataset, recorded as "Commercial Building (ESA)." Finally, we determine the non-domestic electricity consumption for each ESA by multiplying the non-domestic consumption per building and the commercial building in each ESA.

$$\begin{aligned}&\boldsymbol{Non\text{–}Domestic\ Electricity\ Consumption}\\&\qquad = Non\text{–}Domestic\ Electricity\ Consumption\ (per\ Building) \times Commercial\ Building\ (ESA)\end{aligned}$$

- **Domestic Electricity Consumption:**
  This feature is calculated similarly to non-domestic electricity consumption but uses the LSOA dataset and the number of residential buildings. We compute the domestic electricity consumption per building by dividing the total domestic electricity consumption by the number of residential buildings in each LSOA. This result is then aggregated for each ESA.

- **Postcode Domestic Electricity Consumption:**
  For each ESA, sum the domestic electricity consumption from all postcodes within that ESA.

- **Population:**
  We aggregate the population for each ESA based on the number of residential buildings. First, calculate the average population per residential building using the CV and OS datasets. Then, multiply this average by the number of residential buildings in each ESA to determine the total population.

- **Building Number:**
  Using the OS dataset, we calculate the total number of buildings in each ESA, regardless of their type. This count includes all buildings within the ESA boundary.

- **Building Density:**
  We calculate building density by dividing the total number of buildings in each ESA by the size of the area in square kilometres (km^2). This helps assess the density of buildings within each area.

To ensure that features with different scales do not affect the model unfairly, we standardise the features where necessary. We use z-score normalisation for this. This method changes the features so that they have a mean of 0 and a standard deviation of 1. The formula is as follows:

$$z = \frac{X - \mu}{\sigma}$$

Where X represents the original value of the feature, μ is the mean of the feature, σ is the standard deviation of the feature, and z is the normalised value.

3.3 Exploratory Data Analysis

The main dataset includes 6,007 data points with two target variables, Capacity Rating and Peak Load. It also contains geometric information for each data point. Among these, 4,720 data points are complete data, while 1,287 data points are missing in both target variables. Figure 3.1 illustrates the geographical distribution of these data points, where missing data points are highlighted in pink, and complete data points are shown in blue.

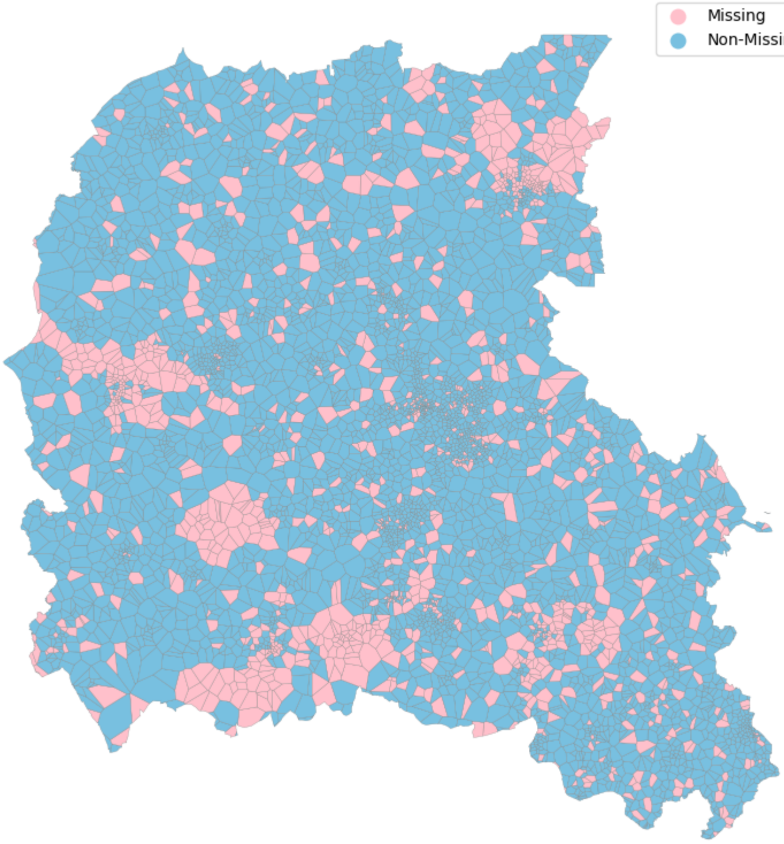


Figure 3.1 Geographical Distribution of Missing and Non-missing Data.

Capacity rating represents the maximum amount of electricity each supply area can deliver under optimal conditions. Peak load measures the highest amount of electricity consumed within each supply area. Table 3.1 presents the summary statistics for capacity rating and peak load from non-missing data. The mean capacity rating is 271.21, with a standard deviation of 254.52, showing that the values vary widely. For peak load, the mean is 95.73, with a higher relative standard deviation of 122.75, showing more variability in demand across different areas. An important observation is that while most peak load values are below 1,000, there is one extreme outlier with a maximum value of 4,053.34. This extreme value could affect the analysis, so it will be excluded in the next steps to prevent it from distorting the results.

	Capacity Rating	**Peak Load**
Count	4720	4720
Mean	271.21	95.73
Std	254.52	122.75
Min	1	0
25%	50	10.88
50%	200	51.72
75%	500	149.50
Max	1500	4053.34

Table 3.1: Summary Statistics for Capacity Rating and Peak Load on Non-missing Data Points.

Figure 3.2 shows the distribution of the target variables after excluding the extreme data point. The histograms reveal that both variables show skewness. For capacity rating, there is a noticeable concentration of values around 500, with a second peak at this value. This skewness and concentration may make prediction more difficult because the model might not handle the variety of values well. Peak load also shows a skewed distribution, which could similarly make predictions less accurate across the range of values.

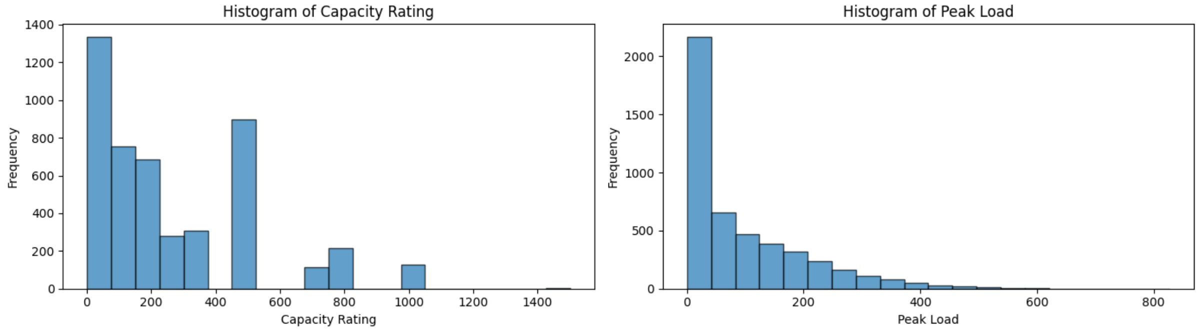


Figure 3.2 Histogram of Capacity Rating and Peak Load on Non-missing Data Points.

Figure 3.3 shows the heatmap of the correlation matrix for the target variables and features. The heatmap reveals relationships between these variables. Building Number and Postcode Domestic Electricity Consumption have the strongest correlations with both capacity rating and peak load, with values of 0.59 and 0.64, and 0.51 and 0.63, respectively. This indicates that these features significantly influence the target variables. Conversely, Non-Domestic Electricity Consumption and Building Density have more moderate correlations, suggesting a weaker impact on target variables. The very high correlation between Building Number and Postcode Domestic Electricity Consumption may affect models like linear regression, which could be influenced by multicollinearity. This needs to be addressed to ensure accurate and reliable model performance.

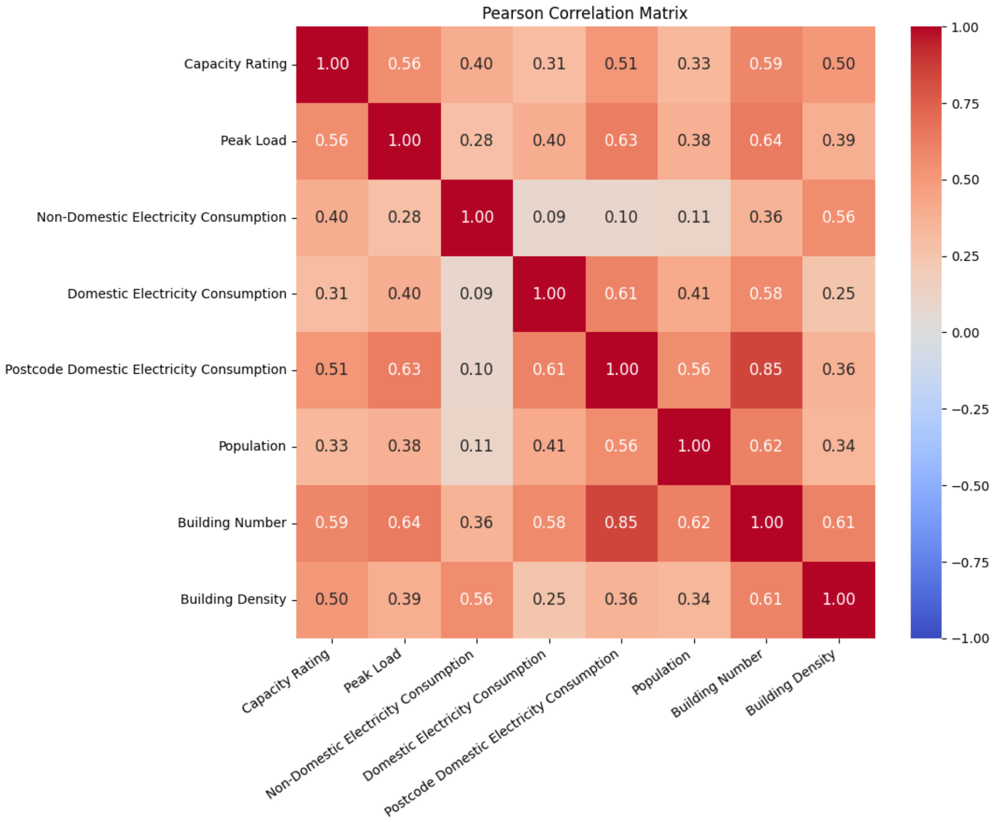


Figure 3.3 Heatmap of Pearson Correlation for Target Variables and Features.

3.4 Implementation

This project will use non-missing data points to train and test different imputation methods. The goal is to identify the best model for imputing missing values. The dataset with non-missing values will be split into three subsets, 70% for training, 15% for validation, and 15% for testing. The training set will be used to build and train the models. The validation set will be used in tuning the parameters by evaluating their performance and selecting the best settings for each method. And the test set will be used to select the overall best-performing model.

Three evaluation metrics will be used in this study, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These metrics assess model performance by comparing predicted values with actual values. RMSE calculates the square root of the average squared differences. MAE calculates the average absolute differences. And MAPE expresses prediction errors as a percentage of actual values. The formulas are as follows:

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

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

$$MAPE = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$$

Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of data points.

Because the peak load data includes values of 0, using MAPE directly would result in division by zero, which leads to undefined or infinite values. To address this, we will add 1 to the actual values when calculating MAPE to avoid such errors. Additionally, since both capacity rating and peak load contain many small values, MAPE may not be as reliable as RMSE or MAE for evaluating model performance. Therefore, RMSE and MAE will be the primary metrics used in this study, with MAPE serving as a supplementary metric for comparison.

During model training, we will closely evaluate performance on the validation set using these metrics. By comparing the results, we will identify the best parameter settings for each model to achieve optimal results before comparing different methods on the test set. Finally, we will compare the models' performance on the test set to determine the best one. The model with the lowest RMSE, MAE, and MAPE on the test set will be considered the best-performing model and selected for imputing the actual missing values.

3.5 Imputation Methods

This section detailed the processes used for each imputation method in this study. Building on the implementation and feature creation discussed earlier to ensure a fair comparison of these methods and identify the best performance model for the dataset.

3.5.1 Mean / Median / Mode Imputation

With this method, we first calculate the mean, median, and mode using the training set from the non-missing data. After determining these statistics, we apply them in the validation set. We then compare the predicted values with the actual values in the validation set to assess the performance of each statistic. The method that performs best will be used as a baseline for comparison with other imputation methods applied to the test set.

3.5.2 *k*-Nearest Neighbour Imputation

The *k*-Nearest Neighbour (KNN) estimates missing values by using the most similar *k* data points. In this study, six features are used to determine similarity. The Euclidean distance is calculated to identify the closest *k* neighbours, and the average of these nearest neighbours' values is used for imputation. Features are standardised before calculating the distance to ensure a fair comparison. The validation set is used to select the optimal *k* value with the best performance from 3, 5, 7, and 9. This setting is then applied to the test set to compare KNN with other imputation methods.

3.5.3 Linear Regression Imputation

Linear regression estimates missing values by modelling the relationship between the target variable and other features. This approach uses the linear associations among variables to predict missing data. In this study, we first address multicollinearity by performing a Variance Inflation Factor (VIF) analysis on the six features to ensure reliable estimates (Martin, 2022). After addressing multicollinearity, linear regression is applied to impute missing values. The performance of the imputation is then evaluated using RMSE, MAE, and MAPE.

3.5.4 Random Forest Imputation

Random Forest builds multiple decision trees on different subsets of the data and averages their results to predict missing values. This method captures complex patterns and enhances prediction accuracy. In this study, all six features will be used for prediction. Random Forest will be tested with varying forest sizes (10, 100, and 500 trees) and maximum depths (15 and 30). By comparing the evaluation results from these settings on the validation set, we aim to identify the best combination of forest size and depth for accurate imputation. The optimal configuration will then be applied to the test set for comparison with other methods.

3.5.5 Multilayer Perceptrons Imputation

Multilayer Perceptrons (MLP) is a neural network model that learns complex patterns through backpropagation and adjusts weights based on errors. In this study, all six features will be used in the models for prediction, the MLP will use the ReLU (Rectified Linear Unit) activation function and the Adaptive Moment Estimation (Adam) optimiser. The model will be trained for a maximum of 1000 iterations. Early stopping will be employed during training with a validation fraction of 0.1 to prevent overfitting. Various configurations will be tested, including different numbers of neurones in the hidden layers ((5, 5), (15, 15), and (30, 30)) and learning rates (0.005, 0.01, 0.05, and 0.1). The best-performing configuration, based on validation results, will be applied to the test set for comparison with other imputation methods.

3.5.6 Inverse Distance Weighting Imputation

Inverse Distance Weighting (IDW) is a spatial imputation method that uses geometric distances to estimate missing values from known data points. In this study, the centre of each geometric region is extracted as the location of the data points, and distances are calculated using Euclidean distance. The weighting is determined by the inverse of the squared distances, and the weights are normalised to predict the missing values. The formula used to calculate the weight for each point before normalisation is as follows:

$$w_i = \frac{1}{d_i^2}$$

Where:

- w_i represents the weight of the known point P_i,
- d_i is the Euclidean distance between the target point and P_i.

Different numbers of neighbours (3, 5, 7, and 9) will be tested. The configuration with the best performance on the validation set will be applied to the test set for later comparison.

3.5.7 Geometrically Weighted Regression Imputation

Geographically Weighted Regression (GWR) is a spatial imputation method that combines geographic information with feature relationships to predict missing values. GWR benefits in estimating separate coefficients for each location, allowing for localised modelling. In this study, the centroid of each geometric region was extracted to represent the location of data points, similar to IDW. And an adaptive bi-square kernel is used for spatial weighting.

The bandwidth is selected based on performance metrics from the validation set. Different bandwidth values (50, 100, 200, and 400) are tested to identify the optimal setting for the imputation task. The configuration with the best performance on the validation set is applied to the test set for comparison. The results are evaluated using metrics such as RMSE, MAE, and MAPE to determine the effectiveness of the GWR model.

3.6 Applying Imputation Methods to Real Missing Values

After determining the best-performing model through the validation and testing process, the next step is to apply the imputation method to the real missing values in the dataset. To capture the full distribution and relationship within the dataset, we will retrain the models using all non-missing data.

Since the true values for the missing data are unknown, directly evaluating the imputation accuracy is not possible. In this study, we will compare the imputed results to known datasets to assess consistency. We will also examine whether the combined data, which includes both imputed and non-missing values, is influenced by the imputed results.

The headroom percentage will also be evaluated to check if the imputed values are realistic. Specifically, we will ensure that the capacity rating remains higher than the peak load, as expected in real-world scenarios. Additionally, we will assess whether the combination of the two target variables aligns with practical expectations after imputation.

$$Headroom\,\% = {}^{(Capacity\ Rating - Peak\ Load)}\!/_{Capacity\ Rating}$$

To validate the imputed peak load values, we used data from the PESA and compared it with the ESA. We summed the peak load values from all ESAs associated with each PESA and compared these totals with the original peak load data from the primary areas. This comparison assessed how closely the imputed data matched the original values. A scatter plot will be used to visualise the relationship between the aggregated secondary peak loads and the primary peak loads.

3.7 Conclusion

This chapter detailed the methods for imputing missing values in the secondary electricity supply area dataset. It covered data gathering, preprocessing steps, and the implementation of several imputation methods, including Mean/Median/Mode, KNN, Linear Regression, Random Forest, MLP, IDW, and GWR. We also discussed the evaluation metrics used to compare these methods and the approach to applying the best-performing model to real missing values. In the next chapter, we will present the results of these imputation methods, analyse their performance, and discuss how well they addressed the missing values in the dataset.

Chapter 4 Results

This chapter presents the results of the imputation methods applied to the dataset. The non-missing data was divided into training, validation and testing sets. The training set was used to train models, while the validation set was used to fine-tune the parameters for each method. The best models were then tested on the test set to find the one with the best performance. The chosen model will be used to handle the actual missing values in the dataset.

Performance was evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). RMSE and MAE were the primary metrics due to their sensitivity to the distribution of values, while MAPE provided additional support for comparison.

4.1 Results for Capacity Rating Imputation

This section presents the results of the evaluation of various imputation methods applied to the capacity rating. The aim is to assess the effectiveness of each method in predicting missing values and to determine the optimal approach based on performance metrics.

4.1.1 Mean / Median / Mode Imputation

The mean, median, and mode were calculated from the training set, resulting in values of 272.56, 200, and 500, respectively. These statistics were then used to predict values for the validation set, and the predicted values were compared with the actual values.

Method	***Value***	**RMSE**	**MAE**	**MAPE**
Mean	272.56	245.90	202.97	328.62%
Median	200.00	254.39	189.99	236.38%
Mode	500.00	339.69	293.65	633.97%

Table 4.1: Evaluation Metrics for Mean, Median, and Mode Imputation Methods on Capacity Rating.

Table 4.1 displays the evaluation results for these methods. Mean imputation resulted in an RMSE of 245.90, an MAE of 202.97, and a MAPE of 328.62%. While mean imputation provided reasonable performance, its error metrics were relatively high. Median imputation achieved an RMSE of 254.39, an MAE of 189.99, and a MAPE of 236.38%, showing slightly better performance. Mode imputation had the highest error metrics: RMSE of 339.69, MAE of 293.65, and MAPE of 633.97%. Among these, median imputation performed the best and will be used as the baseline model for comparison on the test set.

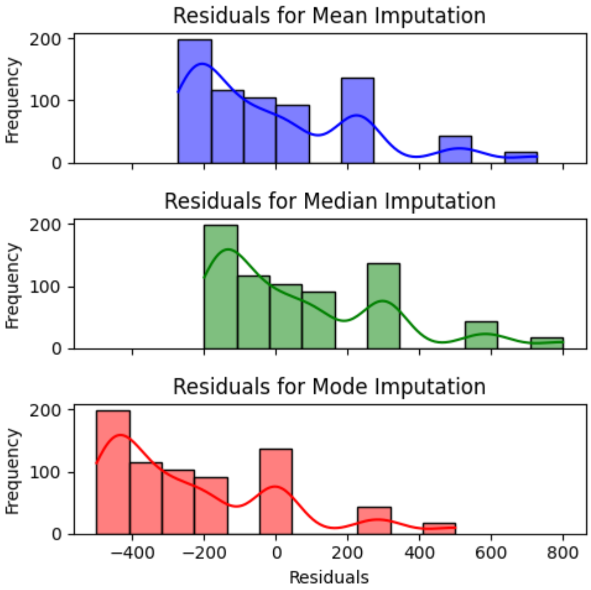


Figure 4.1 Histograms of Residuals for Mean, Median, and Mode Imputation Methods on Capacity Rating.

In Figure 4.1, the residual plots for these methods show similar patterns, reflecting the overall dataset distribution with a vertical shift based on the imputation values. Mean and median imputation residuals are centred around zero, while mode imputation shows a larger shift, with most residuals falling below zero.

4.1.2 *k*-Nearest Neighbour Imputation

The performance of different values of *k* in *k*-Nearest Neighbour (KNN) imputation was evaluated to determine the most effective configuration. Table 4.2 shows that *k = 7* results in the lowest RMSE (162.81) and MAPE (129.88%), and its MAE (114.87) is very close to the lowest value. This indicates that *k = 7* provides the best overall performance for KNN imputation.

k	**RMSE**	**MAE**	**MAPE**
3	174.02	121.96	136.55%
5	165.99	117.88	132.14%
7	162.81	114.87	129.88%
9	164.59	114.55	134.62%

Table 4.2: Evaluation Metrics for Different Values of 'k' in KNN on Capacity Rating.

Figure 4.2 illustrates that *k = 7* has residuals more evenly distributed around zero and fewer extreme residuals compared to other *k* values. In contrast, *k = 3* performs worse, showing the highest evaluation metrics and a wider distribution of residuals in the histogram. This wider spread indicates higher variability and less stability in predictions. Thus, *k = 7* is identified as the optimal choice for KNN imputation on Capacity Rating and will be applied to the test set for comparison with other methods.

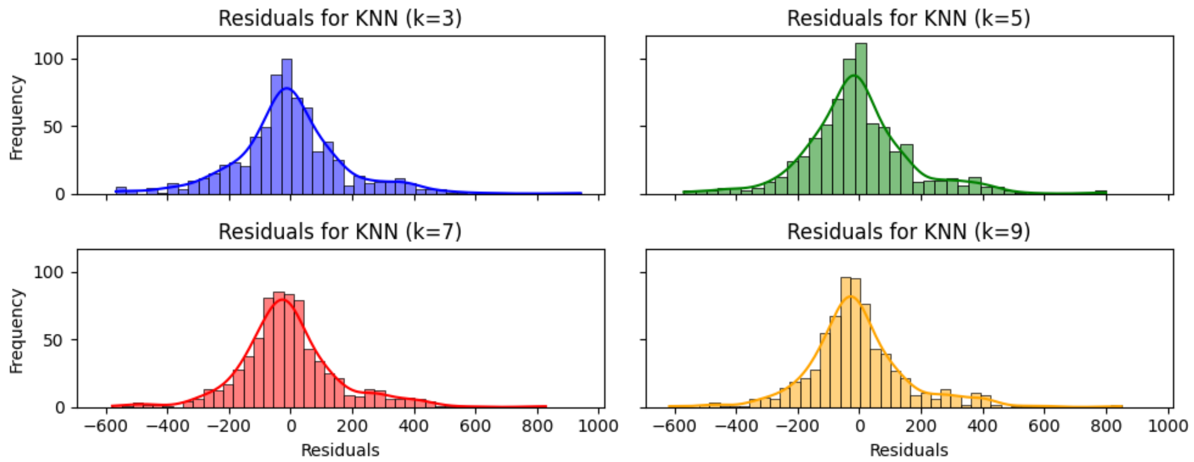


Figure 4.2 Histograms of Residuals for KNN with Different 'k' Values on Capacity Rating.

4.1.3 Linear Regression Imputation

Before training the Linear Regression (LR) model, it is crucial to address multicollinearity among features to ensure unbiased predictions. Variance Inflation Factor (VIF) analysis was performed on the features to identify and mitigate multicollinearity issues. Table 4.3 displays the VIF values for each feature before and after adjustments.

Features	Initial VIF	Updated VIF
Non-Domestic Electricity Consumption	1.83	2.67
Domestic Electricity Consumption	2.46	2.39
Postcode Domestic Electricity Consumption	8.49	2.90
Population	1.58	1.54
Building Number	11.85	Removed
Building Density	2.56	2.08

Table 4.3: Variance Inflation Factor (VIF) for Features before and after adjustments.

Initially, "Building Number" had a VIF of 11.85, indicating severe multicollinearity, and was removed from the analysis (Martin, 2022). " Postcode Domestic Electricity Consumption" also had a high initial VIF of 8.49, which was reduced to 2.90 after adjustments. The updated VIF values reflect a significant reduction in multicollinearity, making the dataset more suitable for linear regression modelling.

RMSE	MAE	MAPE
187.17	134.10	194.44%

Table 4.4: Evaluation Metrics for Linear Regression on Capacity Rating.

Table 4.4 shows the performance metrics for the linear regression imputation on capacity rating. The model yields an RMSE of 187.17, MAE of 134.10, and MAPE of 194.44%. These high error numbers indicate that the linear regression model may not perform well for this dataset.

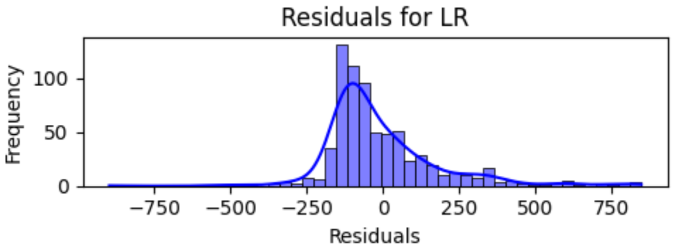


Figure 4.3 Histogram of Residuals for Linear Regression on Capacity Rating.

The histogram of residuals further illustrates that the distribution is a leftward skew from zero. This suggests that the model tends to overestimate the target values, which could explain the high MAPE. Overall, the results suggest that linear regression may not be the most suitable method for this dataset.

4.1.4 Random Forest Imputation

The Random Forest (RF) imputation performance was assessed with various forest sizes and maximum depths for capacity rating. Table 4.5 shows that a forest size of 500 with a max depth of 15 achieves the lowest RMSE (152.32), MAE (109.27), and MAPE (128.99%). This configuration provides the best overall performance and will be used for the test set.

Forest Size	***Max Depth***	**RMSE**	**MAE**	**MAPE**
10	15	157.75	112.85	135.12%
10	30	159.04	114.50	138.26%
100	15	153.39	110.37	130.70%
100	30	154.36	110.69	133.03%
500	15	152.32	109.27	128.99%
500	30	153.52	110.31	131.87%

Table 4.5: Evaluation Metrics for Different Forest Size and Max Depth in Random Forest on Capacity Rating.

Increasing the forest size generally improves performance. With 500 trees, both RMSE, MAE, and MAPE are notably lower compared to configurations with 10 or 100 trees. This suggests that a larger forest size allows for better pattern capture and model accuracy. Conversely, a maximum depth of 15 yields better results compared to a depth of 30. Deeper trees, with a max depth of 30, show minimal improvement or even worse accuracy, which could indicate potential overfitting.

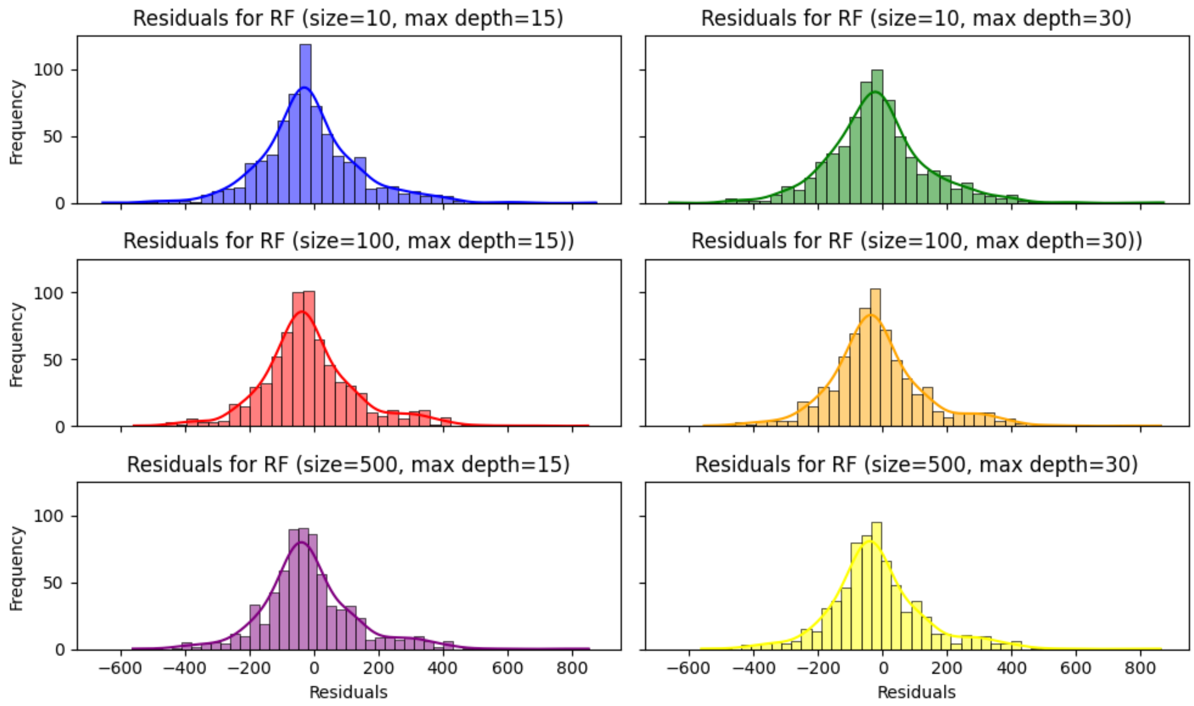


Figure 4.4 Histograms of Residuals for Random Forest with Different Forest Size and Max Depth on Capacity Rating.

Figure 4.4 shows that with a forest size of 10, the residuals have the widest distribution, corresponding to the poorest prediction accuracy. The histograms for larger forest sizes (100 and 500) and different max depths are similar, making it hard to see differences between them based on residuals alone.

4.1.5 Multilayer Perceptrons Imputation

The performance of Multilayer Perceptrons (MLP) imputation was evaluated using various configurations of hidden layer neurones and learning rates. Table 4.6 summarises the evaluation metrics for different setups.

The results show that different hidden sizes do not lead to a clear trend in performance. Configurations with (15, 15) neurones consistently reached higher RMSE and MAE values, indicating poorer performance. The learning rate has a more noticeable impact on results. Both the smallest (0.005) and largest (0.1) learning rates produce less accurate results. In contrast, learning rates of 0.01 and 0.05 deliver better outcomes. This suggests that moderate learning rates are more effective for MLP imputation in this dataset, while the choice of hidden size has less impact on performance.

Overall, the best configuration for RMSE is (30, 30) neurones with a learning rate of 0.01, achieving an RMSE of 153.00. For MAE, the optimal setup is also (30, 30) neurones but with a learning rate of 0.05, resulting in an MAE of 108.22. To compare these configurations, MAPE was used. The setup with (30, 30) neurones with a learning rate of 0.05 was selected as the best performance setup and will be applied to the test set.

Hidden Size	*Learning Rate*	**RMSE**	**MAE**	**MAPE**
(5, 5)	0.005	156.13	110.86	127.01%
(5, 5)	0.01	154.23	110.37	131.36%
(5, 5)	0.05	154.63	112.03	128.70%
(5, 5)	0.1	158.27	110.54	114.27%
(15, 15)	0.005	171.88	120.51	150.46%
(15, 15)	0.01	155.34	115.36	155.77%
(15, 15)	0.05	159.23	111.22	124.46%
(15, 15)	0.1	163.10	113.59	130.25%
(30, 30)	0.005	154.71	112.64	141.19%
(30, 30)	0.01	153.00	111.83	143.08%
(30, 30)	0.05	155.18	108.22	124.66%
(30, 30)	0.1	160.37	110.21	109.30%

Table 4.6: Evaluation Metrics for Different Hidden Neurones and Learning Rates in MLP on Capacity Rating.

4.1.6 Inverse Distance Weighting Imputation

The Inverse Distance Weighting (IDW) imputation performance was assessed with different numbers of neighbours. Table 4.7 shows how changing the number of neighbours affects the imputation accuracy, as measured by RMSE, MAE, and MAPE.

Neighbours	**RMSE**	**MAE**	**MAPE**
3	185.25	136.78	187.16%
5	180.34	134.03	191.08%
7	177.70	133.16	188.72%
9	175.66	132.41	191.03%

Table 4.7: Evaluation Metrics for Different Number of Neighbours in IDW on Capacity Rating.

As the number of neighbours increases, both RMSE and MAE improve. MAPE remains similar across all settings. The improvement with more neighbours suggests that incorporating a broader view of the data enhances imputation accuracy and stability. The best results are achieved with 9 neighbours, providing the lowest RMSE (175.66) and MAE (132.41). However, the overall high evaluation metrics suggest that the capacity rating may not be suitable for prediction based solely on location.

Figure 4.5 shows the hexagonal binning plots of residuals for IDW imputation with varying numbers of neighbours. There appears to be no clear pattern in either extreme values or geometric distribution in the residuals.

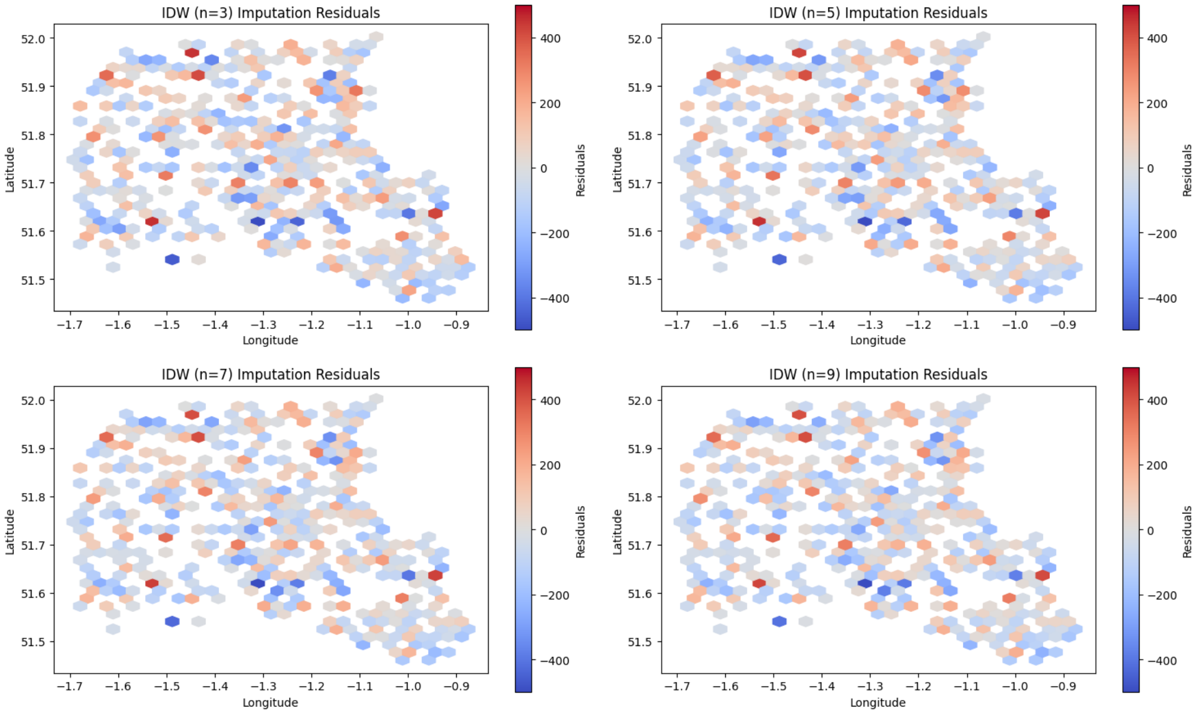

IDW (n=3) Imputation Residuals
IDW (n=5) Imputation Residuals
IDW (n=7) Imputation Residuals
IDW (n=9) Imputation Residuals
Latitude
Longitude
Residuals
52.0
51.9
51.8
51.7
51.6
51.5
−1.7
−1.6
−1.5
−1.4
−1.3
−1.2
−1.1
−1.0
−0.9
400
200
0
−200
−400

Figure 4.5 Hexagonal Binning Plot of Residuals for IDW with Different Number of Neighbours on Capacity Rating.

4.1.7 Geometrically Weighted Regression Imputation

Geometrically Weighted Regression (GWR) imputation combines spatial information and feature relationships to predict missing values. Table 4.8 displays the evaluation metrics for different bandwidths used in GWR for capacity rating. A bandwidth of 200 achieved the best performance, with an RMSE of 164.58 and an MAE of 116.75.

Bandwidth	**RMSE**	**MAE**	**MAPE**
50	187.66	127.49	147.57%
100	168.47	119.21	143.51%
200	164.58	116.75	145.10%
400	165.07	117.63	151.60%

Table 4.8: Evaluation Metrics for Different Bandwidths in GWR for Capacity Rating.

The smallest bandwidth (50) performed significantly worse, showing the highest RMSE and MAE, which indicates poor accuracy. A bandwidth of 200 provided the best balance of performance metrics. However, increasing the bandwidth to 400 slightly reduced accuracy, suggesting that extremely large bandwidths may lead to less precise predictions. This underscores the need to select the right bandwidth to prevent both overfitting and underfitting.

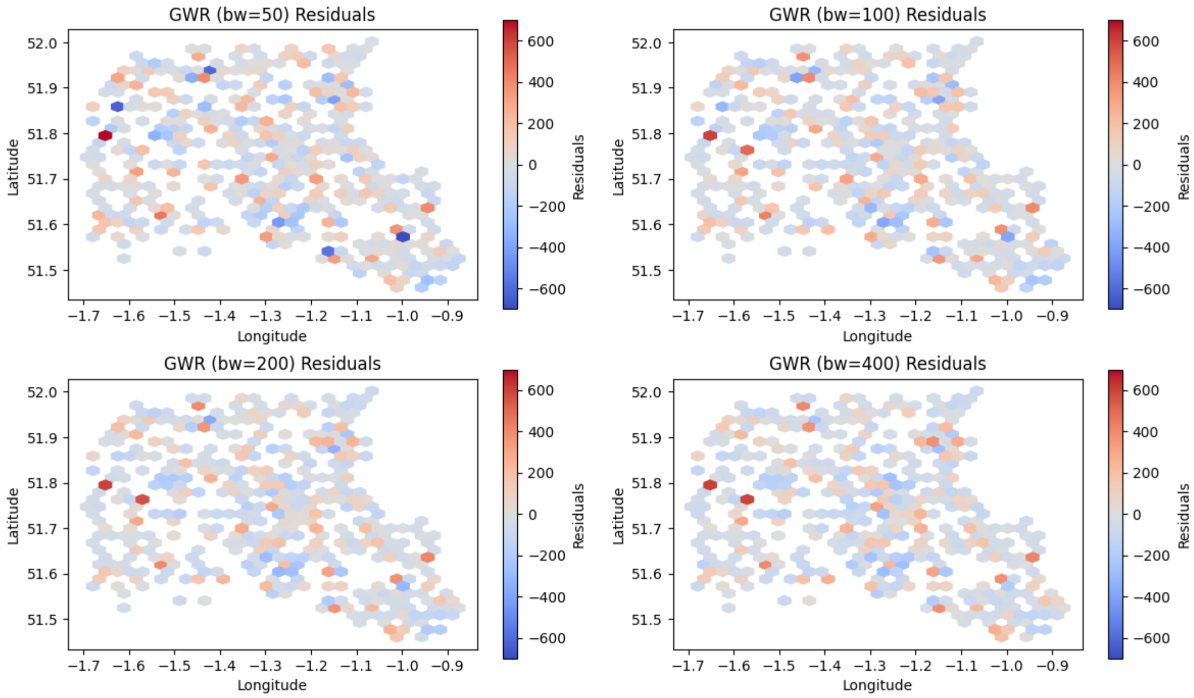


Figure 4.6 Hexagonal Binning Plot of Residuals for GWR with Different Bandwidths on Capacity Rating.

Figure 4.6 shows the hexagonal binning plots of residuals for different bandwidths. With smaller bandwidths, such as 50, there are more extreme negative residuals. As the bandwidth increases, the distribution of residuals becomes more balanced. However, when the bandwidth reaches 400, a few positive extreme residuals start to appear. This indicates that very small bandwidths may lead to significant negative errors. Very large bandwidths, although reducing some extreme values, might introduce a few positive extremes and slightly reduce accuracy.

4.1.8 Methods Comparison on Test Set

The performance of different imputation methods for capacity rating is summarised in Table 4.9. Evaluation metrics were used to assess the accuracy of each method on the test set.

Method	RMSE	MAE	MAPE
Median	266.14	196.31	207.37%
KNN	169.65	117.96	118.99%
LR	196.35	139.71	170.60%
RF	163.99	112.62	118.20%
MLP	163.77	109.75	107.64%
IDW	180.24	134.80	164.69%
GWR	246.16	166.63	201.47%

Table 4.9: Evaluation Metrics for Various Imputation Methods on the Test Set for Capacity Rating.

Among the methods, MLP showed the best overall performance with the lowest RMSE (163.77), MAE (109.75), and MAPE (107.64). Random Forest also performed well, closely following MLP. In contrast, the GWR method had higher values in evaluation metrics even closer to median imputation, indicating poor accuracy performance compared to other methods.

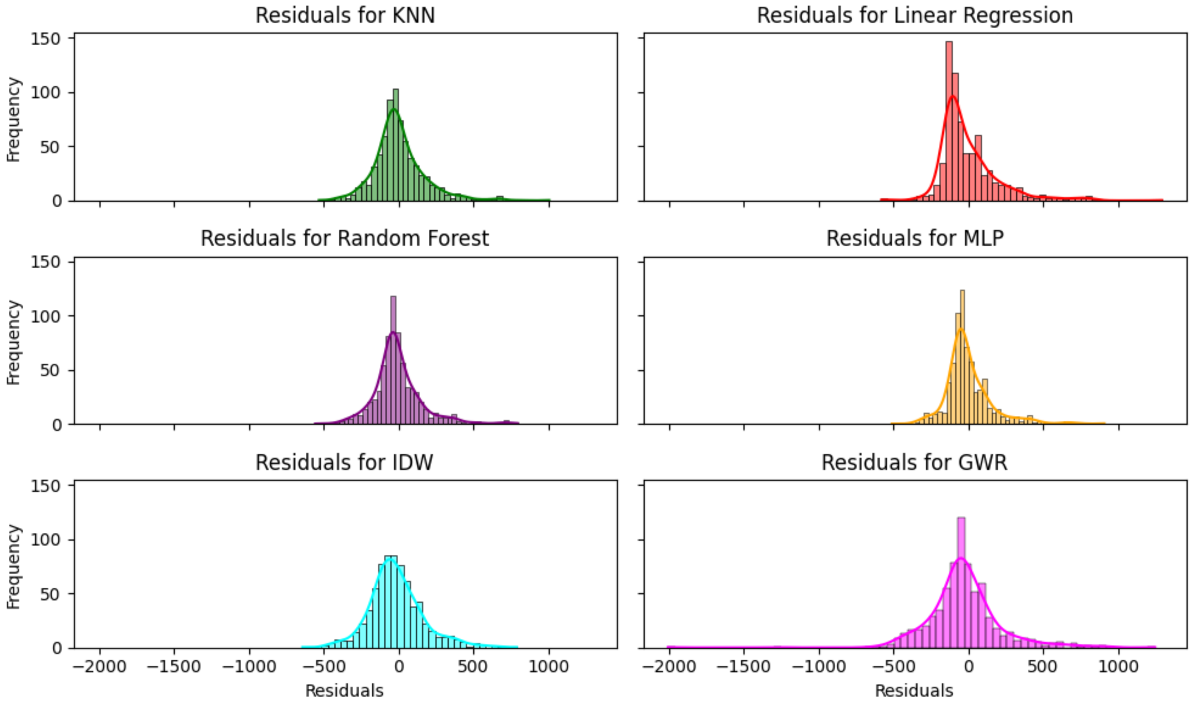


Figure 4.7 Histograms of Residuals with Different Imputation Methods on Capacity Rating.

Figure 4.7 shows the histograms of residuals reveal several patterns. RF and MLP show residuals more concentrated around zero, indicating accurate predictions. KNN also performs well but is less concentrated than RF and MLP. In contrast, IDW exhibits a wider residual distribution, suggesting less precision. The residuals of LR are notably leftward from zero, indicating a tendency to overestimate. Interestingly, the residuals of GWR are similar to KNN near zero but show extreme negative values up to -2000, which might explain the poor performance on evaluation metrics. Both LR and GWR also display extremely positive residuals, highlighting their poor prediction accuracy in certain cases.

4.2 Results for Peak Load Imputation

Similar to the previous section, this part presents the evaluation results for peak load. The same imputation methods applied to the capacity rating are used here to assess and determine the best-performing model on the validation set for peak load. The selected models will then be applied to the test set to decide the best model for peak load imputation.

4.2.1 Mean / Median / Mode Imputation

The mean, median, and mode were calculated from the training set for Peak Load, resulting in values of 95.41, 53.51, and 3.90, respectively. These statistics were then used to predict values for the validation set.

Method	Value	RMSE	MAE	MAPE
Mean	95.41	110.93	87.34	631.93%
Median	53.51	119.03	80.74	348.79%
Mode	3.90	144.61	93.13	82.00%

Table 4.10: Evaluation Metrics for Mean, Median, and Mode Imputation Methods on Peak Load.

Table 4.10 shows the evaluation results for Peak Load imputation. Mean imputation resulted in an RMSE of 110.93, an MAE of 87.34, and a high MAPE of 631.93%. Median imputation performed similarly, with an RMSE of 119.03, an MAE of 80.74, and a lower MAPE of 348.79%. Mode imputation had the highest RMSE (144.61) and MAE (93.13), but the best MAPE (82.00%), which could be due to most Peak Load values being close to zero. Overall, median imputation is selected as the baseline model due to its balanced performance.

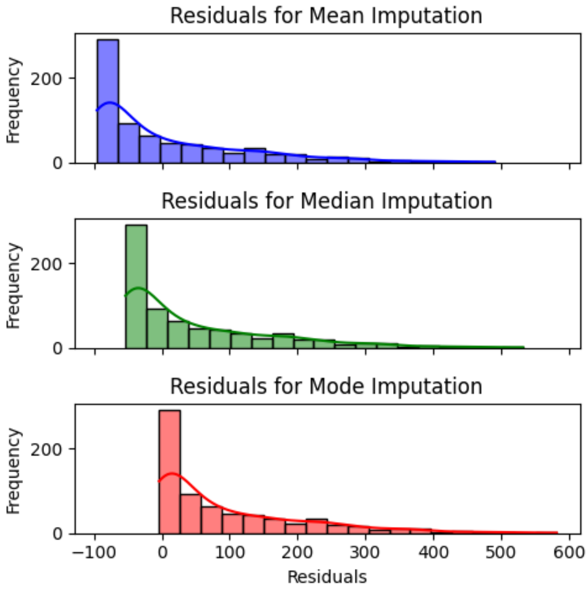


Figure 4.8 Histograms of Residuals for Mean, Median, and Mode Imputation Methods on Peak Load.

In Figure 4.8, the residual plots for Peak Load imputation show similar patterns to those for Capacity Rating. The overall distribution is reflected, with vertical shifts based on the imputation values. While mean and median residuals are centred around zero, mode imputation shows a smaller shift, with most residuals skewing higher.

4.2.2 *k*-Nearest Neighbour Imputation

Table 4.11 shows the performance of *k*-Nearest Neighbour (KNN) imputation for Peak Load with different values of *k*.

k	**RMSE**	**MAE**	**MAPE**
3	71.58	44.83	220.48%
5	67.79	41.93	219.56%
7	67.44	41.89	228.48%
9	67.34	41.54	227.82%

Table 4.11: Evaluation Metrics for Different Values of 'k' in KNN on Peak Load.

As number of neighbours increases, both RMSE and MAE improve. However, MAPE performs better with smaller values of *k*. Based on the results, *k = 9* was chosen due to its lowest RMSE (67.34) and MAE (41.54), despite its high MAPE.

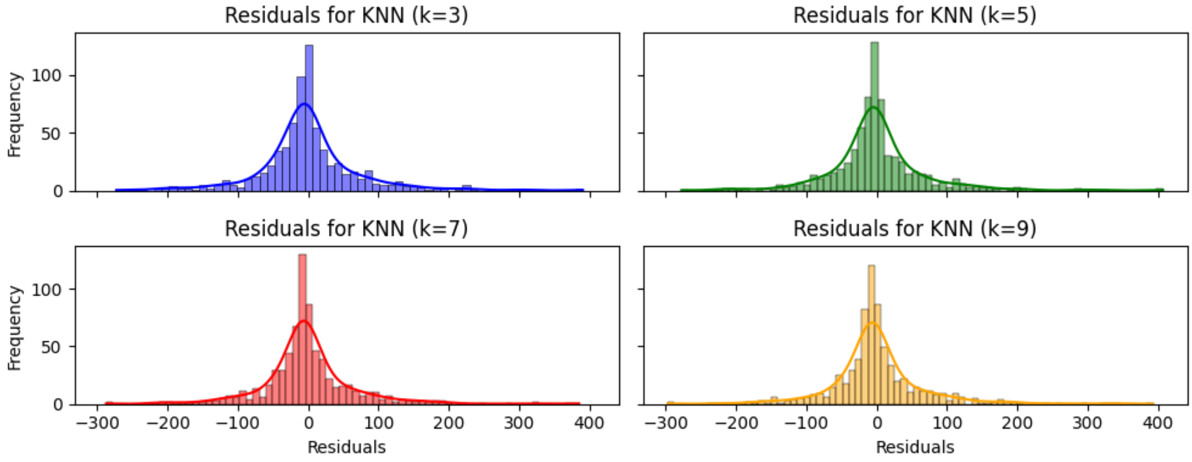


Figure 4.9 Histograms of Residuals for KNN with Different 'k' Values on Peak Load.

Figure 4.9 shows the histograms of residuals for each *k* value. Although *k = 9* offers a more even distribution of residuals, it also has more extreme values compared to other *k* values. This broader spread and presence of outliers may explain why MAPE is higher with *k = 9* despite improvements in other metrics.

4.2.3 Linear Regression Imputation

The Linear Regression (LR) imputation for peak load inherits the VIF results previously shown for Capacity Rating in Table 4.3 The remaining five features, after addressing multicollinearity, were used for this imputation. The evaluation metrics for this approach are detailed in Table 4.12, and the residuals histogram is presented in Figure 4.10.

RMSE	MAE	MAPE
71.76	41.27	304.67%

Table 4.12: Evaluation Metrics for Linear Regression on Peak Load.

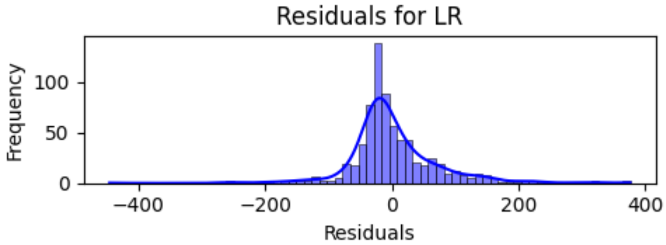


Figure 4.10 Histogram of Residuals for Linear Regression on Peak Load.

The evaluation metrics for linear regression imputation on peak load are RMSE = 71.76, MAE = 41.27, and MAPE = 304.67%. These high values suggest that this method may not be the most accurate for peak load imputation. Although the histogram of residuals does not show a significant skew, there is still a slight leftward skew. This indicates that while the residuals' distribution is relatively balanced, predictions tend to be slightly higher than the actual values. Even after addressing multicollinearity, linear regression did not achieve the lowest error metrics compared to other imputation methods.

4.2.4 Random Forest Imputation

The Random Forest (RF) imputation performance was evaluated with various forest sizes and maximum depths for peak load.

Forest Size	***Max Depth***	**RMSE**	**MAE**	**MAPE**
10	15	72.65	44.65	242.26%
10	30	73.63	45.90	240.28%
100	15	69.39	43.33	224.05%
100	30	69.15	43.35	225.10%
500	15	69.29	43.29	226.21%
500	30	69.33	43.48	226.46%

Table 4.13: Evaluation Metrics for Different Forest Size and Max Depth in Random Forest on Peak Load.

Table 4.13 shows that a forest size of 100 with a max depth of 30 achieves the lowest RMSE (69.15). Meanwhile, a forest size of 500 with a max depth of 15 achieves the lowest MAE (43.29). To determine the best overall model, MAPE was used for comparison between these two configurations. The forest size of 100 with a max depth of 30 performs slightly better in terms of MAPE and is selected for further analysis.

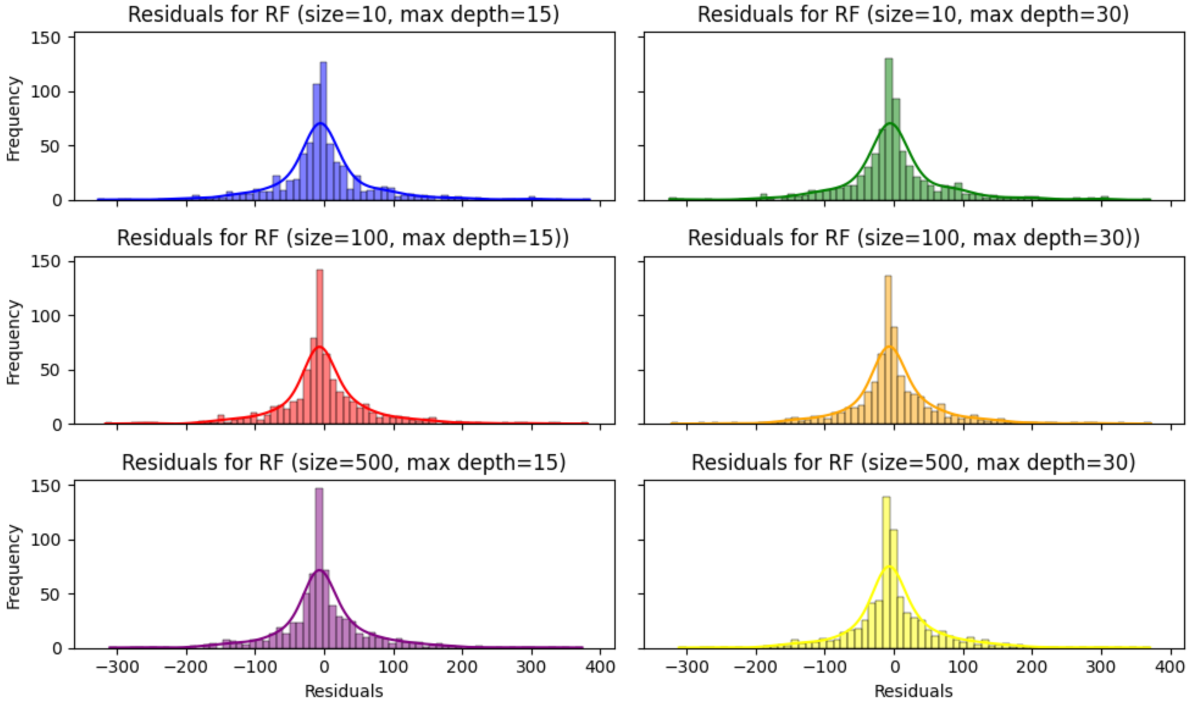


Figure 4.11 Histograms of Residuals for Random Forest with Different Forest Size and Max Depth on Peak Load.

Figure 4.11 displays the histograms of residuals for different forest sizes and max depths. The forest size of 10 shows the lowest performance, with residuals less concentrated around zero, indicating less accurate predictions. In contrast, the forest sizes of 100 and 500, and the max depths of 15 and 30, display similar performance with residuals more concentrated around zero, reflecting better prediction accuracy.

4.2.5 Multilayer Perceptrons Imputation

Table 4.14 displays the performance results for Multilayer Perceptrons (MLP) imputation using different hidden sizes and learning rates on peak load. The configuration with 30 neurones in both layers and a learning rate of 0.005 achieves the lowest RMSE at 66.14. However, for MAE and MAPE, the best performance is observed with (5, 5) neurones and a learning rate of 0.05, reaching values of 40.75 and 182.85%, respectively.

Learning rates do not show a clear trend in performance. Comparing different hidden sizes, (5, 5) neurones achieve the best MAE and MAPE. Meanwhile, (30, 30) neurones provide the lowest RMSE but perform worse in terms of MAPE. The (15, 15) neurones deliver moderate results across all metrics. Overall, (5, 5) neurones with a learning rate of 0.05 is chosen for its balanced performance.

Hidden Size	*Learning Rate*	RMSE	MAE	MAPE
(5, 5)	0.005	68.68	41.48	208.34%
(5, 5)	0.01	68.95	42.12	222.06%
(5, 5)	0.05	67.71	40.75	182.85%
(5, 5)	0.1	68.04	41.99	218.59%
(15, 15)	0.005	67.13	42.26	234.43%
(15, 15)	0.01	67.42	42.80	244.91%
(15, 15)	0.05	68.06	43.26	240.66%
(15, 15)	0.1	68.43	42.32	227.52%
(30, 30)	0.005	66.14	42.13	259.56%
(30, 30)	0.01	66.39	43.67	279.84%
(30, 30)	0.05	66.93	41.59	231.72%
(30, 30)	0.1	67.93	43.48	265.75%

Table 4.14: Evaluation Metrics for Different Hidden Neurones and Learning Rates in MLP on Peak Load.

4.2.6 Inverse Distance Weighting Imputation

The Inverse Distance Weighting (IDW) imputation performance was evaluated with different numbers of neighbours. Table 4.15 shows how changing the number of neighbours affects accuracy. The configuration with 9 neighbours achieves the best performance, with RMSE at 91.22 and MAE at 65.45. MAPE is relatively low at 442.91%.

Neighbours	RMSE	MAE	MAPE
3	99.06	68.08	454.81%
5	93.64	66.15	439.07%
7	92.70	65.46	447.06%
9	91.22	65.45	442.91%

Table 4.15: Evaluation Metrics for Different Number of Neighbours in IDW on Peak Load.

Similar to the capacity rating, the peak load results improve as the number of neighbours increases. However, the performance remains poor across all settings. This suggests that peak load imputation should not rely heavily on spatial information.

Figure 4.12 presents hexagonal binning plots of residuals for IDW with varying numbers of neighbours. Fewer neighbours produce more extreme negative residuals, but the positive residuals remain consistent. There is no clear spatial pattern in the residuals across different settings.

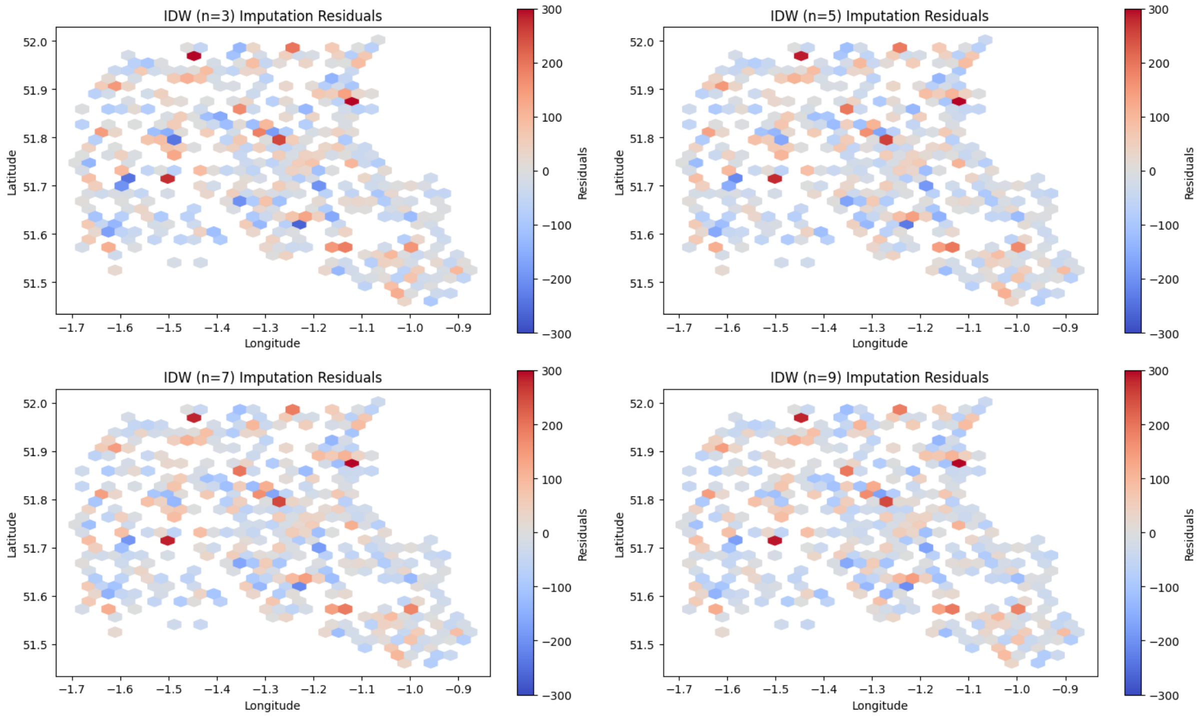


Figure 4.12 Hexagonal Binning Plot of Residuals for IDW with Different Number of Neighbours on Peak Load.

4.2.7 Geometrically Weighted Regression Imputation

Geometrically Weighted Regression (GWR) imputation was evaluated with different bandwidths, and the results are presented in Table 4.16 The optimal performance in terms of RMSE is achieved with a bandwidth of 400, recording an RMSE of 70.14. Meanwhile, the best MAE (43.72) and MAPE (250.89%) are observed with a bandwidth of 200. Given these results, a bandwidth of 200 is selected for its overall superior performance in MAE and MAPE.

Bandwidth	**RMSE**	**MAE**	**MAPE**
50	84.91	50.13	259.00%
100	74.53	45.69	253.49%
200	70.29	43.72	250.89%
400	70.14	43.83	256.84%

Table 4.16: Evaluation Metrics for Different Bandwidths in GWR for Peak Load.

The smallest bandwidth, 50, performed the worst, with the highest metrics values. As the bandwidth increased, the performance improved and stabilised with similar results between 200 and 400. The minimal performance difference between 200 and 400 implies that spatial information may not be as crucial for peak load imputation, where general trends may matter more.

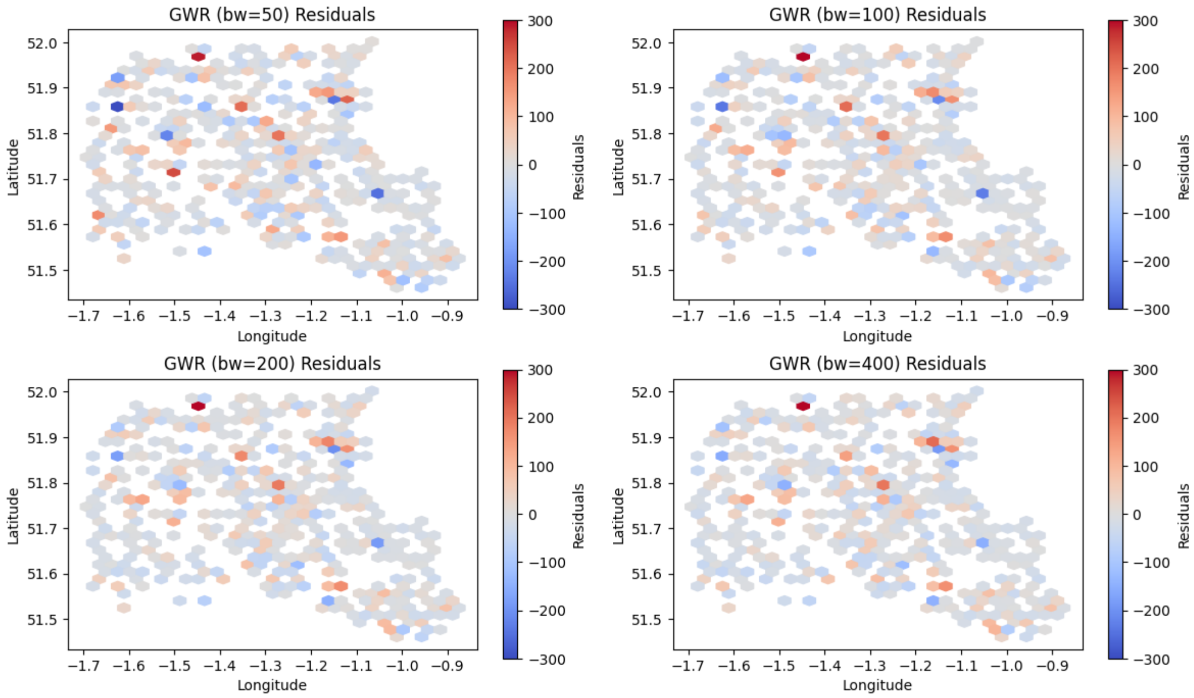


Figure 4.13 Hexagonal Binning Plot of Residuals for GWR with Different Bandwidths on Peak Load.

Like capacity rating, the peak load imputation also shows more extreme negative residuals with smaller bandwidths in GWR. As shown in Figure 4.13, the hexagonal binning plot highlights this pattern clearly. Residuals with bandwidths of 200 and 400 are quite similar, indicating minimal differences between these settings. This suggests that increasing the bandwidth beyond 200 does not significantly impact the residuals, as the model's performance stabilises at these values.

4.2.8 Methods Comparison on Test Set

The performance of different imputation methods for Peak Load, evaluated using the best configurations selected from earlier analyses, is summarised in Table 4.17.

Method	RMSE	MAE	MAPE
Median	111.80	75.75	343.55%
KNN	66.59	41.02	187.38%
LR	69.17	45.89	265.42%
RF	66.90	42.09	195.65%
MLP	65.15	38.82	131.85%
IDW	87.82	63.65	404.08%
GWR	115.72	53.87	270.06%

Table 4.17: Evaluation Metrics for Various Imputation Methods on the Test Set for Peak Load.

Among all the methods, MLP achieved the best overall performance with the lowest RMSE (65.15), MAE (38.82), and MAPE (131.85%). Both KNN and RF also performed well, with KNN slightly ahead of RF. In contrast, the IDW and GWR produced notably higher error metrics, indicating poorer performance. Notably, GWR had an RMSE (115.72) even worse than median imputation (111.80), although its MAE (53.87) was lower than IDW (63.65), suggesting that extreme residuals might have contributed to its elevated error values.

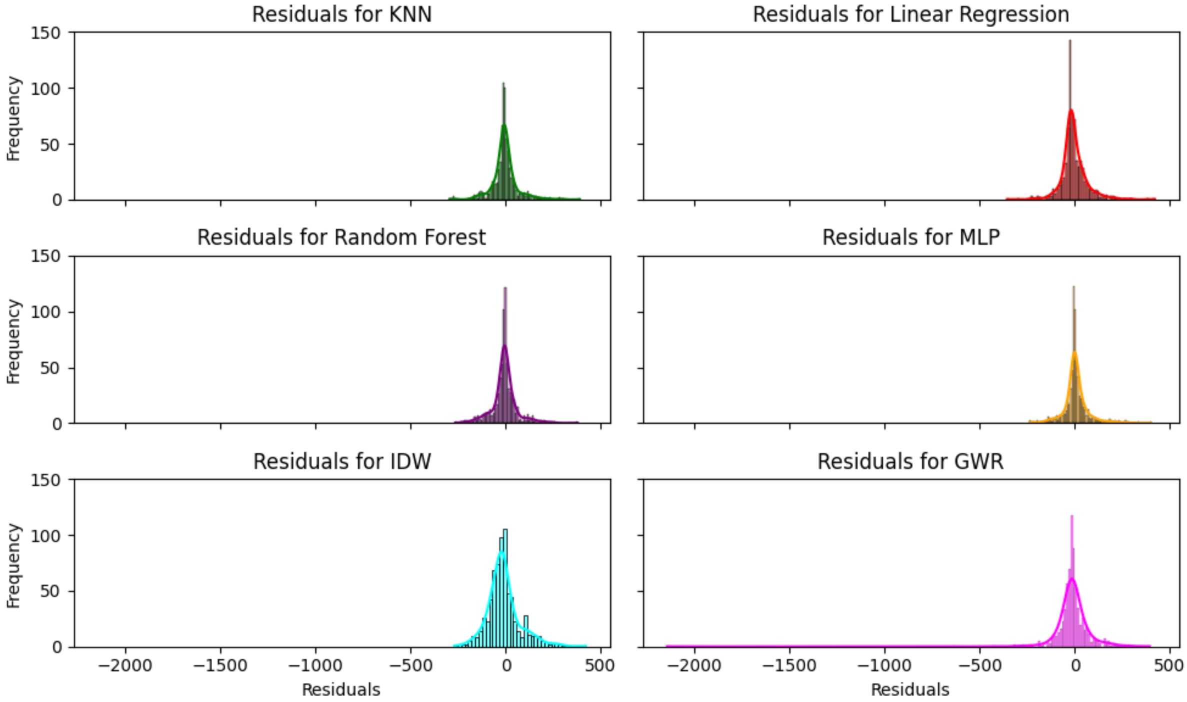


Figure 4.14 Histograms of Residuals with Different Imputation Methods on Peak Load.

Figure 4.14 presents the histograms of residuals for each imputation method. GWR exhibits extreme negative residuals, likely contributing to its highest RMSE. For a clearer comparison between methods, Figure 4.15 focuses on residuals within the range of -500 to 500, allowing for a more precise visual interpretation.

In Figure 4.15, the residuals for MLP exhibit the highest frequency near zero and the narrowest distribution compared to other methods, indicating more accurate predictions. KNN and RF also display similar patterns, though with slightly lower frequencies near zero and a broader spread. Both LR and GWR show a leftward shift from zero, along with a much wider distribution, indicating more significant errors. Meanwhile, IDW shows the widest spread of residuals, reflecting less accurate imputation.

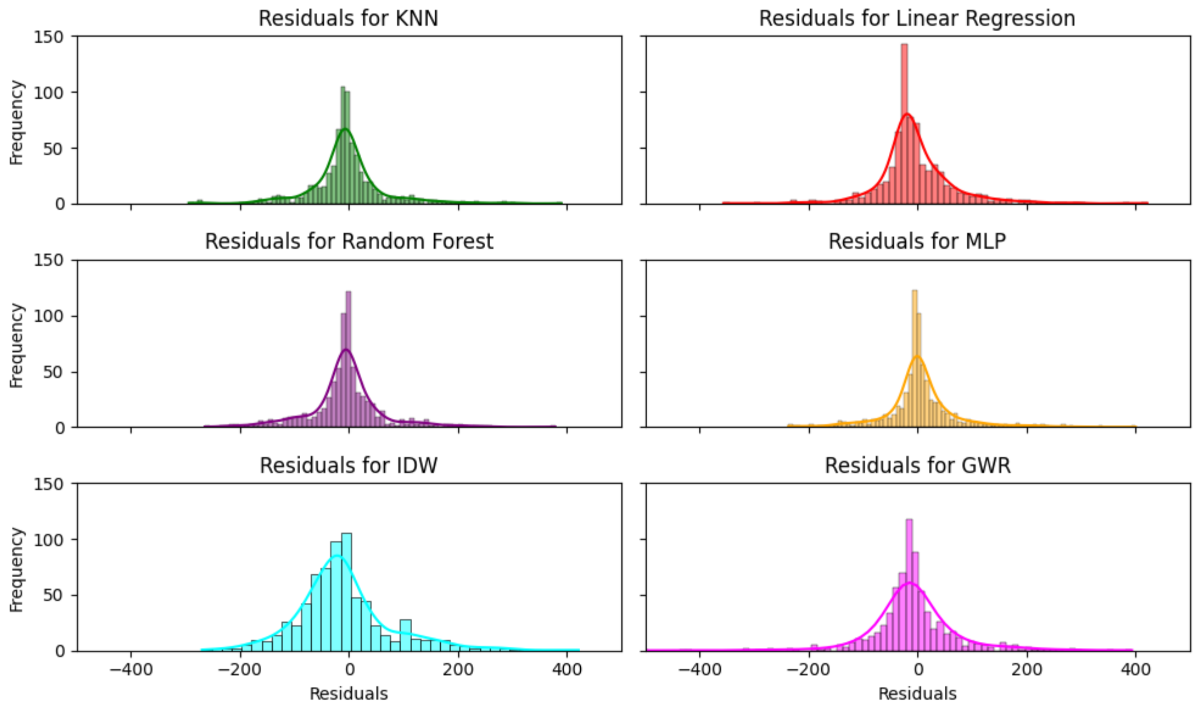


Figure 4.15 Histograms of Residuals with Different Imputation Methods on Peak Load (Focusing on Range -500 to 500).

4.3 Imputation Results on Real Missing Values

In this section, we apply the best models identified from the previous analysis to the real missing values in the dataset. To ensure the best possible imputation, we retrained the models using all available non-missing data to capture the full distribution and relationships within the dataset. Additionally, we compared the imputed data at the Primary Electricity Supply Area level in peak load to further validate the imputed results.

4.3.1 Imputation Performance Comparison

Table 4.18 shows the imputation results for capacity rating and peak load that compared the statistics of non-missing data, imputed data, and all data combined. For capacity rating, the best model was identified as a Multilayer Perceptrons (MLP) with two hidden layers of 30 neurones each, using a learning rate of 0.05. Similarly, for peak load, the best-performing model was also an MLP, but with two hidden layers of 5 neurones each, and the same learning rate of 0.05.

For capacity rating, the imputed mean (266.58) is slightly lower than the non-missing mean (271.25). However, the combined mean (270.25) remains close to the non-missing data. The imputed data also shows a lower standard deviation (212.44) compared to the non-missing data (254.53). This indicates that the imputed values vary less. Additionally, the imputed data lacks smaller values, with a minimum of 86.53 compared to the minimum of 1 in the non-missing set. The overall distribution of imputed values is more centred. This is shown by the lower standard deviation, and as seen in Figure 4.16, suggesting that the imputation model smoothed out extremes and predicted more conservatively.

For peak load, the imputed mean (70.69) is notably lower than the non-missing mean (94.89). The combined mean (89.71) reflects this shift. The standard deviation for the imputed data (75.05) is smaller compared to the non-missing data (108.39), indicating less variability. Figure 4.16 also shows this difference. The interquartile range (IQR) for the imputed values, with the 25th percentile at 16.20 and the 75th percentile at 114.42, is narrower compared to the non-missing data. This suggests that the imputed values are generally lower and less spread out, compared to the wider range of the non-missing data.

	Capacity Rating			**Peak Load**		
	Non-missing	**Imputed**	**All**	**Non-missing**	**Imputed**	**All**
Mean	271.25	266.58	270.25	94.89	70.69	89.71
Std	254.53	212.44	246.11	108.39	75.05	102.65
Min	1	86.53	1	0	8.44	0
25%	50	105.29	94.68	10.88	16.20	12.88
50%	200	150.69	200	51.72	32.03	46.99
75%	500	406.42	500	149.39	114.42	140.23
Max	1500	1045.78	1500	827.45	665.65	827.45

Table 4.18: Summary Statistics for Capacity Rating and Peak Load: Non-Missing, Imputed, and All Data.

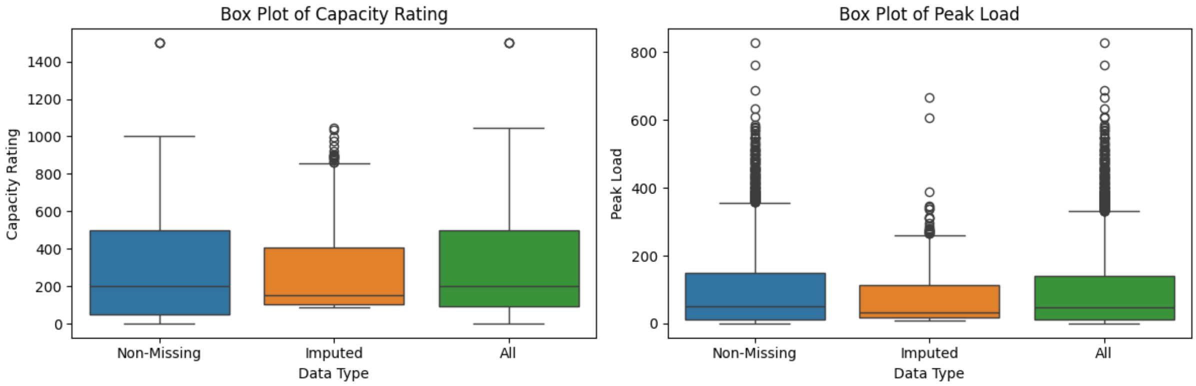


Figure 4.16 Box Plots for Capacity Rating and Peak Load: Comparison of Non-Missing, Imputed, and All Data.

To further analyse the imputation results, the headroom percentage, defined as the difference between capacity rating and peak load expressed as a percentage of capacity, was compared across non-missing, imputed, and combined datasets. The headroom percentage indicates how much capacity exceeds the peak load, reflecting the available margin relative to the system design.

Headroom %	**Non-missing**	**Imputed**	**All**
Mean	58.53	76.84	62.45
Std	113.23	11.88	100.80
Min	-5818.08	27.62	-5818.08
25%	47.01	67.32	53.50
50%	68.88	81.02	72.47
75%	65.66	86.41	86.03
Max	100.00	93.49	100.00

Table 4.19: Summary Statistics for Headroom Percentage: Non-Missing, Imputed, and Combined Data

Table 4.19 presents the summary statistics for headroom percentage across datasets. The minimum headroom percentage in the imputed data is 27.62%, indicating that the capacity rating is consistently higher than the peak load across all imputed data. This trend is also observed in the non-missing data, though it does not hold in a few extreme cases.

The mean headroom percentage is higher in the imputed data (76.84%) compared to the non-missing data (58.53%), and the standard deviation is significantly lower in the imputed data (11.88%) than in the non-missing data (113.23%). This suggests that the imputed data provides a more stable and higher margin between capacity and peak load. These results further support that peak loads are more likely to be underestimated in the imputed data.

4.3.2 Comparison of Peak Load with Primary Electricity Supply Area Data

In addition to further validate the imputed results, we incorporated the primary electricity supply area (PESA) data. For this, we summed the peak load values from all secondary electricity supply areas (ESA) corresponding to each PESA and compared the aggregated values with the original peak load data available for the primary areas.

Figure 4.17 provides a scatter plot that illustrates the relationship between the aggregated secondary area peak load values and the primary area peak loads. While there is a general correlation between the two, the scatter plot shows several cases where the predicted secondary totals fall short of the primary data. This aligns with the earlier finding that the imputed peak load values tend to be more conservative, with lower means and a narrower spread than the non-missing values.

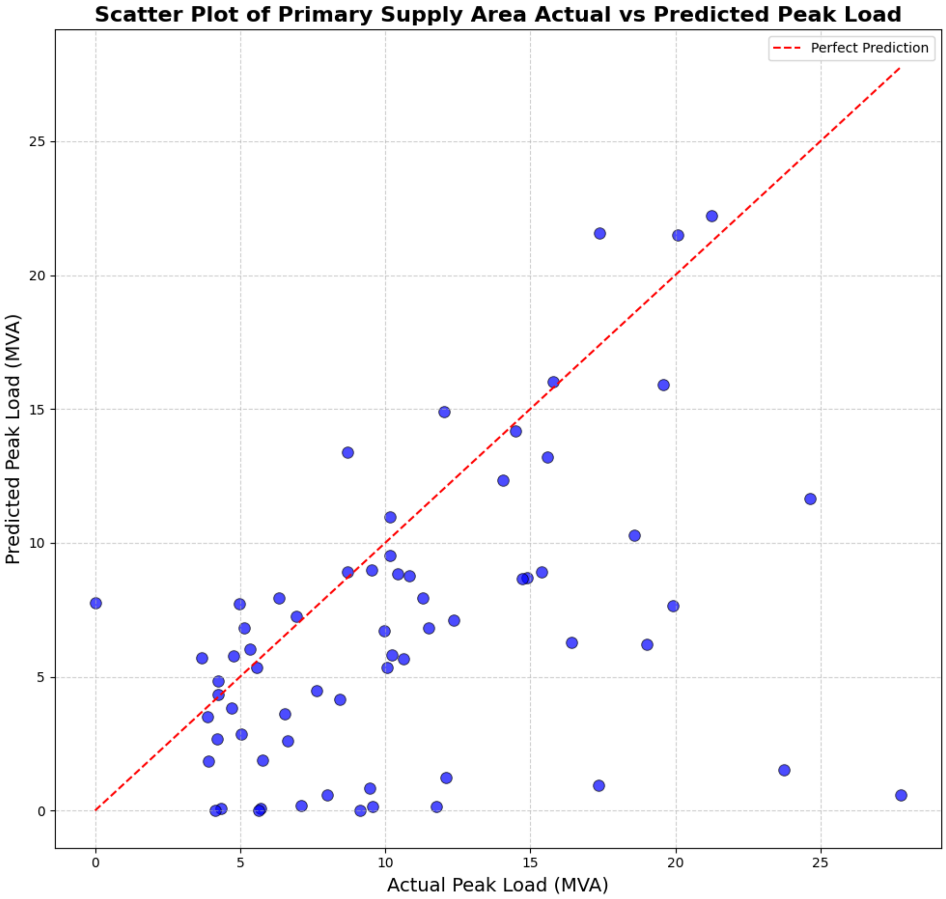


Figure 4.17 Scatter Plot of Primary Supply Area Original vs Predicted Peak Load.

4.4 Conclusion

This chapter presented the results of various imputation methods for capacity rating and peak load. The MLP models, with specific configurations, showed promising results. For capacity rating, the imputed values were more centred and less variable compared to non-missing data. In contrast, peak load imputation led to generally lower and less dispersed values. These findings will be discussed further in the next chapter and will delve into the implications of these results and explore potential improvements.

Chapter 5 Discussion

This chapter evaluates the results presented in the previous chapter. We assess the effectiveness of various imputation methods, compare the best-performing model with the baseline model, analyse the imputed results, and discuss factors influencing prediction accuracy.

5.1 Comparison of Imputation Methods

The seven imputation methods used in this study are Median, k-Nearest Neighbours (KNN), Linear Regression (LR), Random Forest (RF), Multi-Layer Perceptron (MLP), Inverse Distance Weighting (IDW), and Geographically Weighted Regression (GWR). As shown in Table 4.9 and Table 4.17 these methods displayed varying performance across the dataset, revealing their strengths and weaknesses in handling missing data.

Median imputation, the simplest method, provided the least accurate results as expected. Because of its simplicity, it was unable to capture the complex patterns present in the dataset. KNN, on the other hand, showed significant improvement over Median imputation. This improvement suggests that the dataset contains some underlying similarities between data points. Although LR showed an improvement over Median imputation, it was still outperformed by KNN. LR assumes a linear relationship between features, but in this dataset, the relationships might be more complex than linear. This indicates that the data likely do not follow a simple linear pattern but show some level of similarities.

RF and MLP achieved the best results among the methods tested. Both methods are capable of capturing complex and non-linear relationships in the data. This suggests that identifying these complex patterns is crucial for accurate imputation of missing values in capacity rating and peak load. However, both methods require more sophisticated tuning and higher computational resources.

IDW and GWR did not perform well, and their results were close to those of Median imputation. This indicates that spatial information alone, or even when combined with other features, did not significantly improve imputation accuracy. Neither IDW nor GWR added meaningful value in this context. Furthermore, increasing the number of neighbours in IDW and the bandwidth in GWR resulted in slight improvements, suggesting that these methods struggled to capture meaningful local patterns. This emphasises that feature quality is more critical than spatial information for achieving better imputation results in this dataset.

When comparing LR and GWR, both methods struggled to identify useful linear relationships, either globally or locally. In the validation set, GWR showed a slight improvement over LR, but this was not consistent across different splits of the dataset. In the test set, GWR performed much worse than LR, which could be due to random variations in the dataset splits leading to unstable results.

In summary, MLP emerged as the most effective method for imputing missing values, due to its ability to capture complex non-linear relationships. RF also performed well and can be a strong alternative. KNN and LR showed intermediate results. The simplest method, Median imputation, and advanced spatial methods like IDW and GWR did not significantly enhance accuracy. The findings suggest that selecting the right imputation method depends on the data complexity and the nature of the relationships between features.

5.2 Performance of Multilayer Perceptrons Compared to Median Imputation

Both capacity rating and peak load analyses identified the Multilayer Perceptrons (MLP) as the most effective method for imputing missing values. The MLP used (30, 30) neurons with a learning rate of 0.05 for capacity rating, and (5, 5) neurons with the same learning rate for peak load.

Metric	Median (Baseline)	MLP (Best Model)	Improvement
Capacity Rating			
RMSE	266.14	163.77	38.46%
MAE	196.31	109.75	44.09%
MAPE	207.37%	107.64%	48.09%
Peak Load			
RMSE	110.80	65.15	41.20%
MAE	75.75	38.82	48.75%
MAPE	343.55%	131.85%	61.62%

Table 5.1: Performance Comparison Between Median and MLP Models for Capacity Rating and Peak Load.

The results in Table 5.1 highlight the significant improvements of the MLP model over the baseline Median imputation method. For capacity rating, the MLP model reduced RMSE from 266.14 to 163.77, an improvement of 38.46%. MAE and MAPE also decreased by 44.09% and 48.09%, respectively. These improvements indicate that the MLP model better captures the underlying data patterns, leading to more accurate predictions.

For peak load, the MLP model showed even more notable improvements. RMSE decreased from 110.80 to 65.15, representing a 41.20% improvement. MAE and MAPE saw reductions of 48.75% and 61.62%, respectively. The substantial reduction in MAPE demonstrates that the MLP model significantly reduced the relative error in predicting peak load values.

Despite these improvements, both capacity rating and peak load still have error rates that exceed acceptable thresholds for electricity applications. Ideally, the RMSE and MAE should be under 30% of the mean values, which is around 80 for capacity rating and 30 for peak load. This 30% threshold is a suggestion from Advanced Infrastructure Technology Ltd (AITL), the company providing the data, though it has not been independently justified. The results show that even the best-performing MLP models have not yet achieved these levels of accuracy.

5.3 Imputed Results in Capacity Rating and Peak Load

The statistics summary of the imputed results for capacity rating and peak load is shown in Table 4.18, with box plots in Figure 4.16. For capacity rating, the imputation model smooths out extreme values, leading to more conservative predictions. This results in less variability in the imputed values but may also cause an underestimation of true values. For peak load, the MLP model produces a narrower and lower distribution. While this approach reduces variability, it also makes predictions more conservative, with imputed peak load values tending to underestimate the actual values.

Additional analysis with the summary statistics for headroom percentage is provided in Table 4.19. The minimum headroom percentage is greater than 0, indicating that capacity ratings are consistently higher than peak loads. This is expected for a well-functioning system. However, the higher mean headroom percentage compared to non-missing data suggests that peak loads are likely underestimated. Thus, the imputed capacity ratings may better reflect true capacity than the peak load estimates.

Figure 4.17 compares imputed peak load values with primary electricity supply area (PESA) data. The scatter plot shows several cases where predicted totals from secondary areas are lower than the primary data. This reinforces the earlier finding that the imputed peak load values are conservative and tend to underestimate actual peak loads.

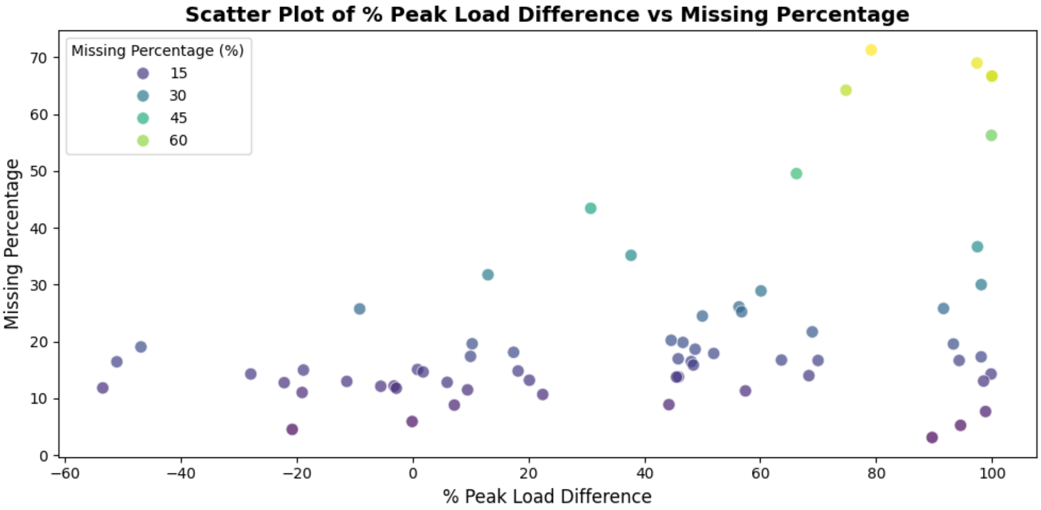


Figure 5.1 Scatter Plot of Peak Load Difference vs. Missing Value Percentage in PESA.

$$\% \, Peak \, Load \, Difference = {(Actual \, Peak \, Load - Non\text{–}missing \, Peak \, Load)}/{Actual \, Peak \, Load}$$

To further analyse how missing values influence the actual data, Figure 5.1 illustrates the relationship between the missing percentage in each PESA and the percentage difference in peak load after imputing the missing values.

The figure indicates that higher missing percentages lead to larger differences in peak load values, which aligns with the expectation that more missing data requires more imputation. However, for areas with lower missing percentages, there is greater variability in these differences. This suggests that while high missing percentages are associated with larger discrepancies between non-missing and actual peak load values, the variability in low missing percentages requires further analysis to understand the underlying factors affecting these differences.

5.4 Factors Impacting Prediction Accuracy

One possible reason for prediction inaccuracies is the lack of diverse and relevant features. In this study, we used features such as non-domestic and domestic electricity consumption, postcode consumption, population, building numbers, and density. However, since many of these features are related to building numbers, they might not fully capture the complexity of the data. Although we tested using building area instead of building numbers for aggregation, building numbers still yielded better results. This indicates that we may be missing other independent features that could improve prediction accuracy. Exploring additional or alternative features, not directly related to building data, might enhance the model's performance.

The postcode data, which has smaller boundaries, showed a higher correlation with the target variables. This suggests that incorporating data with finer boundaries could improve predictions. Using features with smaller, more specific boundaries may provide more detailed and relevant information, potentially enhancing the model's accuracy. Additionally, addressing the unique nature of the two target variables requires identifying features specifically related to each target. For example, capacity rating might improve with features showing overall trends, while peak load could benefit from variables tracking peak usage patterns.

Another factor to consider is the method used to simulate missing values. In this study, random splits were applied to evaluate the model’s prediction, but the actual missing values in the dataset are not spatially random. To better represent the spatial patterns in the data, a more refined approach might involve simulating missing values in a non-missing dataset by creating spatial holes. These holes can then be imputed to assess how well the model performs under more realistic conditions. For this example, five points were randomly selected to create holes with a radius of 0.04, resulting in a total of 602 missing values.

Metric	Test Set	Simulated Holes	Test Set	Simulated Holes
	Capacity Rating		Peak Load	
RMSE	163.77	170.43	65.15	69.36
MAE	109.75	116.52	38.82	46.30
MAPE	107.64%	103.10%	131.85%	140.78%

Table 5.2: Performance Metrics for Predictions on Test Set vs. Simulated Missing Data (Spatial Holes)

Table 5.2 presents the example results from using an MLP model with the same settings as in this study to predict target variables on both the test set and the simulated missing data created by spatial holes. The performance metrics indicate a slight decrease in prediction accuracy for the simulated holes. This emphasizes the challenge of handling spatially clustered missing data. However, these differences may also be influenced by random effects. To further assess stability, additional analyses with more random hole choices are needed.

5.5 Conclusion

In this chapter, we evaluated the performance of seven imputation methods, examining their advantages and limitations. The MLP showed significant improvements over the Median imputation. It notably reduced errors for both capacity rating and peak load. Despite these improvements, MLP did not reach ideal accuracy levels.

Both MLP and RF effectively capture complex data patterns. However, the imputed results still tended to underestimate the actual values. This issue may be due to insufficiently correlated features and challenges in simulating realistic missing data patterns. These problems highlight the need for further investigation and refinement in the imputation methods.

Working with AITL throughout this research allowed for ongoing feedback and adjustments to better meet the specific needs in the industry. Although the results from this project may not provide immediate value for the company they lay the groundwork for future improvements. By refining feature selection and improving imputation accuracy, the techniques explored in this study could lead to more reliable predictions.

In the next chapter, we will evaluate the overall project, reflect on the process, and draw conclusions based on the insights gained.

Chapter 6 Evaluation, Reflections, and Conclusions

This chapter reviews the entire project. It provides a detailed assessment of the work completed and its contributions. It reflects on the methods used, the challenges faced, and the insights gained. This chapter highlights the strengths of the project and identifies areas for improvement. It also suggests future research directions. Finally, it summarises the key conclusion and lessons learned throughout the project.

6.1 Summary of General Conclusions

This project aimed to assess the effectiveness of several imputation methods for handling missing data in the secondary electricity supply area dataset. The results demonstrated clear differences in performance across the methods. The Multilayer Perceptrons (MLP) imputation method proved to be the most effective in both target variables. It significantly reduced error rates compared to the median imputation. Random Forest (RF) also performed well, making it a strong alternative method. However, both methods require more computational resources. On the other hand, methods like Inverse Distance Weighting (IDW) and Geometrically Weighted Regression (GWR), which depend on spatial information, were less effective. This suggests that spatial data alone may not adequately capture the complexity of the dataset.

A key finding was that advanced machine learning methods, such as MLP and RF, performed well in imputing missing values. Their ability to capture non-linear relationships in the data contributed to this success. However, despite these improvements, none of the methods achieved perfect accuracy, particularly when dealing with a high percentage of missing data. Both capacity rating and peak load showed signs of underestimation. This could be due to the lack of highly correlated features or difficulties in simulating realistic missing patterns. These results suggest that advanced imputation techniques improve outcomes, but further refinement in feature selection and data handling is needed for greater accuracy.

In summary, the project demonstrated the significance of selecting the right method for imputation, especially when working with complex datasets. The findings highlight that advanced machine learning methods are useful tools for improving data completeness. However, the results also suggest that there is still room for further optimisation in future studies.

6.2 Future Work

Several opportunities for future work have emerged from the findings of this project. One critical area is improving feature selection. The current feature set, which includes electricity consumption, building density, and population, may not fully capture the complex relationships required for accurate imputation. This is partly because most features were aggregated from building numbers. Future studies should explore adding features that are less correlated with building data and come from smaller geometric regions to obtain more specific information. These changes could improve prediction accuracy.

Another important area is refining the simulation of missing values. In this project, random splits were used to create missing data points. However, real-world missing data often follows spatial or temporal patterns and is not random. Using more advanced approaches, such as creating spatial holes that reflect patterns similar to those in this study, could lead to more realistic test conditions. This approach would provide better insights into model performance by more closely simulating the complexities of real-world missing data.

Additionally, the performance of methods like IDW and GWR suggests that spatial data alone may not be sufficient. Both methods improved when using a larger number of neighbours and a wider bandwidth. This implies there may not be a strong local spatial pattern. The improvements with larger parameters suggest that predictions rely more on distant data points, making local spatial information less relevant for accuracy.

Although GWR performed poorly, this doesn't mean that combining spatial and feature data isn't valuable. It may indicate challenges in detecting local linear relationships. Methods like Geometrically Weighted Artificial Neural Networks (GWANN) (Hagenauer and Helbich, 2022), which blend spatial data with neural networks, could better capture non-linear or complex local patterns. However, GWANN is time-consuming and requires further research to optimise performance.

6.3 Reflection on the Project

Several lessons have been learned throughout the process. Initially, I faced challenges in identifying suitable features due to the diverse geometric boundaries in the datasets. My limited knowledge of the electricity industry also made this process more complex. This difficulty in creating relevant features highlighted the need for collaboration with experts. Gaining deeper insights into how different features interact within the electricity industry might have improved imputation accuracy earlier in the project.

During the literature review, the vast number of imputation methods available was overwhelming. This project revealed that with datasets like ours, which often contain non-linear and complex relationships, simpler methods might not be sufficient. The results suggest that starting with more advanced methods could be more effective. These methods are better equipped to capture complex patterns and may offer improved performance for similar datasets.

Lastly, the approaches to simulating missing values proved more complex than expected. While random splits provided a baseline, they did not fully reflect the challenges of real-world situations. Using approaches that reflect actual data patterns, such as creating spatial holes, might be important. Investing time in understanding and applying these approaches would likely lead to more robust testing of the models.

If I were to start this project again, I would first seek to gain a deeper understanding of the electricity industry to improve feature selection. Also, I would prioritise using advanced imputation methods from the beginning, as they are better suited for datasets with complex relationships. Additionally, I would invest more time in refining the simulation of missing values to better reflect real-world conditions. These changes would likely enhance the robustness of the models and improve overall project outcomes.

Although the results from this project may not yet provide immediate value for the company, they lay the groundwork for future improvements. With further refinement in feature selection and imputation accuracy, the methods explored here could eventually offer useful insights for more reliable predictions. This would make the results more practical and beneficial for real-world applications in the electricity supply industry.

6.4 Final Thoughts

In conclusion, this project has provided valuable insights into the imputation of missing data in complex datasets. It has demonstrated the effectiveness of advanced machine learning methods like MLP and RF in improving data completeness. However, it also highlighted areas for further optimisation, such as feature selection and simulation of missing values. By addressing these areas and refining the approaches used, future work can build on these findings to achieve even greater accuracy and robustness in data imputation.

References

Abidin, N.Z., Ritahani, A., A., N., 2018. Performance Analysis of Machine Learning Algorithms for Missing Value Imputation. International Journal of Advanced Computer Science and Applications 9. https://doi.org/10.14569/IJACSA.2018.090660

Baker, J., White, N., Mengersen, K., 2014. Missing in space: an evaluation of imputation methods for missing data in spatial analysis of risk factors for type II diabetes. Int J Health Geogr 13, 47. https://doi.org/10.1186/1476-072X-13-47

Barudžija, U., Ivšinović, J., Malvić, T., 2024. Selection of the Value of the Power Distance Exponent for Mapping with the Inverse Distance Weighting Method—Application in Subsurface Porosity Mapping, Northern Croatia Neogene. Geosciences (Basel) 14, 155. https://doi.org/10.3390/geosciences14060155

Bastos, L.S.L., Wortel, S.A., Bakhshi-Raiez, F., Abu-Hanna, A., Dongelmans, D.A., Salluh, J.I.F., Zampieri, F.G., Burghi, G., Hamacher, S., Bozza, F.A., de Keizer, N.F., Soares, M., 2024. Comparing causal random forest and linear regression to estimate the independent association of organisational factors with ICU efficiency. Int J Med Inform 191, 105568. https://doi.org/10.1016/j.ijmedinf.2024.105568

Beretta, L., Santaniello, A., 2016. Nearest neighbor imputation algorithms: a critical evaluation. BMC Med Inform Decis Mak 16, 74. https://doi.org/10.1186/s12911-016-0318-z

Bondarenko, I., Raghunathan, T., 2016. Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. Stat Med 35, 3007–3020. https://doi.org/10.1002/sim.6926

Breiman, L., 2001. Random Forests. Mach Learn 45, 5–32. https://doi.org/10.1023/A:1010933404324

Catani, F., Lagomarsino, D., Segoni, S., Tofani, V., 2013. Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Natural Hazards and Earth System Sciences 13, 2815–2831. https://doi.org/10.5194/nhess-13-2815-2013

Chen, G., Zhao, K., McDermid, G.J., Hay, G.J., 2012. The influence of sampling density on geographically weighted regression: a case study using forest canopy height and optical data. Int J Remote Sens 33, 2909–2924. https://doi.org/10.1080/01431161.2011.624130

Chicco, D., Warrens, M.J., Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput Sci 7, e623. https://doi.org/10.7717/peerj-cs.623

Cullen, D.W., Guida, V., 2021. Use of geographically weighted regression to investigate spatial non-stationary environmental effects on the distributions of black sea bass (Centropristis striata) and scup (Stenotomus chrysops) in the Mid-Atlantic Bight, USA. Fish Res 234, 105795. https://doi.org/10.1016/j.fishres.2020.105795 de Myttenaere, A., Golden, B., Le Grand, B., Rossi, F., 2016. Mean Absolute Percentage Error for regression models. Neurocomputing 192, 38–48. https://doi.org/10.1016/j.neucom.2015.12.114

Department for Energy Security and Net Zero, 2024. Lower and Middle Super Output Areas electricity consumption: Lower Super Output Area (LSOA), Middle Super Output Area (MSOA) and Intermediate Geography Zone (IGZ) electricity data.

Department for Energy Security and Net Zero, O.& D. for B.E.& I.S., 2022. Electricity Networks Strategic Framework: Enabling a secure, net zero energy system.

Dou, J., Yunus, A.P., Tien Bui, D., Merghadi, A., Sahana, M., Zhu, Z., Chen, C.-W., Khosravi, K., Yang, Y., Pham, B.T., 2019. Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. Science of The Total Environment 662, 332–346. https://doi.org/10.1016/j.scitotenv.2019.01.221

Emamgholizadeh, S., Shahsavani, S., Eslami, M.A., 2017. Comparison of artificial neural networks, geographically weighted regression and Cokriging methods for predicting the spatial distribution of soil macronutrients (N, P, and K). Chin Geogr Sci 27, 747–759. https://doi.org/10.1007/s11769-017-0906-6

Guo, Z., Xu, L., Ali Asgharzadeholiaee, N., 2022. A Homogeneous Ensemble Classifier for Breast Cancer Detection Using Parameters Tuning of MLP Neural Network. Applied Artificial Intelligence 36. https://doi.org/10.1080/08839514.2022.2031820

Hagenauer, J., Helbich, M., 2022. A geographically weighted artificial neural network. International Journal of Geographical Information Science 36, 215–235. https://doi.org/10.1080/13658816.2021.1871618

Hron, K., Templ, M., Filzmoser, P., 2010. Imputation of missing values for compositional data using classical and robust methods. Comput Stat Data Anal 54, 3095–3107. https://doi.org/10.1016/j.csda.2009.11.023

Jadhav, A., Pramod, D., Ramanathan, K., 2019. Comparison of Performance of Data Imputation Methods for Numeric Dataset. Applied Artificial Intelligence 33, 913–933. https://doi.org/10.1080/08839514.2019.1637138

Kaiser, J., 2014. Dealing with Missing Values in Data. Journal of Systems Integration 5, 42–51.

Khosravi, Y., Balyani, S., 2019. Spatial Modeling of Mean Annual Temperature in Iran: Comparing Cokriging and Geographically Weighted Regression. Environmental Modeling & Assessment 24, 341–354. https://doi.org/10.1007/s10666-018-9623-5

Kokla, M., Virtanen, J., Kolehmainen, M., Paananen, J., Hanhineva, K., 2019. Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. BMC Bioinformatics 20, 492. https://doi.org/10.1186/s12859-019-3110-0

Kuligowski, R.J., Barros, A.P., 1998. USING ARTIFICIAL NEURAL NETWORKS TO ESTIMATE MISSING RAINFALL DATA [1]. JAWRA Journal of the American Water Resources Association 34, 1437–1447. https://doi.org/10.1111/j.1752-1688.1998.tb05443.x

Lawrence, S., Giles, C.L., 2000. Overfitting and neural networks: Conjugate gradient and backpropagation. IEEE Computer Society, Los Alamitos CA, p. Vol1.114-119.

Lin, C.-H., Wen, T.-H., 2011. Using Geographically Weighted Regression (GWR) to Explore Spatial Varying Relationships of Immature Mosquitoes and Human Densities with the Incidence of Dengue. Int J Environ Res Public Health 8, 2798–2815. https://doi.org/10.3390/ijerph8072798

Liu, Z.-N., Yu, X.-Y., Jia, L.-F., Wang, Y.-S., Song, Y.-C., Meng, H.-D., 2021. The influence of distance weight on the inverse distance weighted method for ore-grade estimation. Sci Rep 11, 2689. https://doi.org/10.1038/s41598-021-82227-y

Lodder, P., 2014. To Impute or not Impute, That's the Question, in: Advising on Research Methods: Selected Topics 2013. Johannes van Kessel Publishing.

Mamat, N., Mohd Razali, S.F., 2023. Comparisons of Various Imputation Methods for Incomplete Water Quality Data: A Case Study of The Langat River, Malaysia. Jurnal Kejuruteraan 35, 191–201. https://doi.org/10.17576/jkukm-2023-35(1)-18

Manimekalai, K., Kavitha, A., 2018. MISSING VALUE IMPUTATION AND NORMALIZATION TECHNIQUES IN MYOCARDIAL INFARCTION. ICTACT journal on soft computing 8, 1655–1662.

Martin, P., 2022. Linear Regression : An Introduction to Statistical Models. SAGE Publications, Limited, London, UNITED KINGDOM.

Matthews, S.A., Yang, T.-C., 2012. Mapping the results of local statistics. Demogr Res 26, 151–166. https://doi.org/10.4054/DemRes.2012.26.6

Nevitt, J., Hancock, G.R., 2000. Improving the Root Mean Square Error of Approximation for Nonnormal Conditions in Structural Equation Modeling. The Journal of Experimental Education 68, 251–268. https://doi.org/10.1080/00220970009600095

Osman, M.S., Abu-Mahfouz, A.M., Page, P.R., 2018. A Survey on Data Imputation Techniques: Water Distribution System as a Use Case. IEEE Access 6, 63279–63291. https://doi.org/10.1109/ACCESS.2018.2877269

Parr, C.L., Hjartåker, A., Scheel, I., Lund, E., Laake, P., Veierød, M.B., 2008. Comparing methods for handling missing values in food-frequency questionnaires and proposing k nearest neighbours imputation: effects on dietary intake in the Norwegian Women and Cancer study (NOWAC). Public Health Nutr 11, 361–370. https://doi.org/DOI: 10.1017/S1368980007000365

Pigott, T.D., 2001. A Review of Methods for Missing Data. Educational research and evaluation 7, 353–383.

Schmitt, P., Mandel, J., Guedj, M., 2015. A comparison of six methods for missing data imputation. J Biom Biostat 6, 1.

Shi, M., Su, Q., Zeng, X., 2022. Estimating the Effects of Light Rail Transit (LRT) on Land Price in Kaohsiung Using Geographically Weighted Regression. Transportation in Developing Economies 8, 9. https://doi.org/10.1007/s40890-021-00147-y

Silva-Ramírez, E.-L., Pino-Mejías, R., López-Coello, M., Cubiles-de-la-Vega, M.-D., 2011. Missing value imputation on missing completely at random data using multilayer perceptrons. Neural Networks 24, 121–129. https://doi.org/10.1016/j.neunet.2010.09.008

Singh, A., Halgamuge, M.N., Lakshmiganthan, R., 2017. Impact of Different Data Types on Classifier Performance of Random Forest, Naïve Bayes, and K-Nearest Neighbors Algorithms. International journal of advanced computer science and applications/International journal of advanced computer science & applications.

Sundararajan, A., Sarwat, A.I., 2020. Evaluation of Missing Data Imputation Methods for an Enhanced Distributed PV Generation Prediction. pp. 590–609. https://doi.org/10.1007/978-3-030-32520-6_43

Tkachenko, R., Mishchuk, O., Izonin, I., Kryvinska, N., Stoliarchuk, R., 2019. A Non-Iterative Neural-Like Framework for Missing Data Imputation. Procedia Comput Sci 155, 319–326. https://doi.org/10.1016/j.procs.2019.08.046

Wang, M., He, G., Zhang, Zhaoming, Wang, G., Zhang, Zhengjia, Cao, X., Wu, Z., Liu, X., 2017. Comparison of Spatial Interpolation and Regression Analysis Models for an Estimation of Monthly Near Surface Air Temperature in China. Remote Sens (Basel) 9, 1278. https://doi.org/10.3390/rs9121278

Willmott, C., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res 30, 79–82. https://doi.org/10.3354/cr030079

Zhang, L., Gove, J.H., Heath, L.S., 2005. Spatial residual analysis of six modeling techniques. Ecol Modell 186, 154–177. https://doi.org/10.1016/j.ecolmodel.2005.01.007

Appendix A: Proposal

Research proposal for INM363 MSc Individual Project

Title: Bridging Spatial Gaps in Electricity Network Loading: “A Machine Learning Imputation Framework”

Student: Ling-Yun Huang
Supervisor: Dr Aidan Slingsby

Introduction

Electricity network loading data is important for optimising grid management and planning infrastructure effectively. Having complete spatial data is essential for understanding how grid loads are distributed and patterned. However, missing data is common in real-world scenarios, which poses significant challenges to accurate analysis and decision-making processes. Robust spatial data is critical for predicting and reducing risks and failures, ultimately improving overall grid resilience. Therefore, studying spatial imputation techniques to address these challenges is crucial for gaining valuable insights in electricity network management.

Research Questions

This dissertation addresses the following research questions:

- How do different spatial imputation methods for electricity network loading datasets compare in terms of accuracy, efficiency, and suitability?
- What are the key factors influencing their performance and optimisation to enhance effectiveness?

Objective

This dissertation aims to systematically evaluate and compare various spatial imputation methods applied to electricity network loading datasets, focusing on assessing their accuracy, efficiency, and suitability for real-world applications. Through comprehensive comparative analyses, this research will identify the strengths and weaknesses of different spatial imputation techniques. By investigating the key factors that influence their performance, the goal is to enhance their effectiveness in predicting and filling missing data within electricity network datasets.

Outputs

This dissertation will produce valuable outcomes, including the detailed comparative analysis of different spatial imputation methods applied to electricity network loading datasets. Through this analysis, key factors influencing the performance of these methods will be identified. By systematically comparing different imputation methods and studying their performance factors, this dissertation aims to provide important insights and adopt effective imputation techniques in real-world grid management scenarios. Moreover, by studying the key factors that affect spatial imputation method performance, it will help identify critical considerations for improving these techniques. This includes understanding data characteristics, method parameters, and spatial relationship to enhance the effectiveness of these methods in electricity network datasets.

Beneficiaries

The research outcomes will have valuable insights and practical applications for various individuals involved in electricity network management. For the electricity grid planners, the findings will improve load forecasting accuracy, supporting infrastructure planning decisions, and optimising management strategies. This dissertation will also be a valuable resource for researchers in spatial data analysis. The research will offer guidelines on selecting appropriate spatial imputation methods and optimising their performance for addressing missing data in electricity network datasets.

Critical Context

In electricity network loading datasets, spatial gaps can arise due to various factors such as sensor malfunctions, incomplete data collection, or transmission errors. To address these challenges and ensure comprehensive spatial data coverage, advanced spatial imputation methods are employed.

Spatial Imputation Methods

Various methods exist for managing spatial missing values. Traditional approaches like zero-filling, mean/median imputation may introduce bias and result in misleading outcomes. Advanced spatial imputation techniques such as K-nearest neighbours (KNN), kriging, and Geographically Weighted Regression (GWR) offer more robust alternatives. KNN estimates missing values by leveraging spatial similarity among the k nearest neighbouring nodes (Kim et al., 2017). Kriging utilises optimal spatial linear prediction to estimate missing values in spatial datasets, aiming to minimise the mean squared error of predictions (Cressie, 1993). GWR estimates regression parameters based on neighbouring data points, allowing for localised predictions, and handling spatial missing values effectively (Wheeler and Páez, 2010).

Furthermore, machine learning approaches such as regression models, random forests, and neural networks (Omar et al., 2022) have proved as effective tools for spatial imputation. These spatial imputation methods are crucial for enhancing the accuracy and reliability of electricity network management by addressing spatial gaps and improving overall grid resilience.

Comparative Studies on Spatial Imputation Methods

Numerous studies have conducted comparisons of spatial imputation methods across various applications. For instance, Osman et al. (2018) compared KNN against machine learning methods, while Salmani Ghanbari and Mahmoudi (2022) and Yang et al. (2018) separately investigated the performance of KNN versus kriging in spatial imputation tasks. Additionally, Shin et al. (2016) contrasted GWR with regression models and random forests. These studies employed diverse evaluation metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Normalised Absolute Error (NAE) etc., considering varying parameters such as case numbers and spatial configurations.

However, findings from these studies exhibit notable differences. For example, Salmani Ghanbari and Mahmoudi (2022) reported that KNN achieved higher accuracy than kriging under certain circumstances, while Yang et al. (2018) presented contrasting findings, suggesting that kriging performed better in specific spatial contexts. These conflicting results highlight the importance of considering different factors when evaluating the effectiveness of spatial imputation methods. Further investigation and comprehensive analysis are necessary to determine the optimal spatial imputation approach for specific electricity network management scenarios.

Key Factors Influencing Performance and Optimisation

The effectiveness of spatial imputation methods is influenced by several key factors that must be considered. Baker et al. (2014) emphasize that the most suitable imputation method may not always be the most complex one; rather it should be selected based on the specific data characteristics and application context. Faizin et al. (2019) discuss how the type (completely at random, at random, not at random) and pattern of missing can significantly influence method performance, highlight the importance of choosing an approach tailored to specific missing data scenarios. Additionally, the spatial characteristics of the dataset, such as the density and distribution of sensor nodes, play a crucial role in determining the effectiveness of imputation methods (Decorte et al., 2024).

Methodological considerations also play a crucial role in the performance of spatial imputation techniques. For example, in KNN imputation, the choice of distance metrics directly influences imputation accuracy (Kim et al., 2017; Zhang, 2012). Similarly, in kriging methods, the selection of parameters is essential for achieving accurate predictions (Yang et al., 2018). Furthermore, Omar et al. (2022) discussed that when handling large-scale datasets, the scalability and computational efficiency of spatial imputation methods are essential. These factors are key in ensuring the efficacy of spatial imputation techniques across different applications and datasets.

Approaches

This section outlines the methodology for data preparation, application of spatial imputation methods, evaluation and comparative analysis, identification of key factors influencing method performance, and consideration of ethical and project limitations.

Data Preparation

The dataset planned for use in this project will be provided by the Advanced Infrastructure company (AITL), a complete electricity network loading dataset. This dataset will serve as the foundational dataset for simulating missing values to replicate real-world scenarios. This includes generating three types of missing data: missing completely at random, missing at random, and missing not at random. The approach for simulating missing values will be based on the methodology proposed by Santos et al. (2019), which provides a systematic and validated approach to creating realistic missing data scenarios within the dataset. Through these simulated datasets, the project will proceed to apply and evaluate different spatial imputation methods.

Spatial Imputation Methods Implementation

Following data preparation, the next step involves exploring and applying various spatial imputation methods on the simulated datasets. These methods include a range of approaches designed to address spatial missing values within the electricity network loading datasets effectively. The selected methods include mean/median imputation, KNN imputation, kriging, GWR, and machine learning techniques such as regression models and neural networks.

- **Mean/Median Imputation** is a straightforward method where missing values are replaced with the mean or median of available data for each variable. This approach will serve as a baseline for comparison with more advanced techniques and providing a simple but interpretable imputation method.
- **KNN Imputation** estimates missing values based on the similarity of data points in a feature space, using the values of the nearest neighbours to impute missing value. The selection of the parameter k, representing the number of nearest neighbours, will be optimised during the project to enhance imputation accuracy.
- **Kriging** is a predictive method that analyses spatial relationships and variability among nearby data points. By adjusting its parameters to capture spatial patterns, kriging can efficiently fill in missing values, enhancing the completeness and reliability of data for various tasks.
- **GWR** estimates regression parameters locally based on neighbouring data points. By adapting to the spatial context of the dataset, it can effectively address spatial data and improve the accuracy of predictions.
- **Machine Learning Techniques**, such as regression models and neural networks, offer sophisticated approaches to learn complex relationships and pattern in the data. Regression models excel at capturing both linear and non-linear patterns to estimate missing values and make prediction based on learned patterns. Neural network, on the other hand, utilise layers of interconnected nodes to learned patterns and predict outcomes.

By leveraging these spatial imputation methods, this project aims to comprehensively evaluate their effectiveness in handling various types of spatial missing data scenarios. Each method brings unique strengths and capabilities, contributing valuable insights into optimal strategies for spatial data imputation in the context of electricity network management and infrastructure planning.

Evaluation and Comparative Analysis

The performance of each spatial imputation method will be evaluated using established metrics including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Normalised Absolute Error (NAE). Additionally, three distinct comparative analysis approaches will be employed to assess the effectiveness of these methods, which will be crucial in determining the most effective technique for handling missing values within electricity network loading datasets.

- **Parameter Optimisation:**

  Each spatial imputation method will be tested using the same simulated dataset but with different parameter configurations. Key parameters such as k-value in KNN, variogram model in kriging, bandwidth in GWR, features in regression models, and architecture in neural networks will be adjusted to evaluate their impact on imputation accuracy.

The goal is to identify optimal parameter settings that reach the most accurate imputed values.

- **Cross-Dataset Evaluation:**

  Second, the same spatial imputation method will be evaluated across different datasets that generated under varying missing data scenarios. This comparative analysis will reveal how well each method adapts to diverse dataset characteristics. The goal is to assess their robustness and generalisability of each method across different data distributions and missing data patterns.

- **Method Comparison:**

  Third, compare different spatial imputation methods using the same set of missing values dataset. This comparative analysis will highlight the strengths and weaknesses of each imputation method in handling spatial missing data. The goal is to determine which method reaches the most accurate and reliable imputed values in each missing data type.

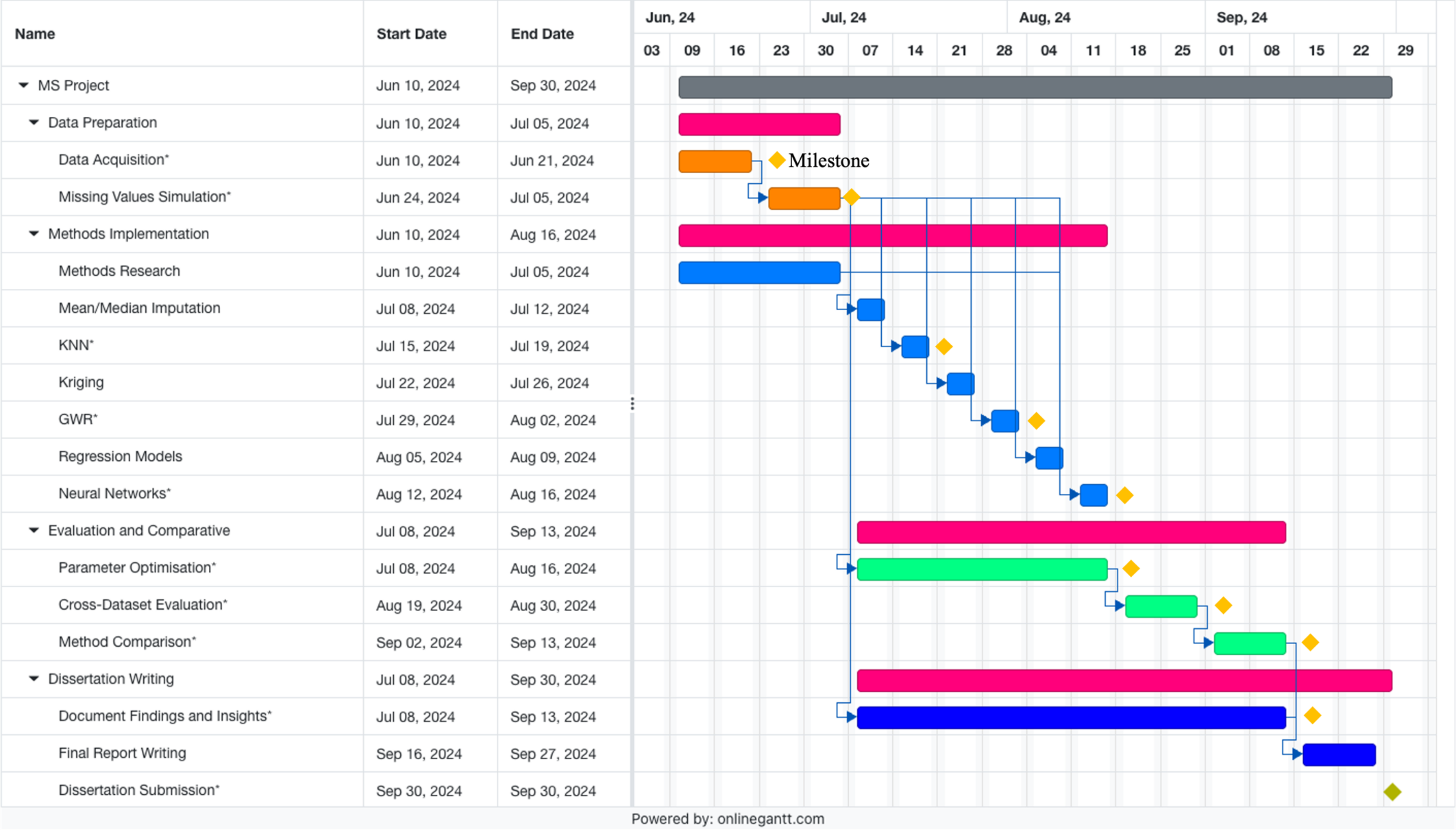
Through these comprehensive comparative analyses, key factors influencing the performance of each spatial imputation method will be identified and analysed. This includes analysing data characteristics, method parameters, and spatial relationships to gain insights into optimal method selection and configuration for electricity network loading datasets. By identifying key factors, this project will inform best practices for improving data quality and supporting effective decision-making in electricity grid management and infrastructure planning.

Ethical and Limitations Consideration

The primary ethical considerations in this project will focus on ensuring data privacy and confidentiality for the dataset provided by AITL. This includes obtaining necessary permissions for data usage and ensuring that the data is anonymized and securely stored throughout the research process. Transparency in the methodology and reporting will also be prioritised in this project. Additionally, this project will address fairness and mitigate biases in the methods employed, making sure the equitable outcomes across the applications of the research.

While this research aims to thoroughly assess spatial imputation methods for electricity network loading datasets, it still has some limitations. The effectiveness and suitability of these methods cloud depend on the quality and characteristic of the dataset from AITL, affecting the broader applicability of the findings. Additionally, complex methods like neural networks may require significant computational resources, potentially limiting their practical implementation. Although this study focuses specifically on evaluating spatial imputation methods, it may not cover all aspects of electricity management and planning. Despite these limitations, the research will offer valuable insights for improving data quality and decision-making in electricity network management.

Work Plan



Name	Start Date	End Date
MS Project	Jun 10, 2024	Sep 30, 2024
Data Preparation	Jun 10, 2024	Jul 05, 2024
Data Acquisition*	Jun 10, 2024	Jun 21, 2024
Missing Values Simulation*	Jun 24, 2024	Jul 05, 2024
Methods Implementation	Jun 10, 2024	Aug 16, 2024
Methods Research	Jun 10, 2024	Jul 05, 2024
Mean/Median Imputation	Jul 08, 2024	Jul 12, 2024
KNN*	Jul 15, 2024	Jul 19, 2024
Kriging	Jul 22, 2024	Jul 26, 2024
GWR*	Jul 29, 2024	Aug 02, 2024
Regression Models	Aug 05, 2024	Aug 09, 2024
Neural Networks*	Aug 12, 2024	Aug 16, 2024
Evaluation and Comparative	Jul 08, 2024	Sep 13, 2024
Parameter Optimisation*	Jul 08, 2024	Aug 16, 2024
Cross-Dataset Evaluation*	Aug 19, 2024	Aug 30, 2024
Method Comparison*	Sep 02, 2024	Sep 13, 2024
Dissertation Writing	Jul 08, 2024	Sep 30, 2024
Document Findings and Insights*	Jul 08, 2024	Sep 13, 2024
Final Report Writing	Sep 16, 2024	Sep 27, 2024
Dissertation Submission*	Sep 30, 2024	Sep 30, 2024

Powered by: onlinegantt.com

Figure 1. Project Gantt Chart

Risk

Risk Description	Likelihood	Impact	Mitigation Strategy
Data Quality Issues	High	High	Implement data preprocessing techniques to identify and address data quality issues early in the project. Communicate closely with AITL to ensure data completeness and employ validation methods to verify the dataset.
Inadequate Understanding of Spatial Imputation Methods	Medium	High	Conduct thorough literature review and preliminary research to build a strong understanding of spatial imputation methods. Regularly engage with the supervisor to seek guidance and feedback
Incomplete Evaluation Due to Time Constraints	Medium	Medium	Follow the work plan, continuously monitor progress, and adjust schedule as needed. If significant delays occur, discuss with the supervisor to prioritise key assessment components, and focus on key objectives within the available timeframe.
Insufficient Computational Resources	Medium	Medium	Prioritise tasks and optimise code. Use cloud computing services or parallel processing to handle large datasets.
Ethical Concerns with Data Privacy	Low	High	Obtain all necessary permissions and approvals for data usage from AITL and ensure secure data storage throughout the research process. Prepare a backup dataset from open-source resources.
Unexpected Methodological Limitations	Low	Medium	Conduct comprehensive literature reviews. Consult with supervisor regularly to help identify potential limitations early.

References

Baker, J., White, N., Mengersen, K., 2014. Missing in space: an evaluation of imputation methods for missing data in spatial analysis of risk factors for type II diabetes. Int J Health Geogr 13, 47. https://doi.org/10.1186/1476-072X-13-47

Cressie, N., 1993. Spatial Prediction and Kriging. pp. 105–209. https://doi.org/10.1002/9781119115151.ch3

Decorte, T., Mortier, S., Lembrechts, J.J., Meysman, F.J.R., Latré, S., Mannens, E., Verdonck, T., 2024. Missing Value Imputation of Wireless Sensor Data for Environmental Monitoring. Sensors 24, 2416. https://doi.org/10.3390/s24082416

Faizin, R.N., Riasetiawan, M., Ashari, A., 2019. A Review of Missing Sensor Data Imputation Methods, in: 2019 5th International Conference on Science and Technology (ICST). IEEE, pp. 1–6. https://doi.org/10.1109/ICST47872.2019.9166287

Kim, M., Park, S., Lee, J., Joo, Y., Choi, J., 2017. Learning-Based Adaptive Imputation Methodwith kNN Algorithm for Missing Power Data. Energies (Basel) 10, 1668. https://doi.org/10.3390/en10101668

Omar, M.B., Ibrahim, R., Mantri, R., Chaudhary, J., Ram Selvaraj, K., Bingi, K., 2022. Smart Grid Stability Prediction Model Using Neural Networks to Handle Missing Inputs. Sensors 22, 4342. https://doi.org/10.3390/s22124342

Osman, M.S., Abu-Mahfouz, A.M., Page, P.R., 2018. A Survey on Data Imputation Techniques: Water Distribution System as a Use Case. IEEE Access 6, 63279–63291. https://doi.org/10.1109/ACCESS.2018.2877269

Salmani Ghanbari, S., Mahmoudi, A.H., 2022. An improvement in data interpretation to estimate residual stresses and mechanical properties using instrumented indentation: A comparison between machine learning and Kriging model. Eng Appl Artif Intell 114, 105186. https://doi.org/10.1016/j.engappai.2022.105186

Santos, M.S., Pereira, R.C., Costa, A.F., Soares, J.P., Santos, J., Abreu, P.H., 2019. Generating Synthetic Missing Data: A Review by Missing Mechanism. IEEE Access 7, 11651–11667. https://doi.org/10.1109/ACCESS.2019.2891360

Shin, J., Temesgen, H., Strunk, J.L., Hilker, T., 2016. Comparing Modeling Methods for Predicting Forest Attributes Using LiDAR Metrics and Ground Measurements. Canadian Journal of Remote Sensing 42, 739–765. https://doi.org/10.1080/07038992.2016.1252908

Wheeler, D.C., Páez, A., 2010. Geographically Weighted Regression, in: Handbook of Applied Spatial Analysis. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 461–486. https://doi.org/10.1007/978-3-642-03647-7_22

Yang, H., Yang, J., Han, L.D., Liu, X., Pu, L., Chin, S., Hwang, H., 2018. A Kriging based spatiotemporal approach for traffic volume data imputation. PLoS One 13, e0195957. https://doi.org/10.1371/journal.pone.0195957

Zhang, S., 2012. Nearest neighbor selection for iteratively kNN imputation. Journal of Systems and Software 85, 2541–2552. https://doi.org/10.1016/j.jss.2012.05.073

Research Ethics Review Form: BSc, MSc, and MA Projects

Computer Science Research Ethics Committee (CSREC)
http://www.city.ac.uk/department-computer-science/research-ethics

Part A: Ethics Checklist

A.1 If you answer YES to any of the questions in this block, you must apply to an appropriate external ethics committee for approval and log this approval as an External Application through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
1.1	Does your research require approval from the National Research Ethics Service (NRES)? *e.g. because you are recruiting current NHS patients or staff?* *If you are unsure try - https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/*	**NO**
1.2	Will you recruit participants who fall under the auspices of the Mental Capacity Act? *Such research needs to be approved by an external ethics committee such as NRES or the Social Care Research Ethics Committee - http://www.scie.org.uk/research/ethics-committee/*	**NO**
1.3	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners and those on probation? *Such research needs to be authorised by the ethics approval system of the National Offender Management Service.*	**NO**
A.2 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee, you must apply for approval from the Senate Research Ethics Committee (SREC) through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
2.1	Does your research involve participants who are unable to give informed consent? *For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf.*	**NO**
2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	**NO**
2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	**NO**
2.4	Does your project involve participants disclosing information about special category or sensitive subjects? *For example, but not limited to: racial or ethnic origin; political opinions; religious beliefs; trade union membership; physical or mental health; sexual life; criminal offences and proceedings*	**NO**
2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study?	**NO**

	Please check the latest guidance from the FCO - *http://www.fco.gov.uk/en/*	
2.6	Does your research involve invasive or intrusive procedures? *These may include, but are not limited to, electrical stimulation, heat, cold or bruising.*	**NO**
2.7	Does your research involve animals?	**NO**
2.8	Does your research involve the administration of drugs, placebos or other substances to study participants?	**NO**
A.3 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee or the SREC, you must apply for approval from the Computer Science Research Ethics Committee (CSREC) through Research Ethics Online - https://ethics.city.ac.uk/ **Depending on the level of risk associated with your application, it may be referred to the Senate Research Ethics Committee.**		*Delete as appropriate*
3.1	Does your research involve participants who are under the age of 18?	**NO**
3.2	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? *This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.*	**NO**
3.3	Are participants recruited because they are staff or students of City, University of London? *For example, students studying on a particular course or module.* *If yes, then approval is also required from the Head of Department or Programme Director.*	**NO**
3.4	Does your research involve intentional deception of participants?	**NO**
3.5	Does your research involve participants taking part without their informed consent?	**NO**
3.5	Is the risk posed to participants greater than that in normal working life?	**NO**
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	**NO**
A.4 If you answer YES to the following question and your answers to all other questions in sections A1, A2 and A3 are NO, then your project is deemed to be of MINIMAL RISK. **If this is the case, then you can apply for approval through your supervisor under PROPORTIONATE REVIEW. You do so by completing PART B of this form.** **If you have answered NO to all questions on this form, then your project does not require ethical approval. You should submit and retain this form as evidence of this.**		*Delete as appropriate*
4	Does your project involve human participants or their identifiable personal data? *For example, as interviewees, respondents to a survey or participants in testing.*	**NO**