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.

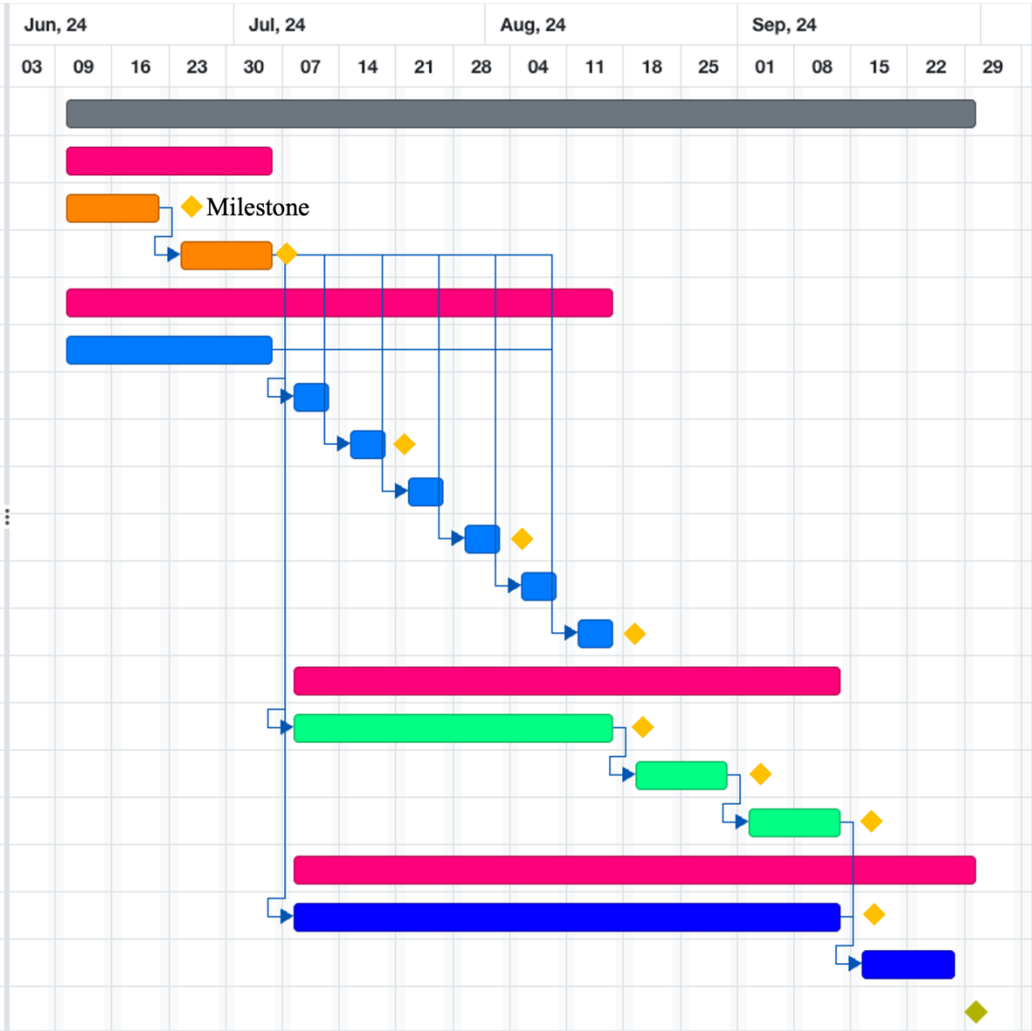
Ethical and Limitations Consideration

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**