A data-driven approach for multi-scale GIS-based domestic building energy modelling for analysis, planning, and support decision-making in Northern Ireland

Final Project Report

Department for the Economy (DfE), Northern Ireland

University College Dublin (UCD), Republic of Ireland

2026

Table of Content

1. Introduction

The report outlines the initial data analysis processes, including data collection and pre-processing, and details the implementation and results of the project. This report provides a comprehensive overview of a project on data collection, data pre-processing, feature selection, statistical analysis, and data-driven machine learning model development.

The project's methodology is explained in detail, along with the results and discussion of each step. The report also discusses the challenges posed by the available datasets and provides a summary of the Phase I assessment. The report elaborates on each step of Phase II with example output results, including EPC GIS mapping, statistical analysis, data pre-processing, and feature selection. The results and discussion of each step are presented, along with recommendations for future work. Finally, the report concludes with a discussion of future work and recommendations for further research.

Rationale – Potential benefits for NI in terms of the Energy Strategy development

This project builds upon a significant body of work that has already been performed for SEAI in the Republic of Ireland. It is envisaged that modelling results are used for spatial modelling at multiple scales, from the individual building level to the national level. Furthermore, these maps are coupled with available spatial resources (social, economic, or environmental data) for energy planning, analysis, and support decision-making. The modelling results identify clusters of buildings that have a significant potential for energy savings within a defined region.

Urban planners and energy policymakers often face significant challenges when implementing sustainable energy analysis and planning on a large scale due to the complexity of the energy system. System complexity increases exponentially from the individual building level to the national level, mainly due to a significant increase in the number of buildings and associated data resources. This project formulates a methodology to integrate various data resources using machine learning algorithms effectively; this can facilitate energy planning at multiple scales. This project uses limited available resources such as energy performance certificates and geographical, spatial, and census data to predict the building energy performance of NI. This work estimates building energy performance from a local to national scale with limited knowledge of building dynamics.

Moreover, planning at the urban scale often involves a significant amount of building data. Urban planners could use this project to integrate various data resources using machine learning techniques; this eventually facilitates planning at multiple scales. Furthermore, urban planners could use the GIS modelling results to develop energy efficiency areas of the urban building stock and identify priority areas to implement energy efficiency solutions in NI.

Objectives – What the proposer would set out to achieve.

The objectives of the project are to implement multi-scale domestic building energy performance analysis in order to help stakeholders (urban planners, energy policymakers, and local authorities) analyse multi-scale building energy performance and identify priority areas for energy planning, analysis, and support decision-making using a GIS-based data-driven approach.

2. Phase I - Assessment phase

This phase of the project discusses the initial data analysis, including the issues and challenges encountered during this phase. The key project outputs are also presented, and the deliverables and outcomes of the project are defined. Finally, recommendations for Phase II are provided.

Anticipated datasets

1. EPC domestic data set for Northern Ireland
2. OSNI POINTER–addressable dataset for Northern Ireland
3. OSNI Fusion - Building footprint for Northern Ireland
4. Domestic Property Data for Northern Ireland
5. OSNI Open data boundaries outline for Northern Ireland
    - Country
    - Counties
    - Local Government Districts
    - District Electoral Areas
    - Super Output Areas
    - Small Area
6. 2011 Census data in Northern Ireland
7. Year of construction band dataset for Northern Ireland
8. House Condition Survey 2016 dataset (Optional)
9. Total final energy consumption at regional and local authority level: 2005 to 2019 (Optional)

Issues and challenges in datasets

The following issues and challenges arose during the initial analysis of all the datasets that will be used in this project.

- The EPC domestic data set for Northern Ireland has multiple issues, such as
    - Multiple entries for one household.
    - Many missing values in the data.
    - The limited number of input variables for data-driven model development
    - Required input variables values are categorial in format, e.g., floor energy efficiency: very good; good; average; poor; very poor.
- The OSNI Fusion Building footprint data are missing some new addresses that are found in the OSNI Pointer addressable dataset for Northern Ireland.

- All the OSNI open data boundaries outline shape files are based on the year 2012.
- Census data required for this project are over 10 years old.
- The House Condition Survey 2016 dataset is only useful for validation, but implementation depends upon matching variables or features with EPC data.
- The total final energy consumption at the regional and local authority level from 2005 to 2019 is useful, but data would be helpful if available at a more granular level, such as a small area.
- Linking between different data sets is a challenging task due to limited variables.

Final prediction variables (Key Project Outputs)

Based on an initial analysis of the EPC data set following variables will be used for prediction.

1. CURRENT_ENERGY_EFFICIENCY (Numeric)
2. ENVIRONMENT_IMPACT_CURRENT (Numeric)
3. CURRENT_ENERGY_RATING (Letter)

Delivery and Output

What would actually be produced at the end of the project, and who would be involved alongside the experience of the staff?

This project implements the data-driven approach using GIS and machine learning models for building energy modelling at multiple scales. The potential output of this project would be spatial modelling of Northern Ireland's building energy performance data. The project would divide into the following deliverables:

1. Data Collection and initial analysis
2. Data Pre-processing, Feature Selection
3. Data-driven Machine Learning Model Development
4. EPC GIS mapping
5. Publish results in a final report and ArcGIS online.
6. Presentation

Tools and Technologies

The following tools and technologies are used in this project.

- Python
- Rapid Miner Studio (Educational Licence)
- ArcGIS Desktop 10.6 (UCD Licence)
- ArcGIS Online (UCD & DfE Licence)
- Microsoft Excel / Google Sheet
- Microsoft Word / Google Doc

Initial Data Analysis

The initial analysis of the data sets indicates a few issues and limitations impacting model development. However, the data provided by DfE is sufficient to achieve the project's goals. Some assumptions must be made when linking the data from one source to another, and will primarily involve using UPRNs. These assumptions will clearly be listed in the project's final report. The EPC 'Energy efficiency' ratings and the EPC 'Environmental Impact' ratings (as recorded in the EPC dataset) are independent output variables. Therefore, the model formulation will include a separate model to predict these variables individually. The developed workflow uses the building archetypes approach to estimate the input of non-EPC properties developed at the small area level for a refined-granular analysis. The building archetypes approach is a method that groups properties with similar physical and construction characteristics into representative categories. As per the initial analysis, building archetypes will be developed based on the building type and year of construction. The archetype variables (building type and year of construction) are extracted from the property data for non-EPC properties.

The statistical modelling procedure will aggregate the EPC properties at a small area level and subsequently map the result for a detailed analysis of the heating type, low-energy lighting, and other energy-efficiency parameters. These assumptions and results could be validated using the existing available census or national statistical data at the small area level.

The 2016 House Condition Survey data could be used for the validation of prediction and statistical results. However, the integration of this dataset is limited by the variables or features in common with the EPC dataset.

Recommendations for Phase II

Following are the recommendations for the project that can help in phase II of the project.

- The UPRN will link the data from different sources.
- The non-EPC properties will be estimated from property value and small area EPC datasets by using the building archetypes approach based on the building type and year of construction.
- The result could further improve if we can use the most recent version of OSNI open data boundaries outlines shape files data.
- The total final energy consumption from 2005 to 2019 can be used for archetypes validation. Similarly, the metered gas connection within the geographical area dataset could also help validate archetypes.
- Finally, continuous feedback is required on each task to complete the project successfully. However, any significant change in the project tasks will affect deliverables and delay the project's completion.

3. Phase II - Implementation phase

Phase 2 - The implementation phase is the development of the methodology for determining the relationships with existing EPC data which could be deployed on buildings with no EPC data to provide an estimation across all Northern Irish buildings ultimately.

The GIS-based mapping of building energy performance follows these steps (Figure 1):

1. Data collection from different resources for NI;
2. Data pre-processing, employ data-driven approaches to improve the quality of the data along with feature selection;
3. EPC modelling, data-driven machine learning modelling at a large scale using a bottom-up approach; The data-driven model development step predicts building energy performance.
4. The energy planning step analyses the modelling results for planning or decision-making using GIS-based EPC analysis.

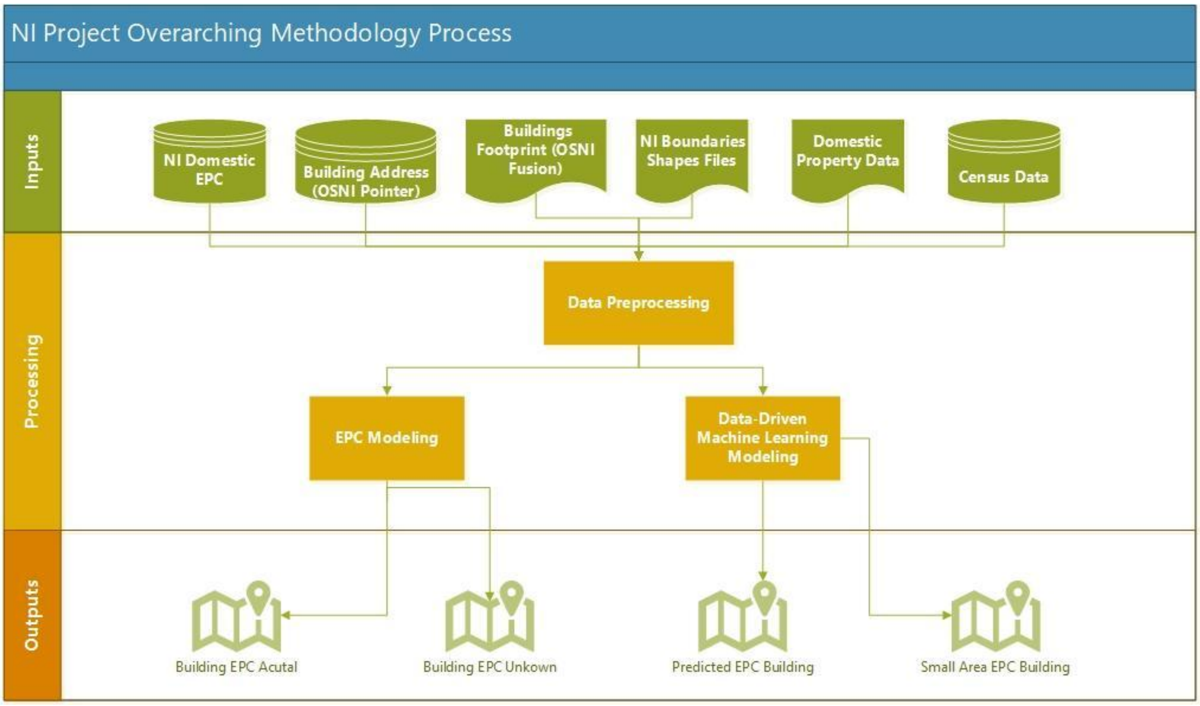

Figure 1 Overarching methodology for GIS-based mapping of multi-scale residential building energy performance using data-driven approaches.

Data Collection

EPC domestic data set for Northern Ireland

The EPC domestic data set for Northern Ireland is a comprehensive collection of information on the energy efficiency and environmental impact of residential properties in the region. This data set includes a range of variables such as property type, age, insulation levels, heating systems, and carbon emissions. This is an essential dataset for policymakers, researchers, and other stakeholders seeking to understand energy use patterns and identify opportunities for energy efficiency improvements in Northern Ireland's housing stock. The EPC data set provides valuable insights into how energy is used in homes across the region and can help inform policy decisions to reduce energy consumption and emissions in the residential sector.

Domestic Pointer Data for Northern Ireland

Pointer is the address database for Northern Ireland and is maintained by Land & Property Services (LPS), with input from Local Councils and Royal Mail (RM). This is the common standard address dataset for every property in Northern Ireland. It is important to note that Pointer is a dataset for addressable buildings in Northern Ireland. Each primary Addressable Object (PAO) building is given a Unique_Building_ID. Each addressable unit within a building, which is a Secondary Addressable Object (SAO), is given a Unique Property Reference Number (UPRN).

OSNI Fusion - Building footprint for Northern Ireland

OSNI Fusion is the definitive large-scale vector map for Northern Ireland and replaces the OSNI Largescale vector. It consists of a more intuitive, simplified data structure and contains 13 main themes, with the existing large-scale feature types migrated and stored as attributes. Each feature has its unique identifier and includes complete life cycle information. Additionally, OSNI Fusion contains a comprehensive coverage of polygons for the ground surface, consisting of water, transport, and land features. In this project, we are only focused on the building theme.

Domestic Valuation for Northern Ireland

The Ordnance Survey of Northern Ireland (OSNI) provides the Domestic Valuation dataset and contains information on the capital value of residential properties in Northern Ireland. The data is compiled by Land and Property Services (LPS) and includes details such as the address, valuation band, and capital value of each property.

OSNI Open data boundaries outline for Northern Ireland

The OSNI Open Data Boundaries Outline for Northern Ireland is a publicly available dataset that provides digital boundaries of administrative areas within Northern Ireland. This dataset includes information on a small area, counties, super output areas, and other administrative areas. The dataset is designed to be easily integrated into mapping and geographic information systems (GIS), making it a valuable resource for spatial analysis and decision-making. The OSNI Open Data Boundaries Outline is updated regularly to reflect changes in administrative boundaries, ensuring it remains a reliable and up-to-date resource for users. This dataset is freely available for download from the OSNI Open Data Portal, managed by Land & Property Services (LPS) in Northern Ireland. Anyone can use it for a wide range of purposes, including research, planning, and policymaking.

Data Pre-processing, Feature Selection

This process follows data collection and involves an initial statistical analysis, pre-processing, and feature selection. Furthermore, this process employs various data-driven techniques, such as data pre-processing, to extract building characteristics and associated energy usage of individual buildings.

This study relies on EPC data publicly available in most countries for building stock development and enhancement. As the EPC data are usually collected from statistical surveys or questionnaires, many anomalies can exist in the dataset. Data-driven methods such as data pre-processing and outlier detection can eliminate/treat these existing anomalies and aid in extracting clean data.

Statistical analysis

The first sub-step of our analysis involves conducting a preliminary statistical examination of the building stock dataset. We use visual analysis techniques, including density plots, histograms, and box plots, to assess the quality of the data and identify appropriate processing techniques. Table 1 presents the overall statistics of the datasets utilized in this project.

Table 1 Overall statistics of the datasets utilized in this project.

Data	Count
Domestic EPC OSNI amended	570,420
Domestic EPC (Remove missing UPRN and Duplicate) + Enhanced Domestic property	452,104
Domestic Valuation OSNI amended	818,347
Domestic Valuation OSNI amended with LPS Filter	809,776
POINTER (Domestic Building Points)	966,592
POINTER (Domestic Building Points) with LPS Filter	867,514
Small Area 2011	4,537
Super Output Area 2011	890
Domestic Valuation and POINTER (Common) with LPS Filter	809,079
Enhanced Domestic property dataset ->Domestic Valuation and POINTER (Combined) with LPS Filter	867,514
EPC Pre-processed	442,298

The Energy Performance Certificate (EPC) provides a rating of a building's energy efficiency and carbon emissions on a scale from A to G, with A being very efficient and G being very inefficient. This rating is used to measure a building's overall energy efficiency, captured through the Energy Efficiency Rating (EER). Figure 2 Energy Efficiency Ratings of EPC data illustrates the breakdown of EER, which ranges from 1 to 100, with 100 representing the most efficient rating and lower energy costs. EER is also grouped into energy efficiency bands from A to G, with A being the most efficient.

The calculation of EER is based on the cost of heating the property, heating water, and lighting, providing a comprehensive assessment of a building's energy efficiency.

Figure 2 Energy Efficiency Ratings of EPC data

The statistical analysis of the Energy Performance Certificate (EPC) domestic data reveals the breakdown of energy efficiency ratings in Northern Ireland. The distribution of ratings appears to be approximately normal, with the majority of households falling into the middle energy efficiency rating (D) and very few properties achieving the highest energy efficiency rating (A) (as shown in Figure 3 and Figure 4).

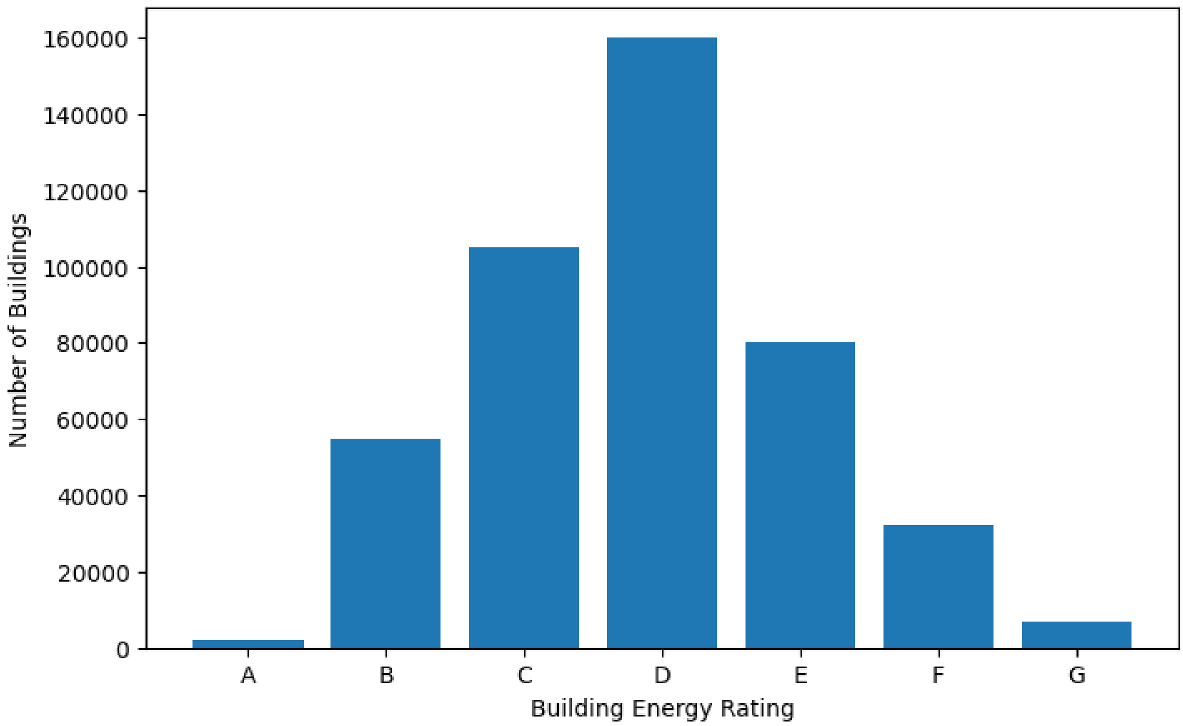

Figure 3 Distribution of building energy rating (EPC) in Northern Ireland

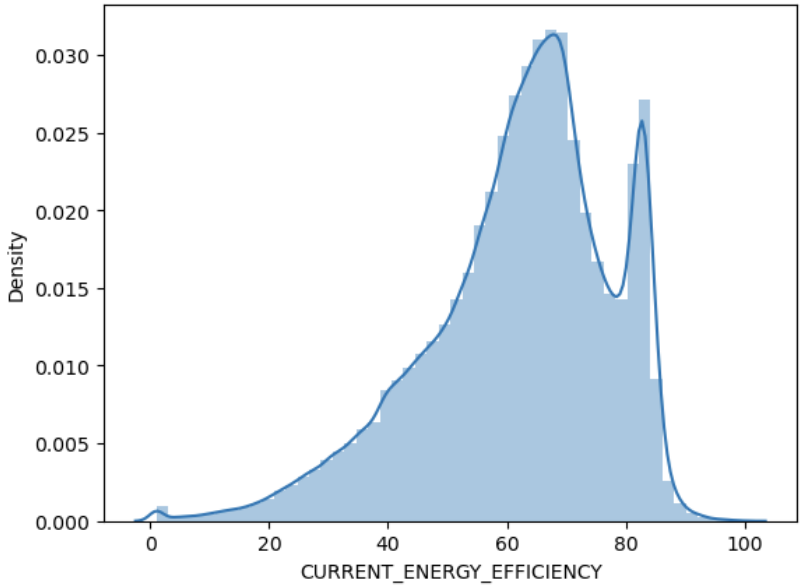

Figure 4 Distribution of Building Current Energy Efficiency Ratings (EER) in Northern Ireland calculated using a histogram with kernel density estimation (KDE). The histogram represents the density of buildings across EER score ranges

Data pre-processing

Real-world data, including data collected through surveys such as EPCs, often contain irrelevant, incomplete, noisy, redundant, and inconsistent information. EPC surveys are conducted by dedicated assessors who follow strict guidelines and use specialized software to complete the calculations. However, in situations where assessors cannot access the entire property, some measures may need to be estimated based on their experience. This can make extracting useful or accurate knowledge from the data challenging. Therefore, data pre-processing is an essential step that follows statistical analysis and eliminates any inconsistencies that are identified before the data is further used.

- **Data cleaning** involves filling in missing or zero values, removing duplicate data, smoothing out noisy data, and resolving inconsistencies. For example, missing or zero building fabric area values may be replaced with an average or median value. If data is still missing, the mode and median of Small Area and building type can be used. Super Output Area and building may also be utilized for handling missing value data.

- **Data integration** combines data from multiple sources, such as redefining building type descriptions according to common standard datatype.

- **Data reduction** removes irrelevant attributes such as IDs, dates, and time.

- **Data transformation** replaces or adds new variables inferred from existing variables, such as the overall building U-value.

- **Data discretization** transforms numeric data by mapping values to interval concept labels and for example, converting the year of construction into age bands.

The first step of this project involves generating an enhanced domestic property dataset by combining data from the domestic valuation and pointer databases to cover the total number of domestic buildings in Ireland ((figure 5). The data cleaning process on domestic valuation involves several steps, including removing irrelevant or erroneous information from the dataset. For example, in this case, we need to remove the property numbers prefixed with 'N' or 'M', which may be incorrect or irrelevant to the analysis. Additionally, we need to remove any Unique Property Reference Numbers (UPRNs) prefixed with 'E,' as they may not be valid or accurate.

Furthermore, we need to remove any missing SA Code, which is a code used to identify a small area where the property is located. This is an important step to ensure the data's accuracy and prevent any biases or errors in the analysis. By removing these specific data points or codes, we can improve the quality and reliability of the domestic valuation dataset and ensure that it provides accurate and useful information for any analysis or decision-making purposes.

Furthermore, data cleaning of domestic pointer datasets involves several steps to ensure the accuracy and reliability of the data. The first step is to remove any data points with a CLASSIFICATION value of NON-POSTAL or DO_OTHERS or start with ND. Next, we need to remove any data points with an ADDRESS_STATUS value of HISTORICAL or REJECTED. Additionally, we need to remove any data points with a BUILDING_STATUS value of DEMOLISHED, DERELICT, or UNDER CONSTRUCTION.

Furthermore, we must remove any BUILDING_NUMBER that starts with 'N' or 'M' followed by a number. We also need to remove any missing UPRN and SA Code to ensure the dataset is complete and accurate. These steps are crucial to ensure that the data is free from errors, inconsistencies, and biases, allowing for precise analysis and decision-making.

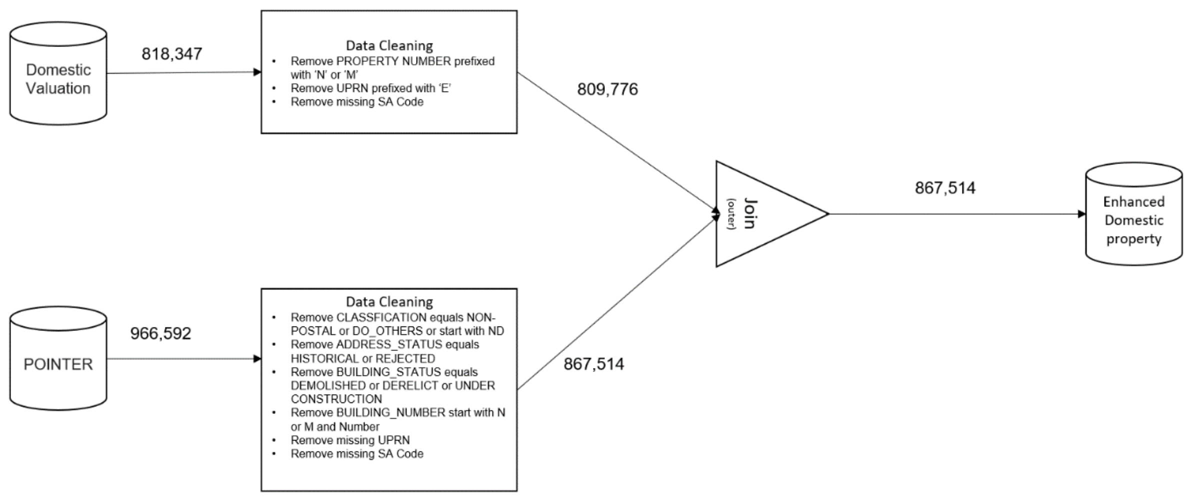

Figure 5 Process diagram of Enhanced Domestic property dataset generation

Feature selection

This project identifies the key features that significantly influence building energy performance at the building stock level and selects the optimal features for retrofit measures. The feature selection process involves removing irrelevant and redundant information and only selecting the set of informative features that influence building energy performance.

Two approaches were implemented to select these features: the engineering approach, which selects features based on existing literature and expert/survey reports, and the data-driven approach, which determines the key features using machine learning algorithms. The data-driven approach is particularly useful for large or complex datasets that contain irrelevant and redundant features, as it reduces the complexity of the model.

Finally, the pre-processed EPC data for modelling was generated using pre-processing and feature selection techniques, which are explained in Figure 6. The EPC domestic dataset initially contained 570,420 rows. After data pre-processing and feature selection, the final dataset was reduced to 433,255 rows used for the modelling process.

The EPC dataset originally had 178 features, but only 45 features were used for modelling after the feature selection process. Additionally, we joined the EPC dataset with the enhanced domestic property dataset, which is an inner join of the domestic valuation and pointer databases. This step was taken to ensure that the dataset was complete and accurate.

These are the pre-processing steps applied on EPC domestic data set:

- Domestic EPC OSNI: (a) removing rows with PrimaryLast ==0 & missing UPRN, and (b) removing AGE_CATEGORY 12 and 13.
- Engineering Feature Selection: A manual feature selection process was performed to select the most relevant features for the analysis.
- Remove wrong energy, env values, SAP05 rows: Rows with incorrect energy and environmental values and SAP05 rows were removed.
- Rows with an unknown U-value were removed.
- Rows with missing Small Area SA2011 values were removed.
- Map Build Type: A mapping process was performed to categorize building types.
- Categorial Variables Reduction: Categorical variables were reduced to a smaller number of categories.
- Numeric Conversion: Numeric variables were converted to a common unit of measurement to ensure consistency across the dataset.

Overall, these pre-processing steps were performed to ensure that the data was consistent, complete, and relevant for the analysis.

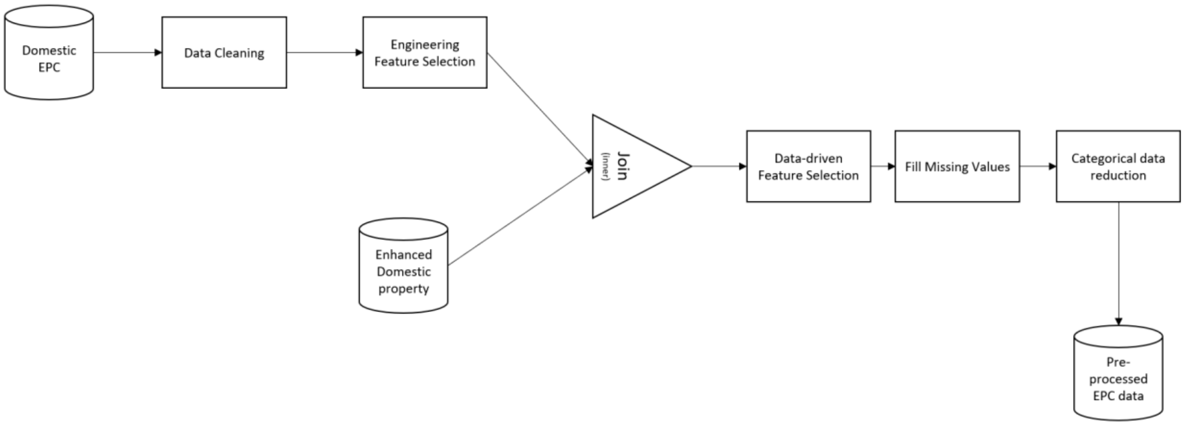

Figure 6 Process diagram of EPC data pre-processing and feature selection.

Data-driven Machine Learning Model Development

The first step in the data-driven machine learning modelling process involves filtering important features that yield the best results for machine learning algorithms (Figure 7). Before implementing these algorithms, the data is split into two subsets to avoid overfitting: a training dataset (used to train the model) and a test dataset (used to test the final trained model). This step also helps evaluate the performance of the machine learning model.

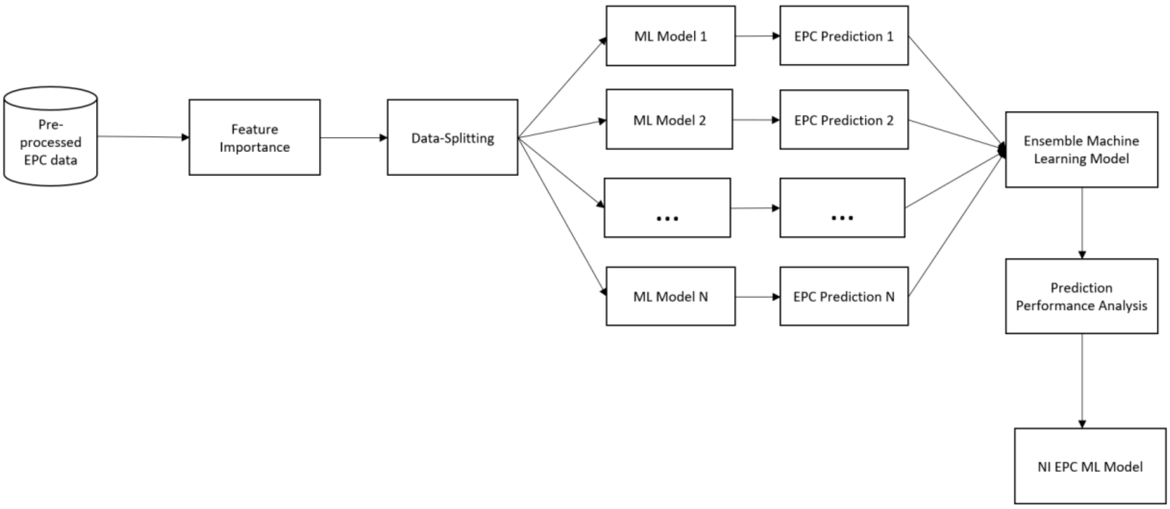

Figure 7 Methodology for data-driven machine learning modelling to predict Energy Use Intensity (EUI)

In this project, we used two data-splitting methods: random data splitting and cross-validation. Random data splitting involves randomly dividing the data into training and test subsets. Cross-validation, on the other hand, involves dividing the data into k subsets, training the model on k-1 subsets, and testing it on the remaining subset. This process is repeated k times, and the results are averaged to evaluate the machine learning model's performance.

The workflow formulates regression machine learning models to predict building energy performance in terms of Energy Use Intensity (EUI). Instead of using a traditional single-machine learning approach, this paper implements ensemble methods, in this case the stacking technique for modelling, to test multiple learning algorithms and obtain better predictive performance. This method aims to reduce the generalization error of different machine learning models.

Ensemble learning techniques typically use multiple machine-learning models with the best prediction performance, such as Linear Regression (LR), Neural Network (NN), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbour (KNN), Support Vector Regression (SVR), and Gradient Boosting (GB).

Linear regression trains the model with coefficients to minimize the residual sum of squares between the observed output in the dataset and the output predicted by the linear approximation. Neural Network, also known as Multi-layer Perceptron regressor, uses artificial neural network architectures. Decision tree builds regression models as tree-like structures, each node representing a splitting rule for one specific attribute.

Random Forest is a meta-estimator that trains several classifying decision trees on various sub-samples of the dataset and uses averaging to improve predictive accuracy. The k-nearest neighbour algorithm is a simple, easy-to-implement approach that can solve classification and regression problems. SVR is a type of Support Vector Machine (SVM) that supports linear and non-linear regression using a subset of training points in the decision function.

Gradient Boosting is a regression-based tree model and one of the most effective machine-learning algorithms for improving the accuracy of trees. It uses a flexible non-linear regression procedure with forward stage-wise methods. There are different implementations of gradient-boosted machine learning algorithms, such as XGBoost (Extreme Gradient Boosting), Histogram-Based Gradient Boosting (HGB), and LGBM (Light Gradient Boosted Machine).

Adopted performance indices, such as R-Squared (R2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), were used to evaluate the effectiveness of each machine learning model in predicting Energy Use Intensity (EUI). The machine learning model with the lowest RMSE and MAE values, and an R-Squared value closer to 1, was considered the best among all models.

R-Squared (R2):

$$R2 = 1 - (SS\_res \,/\, SS\_tot)$$

where SS_res is the sum of squared residuals, and SS_tot is the total sum of squares.

Mean Absolute Error (MAE):

$$MAE = (1/n) * \Sigma|actual - predicted|$$

where n is the number of observations in the dataset.

Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{}(\Sigma(actual - predicted)^2 \,/\, n)$$

where n is the number of observations in the dataset.

Finally, the predicted EUI values were converted into an Energy Performance Certificate (EPC) label or rating to calculate the model's accuracy. By comparing the predicted EPC rating to the actual EPC rating, we determined the accuracy of the machine learning model in predicting building energy performance. In order to evaluate the accuracy of the model for each building rating, an evaluation confusion matrix is utilized. A confusion matrix is a table used to evaluate the performance of a model. It shows the number of true positive, false positive, true negative, and false negative predictions for each class in the dataset. The rows in the matrix represent the actual classes, while the columns represent the predicted classes. The diagonal elements of the matrix represent the correct predictions, while the off-diagonal elements represent the incorrect predictions. The confusion matrix provides a helpful way to evaluate the accuracy, precision, recall, and F1-score of a model, and it can also be used to identify potential issues with the model's performance, such as imbalanced classes or misclassified samples. They are calculated using the following formulas:

Precision = true positives / (true positives + false positives)

Recall = true positives / (true positives + false negatives)

F1-score = 2 * (precision * recall) / (precision + recall)

In the initial phase, we aimed to identify the best machine-learning models for predicting EUI using pre-processed EPC datasets. For this purpose, we tested the dataset with 20 different machine-learning algorithms. Table 2 shows the performance results of these models in terms of predicting building energy performance in the form of Energy Use Intensity (EUI).

The Root Mean Squared Error (RMSE) is used to measure the difference between the predicted and actual values of EUI. The lower the RMSE value, the better the model's predictive accuracy. The XGBRegressor model has the lowest RMSE value of 4.31, followed closely by ExtraTreesRegressor (4.45), HistGradientBoostingRegressor (4.46), LGBMRegressor (4.46), and RandomForestRegressor (4.47).

Table 2 Performance results of 20 different machine learning models that are used to predict building energy performance in terms of Energy Use Intensity (EUI).

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
XGBRegressor	0.92	0.92	4.31	11.48
ExtraTreesRegressor	0.92	0.92	4.45	327.99
HistGradientBoostingRegressor	0.92	0.92	4.46	4.39
LGBMRegressor	0.92	0.92	4.46	2.39
RandomForestRegressor	0.92	0.92	4.47	351.53
MLPRegressor	0.91	0.91	4.54	370.93

BaggingRegressor	0.91	0.91	4.68	40.41
GradientBoostingRegressor	0.90	0.90	4.96	90.18
SVR	0.90	0.90	5.00	12036.08
NuSVR	0.90	0.90	5.01	22067.40
KNeighborsRegressor	0.87	0.87	5.68	19.12
DecisionTreeRegressor	0.84	0.84	6.25	6.51
ExtraTreeRegressor	0.84	0.84	6.29	3.93
BayesianRidge	0.82	0.82	6.60	1.33
TransformedTargetRegressor	0.82	0.82	6.60	1.31
LinearRegression	0.82	0.82	6.60	1.17
RidgeCV	0.82	0.82	6.60	1.62
Ridge	0.82	0.82	6.60	0.80
LassoCV	0.82	0.82	6.60	4.68

The MLPRegressor, BaggingRegressor, and GradientBoostingRegressor models have slightly higher RMSE values ranging from 4.54 to 4.96. The SVR and NuSVR models have even higher RMSE values of 5.00 and 5.01, respectively, indicating that these models are less accurate in predicting EUI.

The KNeighborsRegressor, DecisionTreeRegressor, ExtraTreeRegressor, BayesianRidge, TransformedTargetRegressor, LinearRegression, RidgeCV, Ridge, and LassoCV models have higher RMSE values ranging from 5.68 to 6.60. These models are less accurate in predicting EUI, and their computational times vary from less than 1 second to more than 22,000 seconds.

Overall, the XGBRegressor model has the best predictive accuracy and computational time performance, followed by ExtraTreesRegressor, HistGradientBoostingRegressor, LGBMRegressor, and RandomForestRegressor. The choice of the best model ultimately depends on the specific requirements of the application and the trade-off between accuracy and computational time.

Next, we aim to identify the important features based on their performance. To evaluate the contribution of each feature to the final model's predictions, we rely on SHAP (SHapely Additive exPlanations) feature importance. SHAP is a model-agnostic method used to interpret the output of machine learning models. It provides a way to estimate the contribution of each feature in a given prediction, which can be useful for understanding the model's decision-making process and identifying important features.

The SHAP value for a particular feature of a given observation represents the contribution that the feature made to the final score in terms of the difference in the estimate. The relative importance of each feature to the model is measured by calculating the mean of the absolute values of these SHAP values for each feature.

To compute the SHAP values for a specific prediction, the method generates a set of artificial samples that vary each feature's value while holding the other features constant. The machine learning model is then used to predict the output for each artificial sample. The SHAP values are computed based on the difference between the predicted output and the baseline output (the average prediction across all the samples).

The SHAP values provide an estimate of the contribution of each feature to the prediction output. Positive SHAP values indicate that the feature positively contributed to the prediction output, while negative SHAP values indicate the opposite. The SHAP value's magnitude reflects the contribution's strength, with larger values indicating more significant contributions.

Overall, the SHAP feature importance method is a powerful tool for interpreting the output of machine learning models and identifying important features. It is model agnostic, meaning it can be used with any machine learning model and provides intuitive and easily interpretable results.

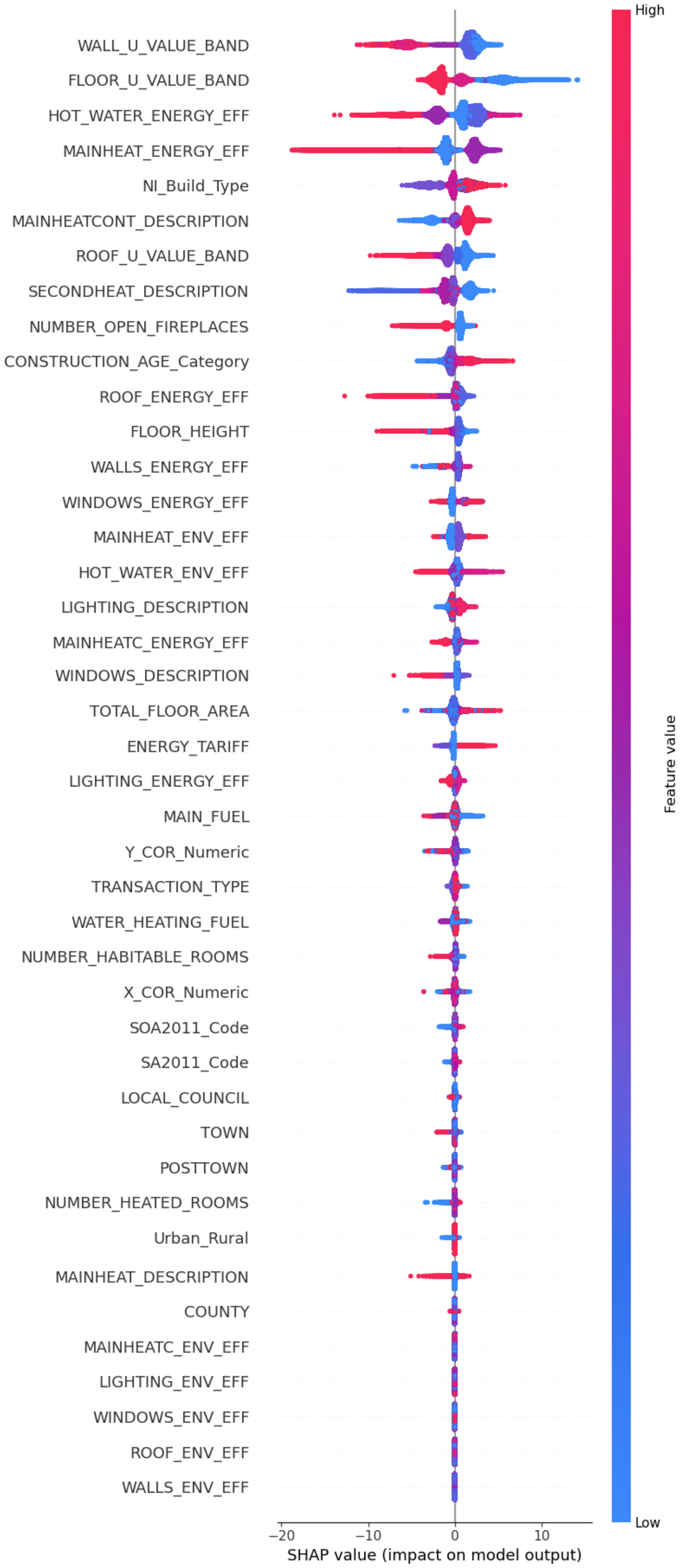

Figure 8 Importance of different features in predicting building energy performance in terms of Energy Use Intensity (EUI), as determined by SHAP (SHapely Additive exPlanations) feature importance

Figure 8 shows the importance of different features in predicting building energy performance in terms of Energy Use Intensity (EUI), as determined by SHAP (SHapely Additive exPlanations) feature importance. The feature "WALL_U_VALUE_BAND" has the highest importance, followed by "FLOOR_U_VALUE_BAND" and "HOT_WATER_ENERGY_EFF". The least important feature is "WINDOWS_ENV_EFF", "WALLS_ENV_EFF", "MAINHEATC_ENV_EFF", "ROOF_ENV_EFF", "LIGHTING_ENV_EFF" with an importance value of 0.00. These results indicate that building envelope features such as wall and floor U-value and hot water energy efficiency have a higher impact on building energy performance compared to other features.

After analysing the feature importance using SHAP values, the most important features were identified and selected for the final machine learning model. Out of the initial 41 inputs features, only 26 features were considered important in predicting building energy performance. These 26 features include:

1. FLOOR_HEIGHT
2. HOT_WATER_ENERGY_EFF
3. HOT_WATER_ENV_EFF
4. LIGHTING_DESCRIPTION
5. LIGHTING_ENERGY_EFF
6. MAIN_FUEL
7. MAINHEAT_DESCRIPTION
8. MAINHEAT_ENERGY_EFF
9. MAINHEAT_ENV_EFF
10. MAINHEATC_ENERGY_EFF
11. MAINHEATCONT_DESCRIPTION
12. NUMBER_HABITABLE_ROOMS
13. NUMBER_HEATED_ROOMS
14. NUMBER_OPEN_FIREPLACES
15. ROOF_ENERGY_EFF
16. SECONDHEAT_DESCRIPTION
17. TOTAL_FLOOR_AREA
18. WALLS_ENERGY_EFF
19. WATER_HEATING_FUEL
20. WINDOWS_DESCRIPTION
21. WINDOWS_ENERGY_EFF
22. CONSTRUCTION_AGE_Category
23. FLOOR_U_VALUE_BAND
24. ROOF_U_VALUE_BAND
25. WALL_U_VALUE_BAND
26. NI_Build_Type

These selected features were used in the final ensemble-based machine learning model to predict building energy performance accurately. After performing the important feature analysis using the SHAP method, the following 15 features were removed from the model development:

1. WALLS_ENV_EFF
2. ROOF_ENV_EFF
3. WINDOWS_ENV_EFF
4. LIGHTING_ENV_EFF
5. MAINHEATC_ENV_EFF
6. COUNTY
7. Urban_Rural
8. POSTTOWN
9. TOWN
10. LOCAL_COUNCIL
11. SA2011_Code
12. SOA2011_Code
13. TRANSACTION_TYPE
14. X_COR_Numeric
15. Y_COR_Numeric

These features were found to have a negligible impact on the final model's predictions and therefore were not included in the final 26 selected features for the ensemble-based machine learning model development.

The best-performing machine learning model is selected for ensemble-based modelling using the stacking regression method in the final step. Table 3 presents the performance metrics of six regression models evaluated to predict building energy performance, including root mean squared error (RMSE), mean absolute error (MAE), and R-squared values. The individual base learners considered for stacking include XGBRegressor, LGBMRegressor, RandomForestRegressor, ExtraTreesRegressor, and HistGradientBoostingRegressor.

As shown in Table 3, the gradient boosting–based models (XGBRegressor, LGBMRegressor, and HistGradientBoostingRegressor) demonstrate strong predictive performance, with relatively low RMSE and MAE values and high R-squared scores. These top-performing models were therefore selected as base learners for the stacking ensemble.

The StackingRegressor, which combines the predictions of these base models, achieves the lowest RMSE (4.15) and MAE (2.88), along with a high R-squared value (0.93), indicating that the ensemble approach provides a marginal but consistent improvement in predictive accuracy compared to individual models.

Table 3 Performance metrics of different regression models used in the project to predict building energy performance

Model Name	RMSE	MAE	R-squared
StackingRegressor	4.15	2.88	0.93
XGBRegressor	4.18	2.89	0.93
LGBMRegressor	4.18	2.94	0.93
HistGradientBoostingRegressor	4.20	2.96	0.93
RandomForestRegressor	4.58	3.18	0.91
ExtraTreesRegressor	4.58	3.19	0.91

Finally, to estimate the accuracy of the model, the predicted EUI values are converted to energy ratings. A confusion matrix in Figure 9 and Table 4 summarizes the performance of a StackingRegressor machine learning model by comparing its predicted labels with the actual labels. Table 4 shows the confusion matrix of a model that predicted energy performance labels for a dataset of 129,977 samples. The classes are labeled from A to G, and the table shows each class's precision, recall, and F1 score, as well as the support (number of samples) for each class. The accuracy of the model is 0.80, which means that 80% of the predicted labels match the actual labels. In this matrix, Class B has the highest precision value of 0.93, indicating that 93% of the positive predictions for Class B were correct. Class A's lowest precision value of 0.75 indicates that only 75% of the positive predictions for Class A were correct.

Conversely, recall is the ratio of correctly predicted positive observations to the total actual positive observations. In this matrix, Class D has the highest recall value of 0.84, indicating that 84% of the actual positive observations for Class D were correctly predicted. Class A has the lowest recall value of 0.08, indicating that only 8% of the actual positive observations for Class A were correctly predicted.

Overall, the matrix shows that the model performed well in predicting Classes B, C, D, and E with relatively high precision and recall values. However, the model had difficulty in predicting Classes A and G, as indicated by the low precision and recall values. The weighted average precision and recall values are both 80%, indicating an overall good performance of the model.

Moreover, in order to develop a model for predicting the current environmental impact, we utilized the same machine-learning models (XGB, LGBM, RandomForest, ExtraTrees, and HistGradientBoosting).

Table 4 Summary of the performance of a StackingRegressor machine learning model by comparing its predicted labels with the actual labels.

Class	Precision	Recall	F1-score	Support
A	0.75	0.08	0.15	183
B	0.93	0.86	0.89	16,262
C	0.83	0.76	0.80	30,947
D	0.80	0.84	0.82	47,700
E	0.71	0.76	0.73	23,934
F	0.70	0.70	0.70	9,184
G	0.78	0.62	0.69	1,767
Accuracy			0.80	129,977

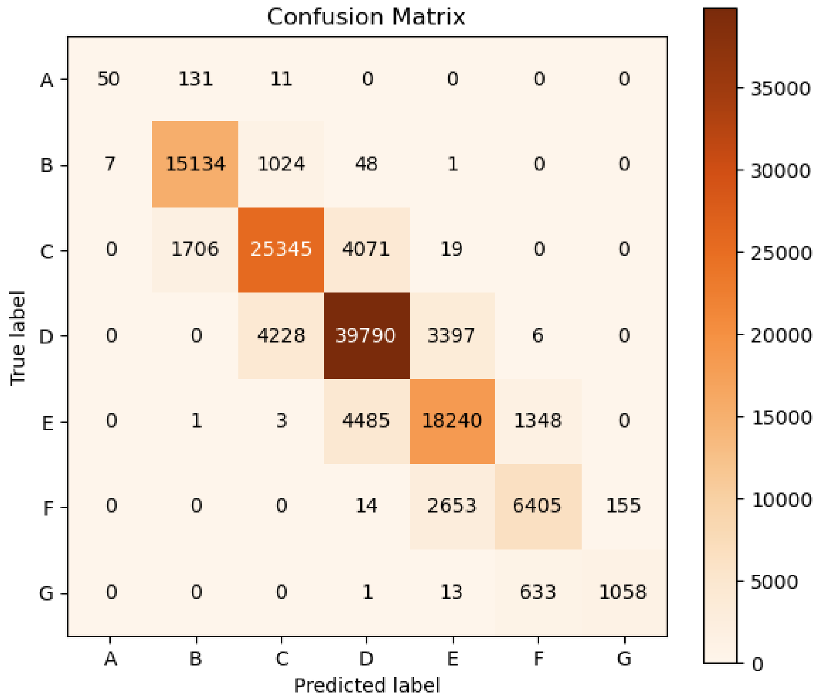

Figure 9 Confusion matrix illustrating strong classification performance of the model in predicting building EPC rating bands (A–G). Most properties are correctly classified, as indicated by the high values along the diagonal, particularly for the dominant mid-range ratings (C, D, and E). Lower accuracy in some outer rating bands is primarily due to the limited number of available samples in those categories

The final objective of this project is to develop a machine-learning model that could predict the energy performance of non-EPC domestic buildings. The proposed methodology utilizes pre-processed EPC data, enhanced property data, actual EPC data, and a developed machine-learning model (Figure 10). To generate the input data for these buildings, an archetypes approach was used to generate 26 feature values for each building. The approach considered the aggregation of building type and year of construction in the primary thorofare area, which was defined as the street where the building was located and its surrounding neighbourhood. However, it should be noted that the archetypes approach may not be representative of all types of buildings, which could introduce bias into the predictions. At the same time, the model could predict 370,827 domestic buildings that did not have EPCs, and approximately 50,879 buildings could not be predicted due to the unavailability of input data.

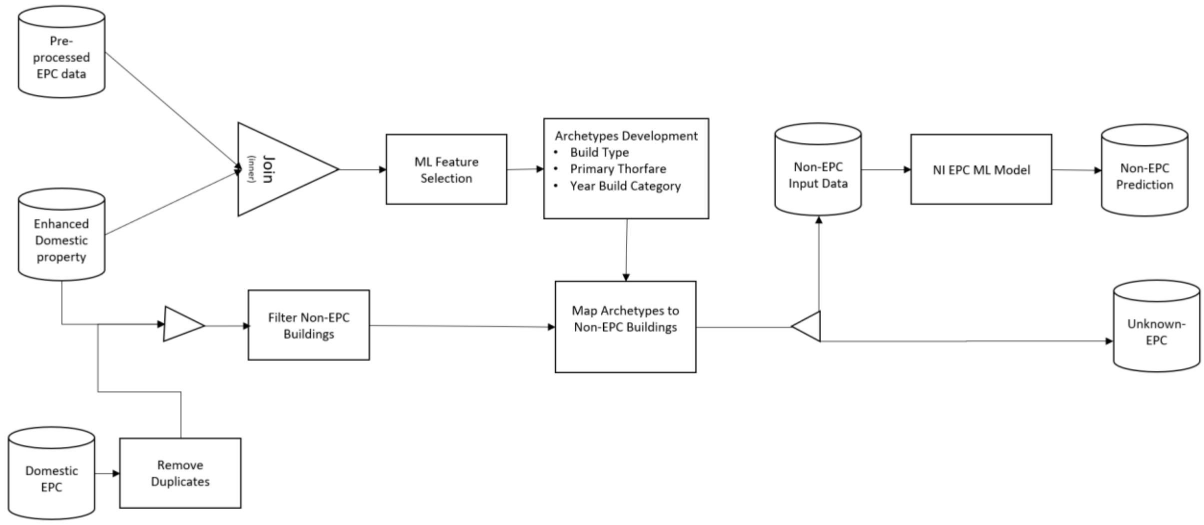

Figure 10 Methodology for data-driven machine learning prediction of Non-EPC domestic buildings

EPC GIS Mapping

After developing the machine learning model to predict EPC for non-EPC buildings, the final step involves GIS mapping of all results for detailed analysis by policymakers. The proposed data, including domestic EPC, non-EPC prediction, unknown EPC, and enhanced property datasets, are used for final mapping (Figure 11). GIS features available for public or internal use are defined and shown in pop-up boxes. Additionally, small area maps are created to display aggregated results of building energy performance. This bottom-up approach aggregates energy performance prediction results from individual buildings to a small area scale, helping to identify clusters of buildings with significant potential for energy savings in specific regions. Geographic information system-based modelling aids stakeholders in identifying priority areas for implementing energy efficiency measures, allowing them to target local communities for retrofit campaigns and enhance the implementation of sustainable energy policy decisions.

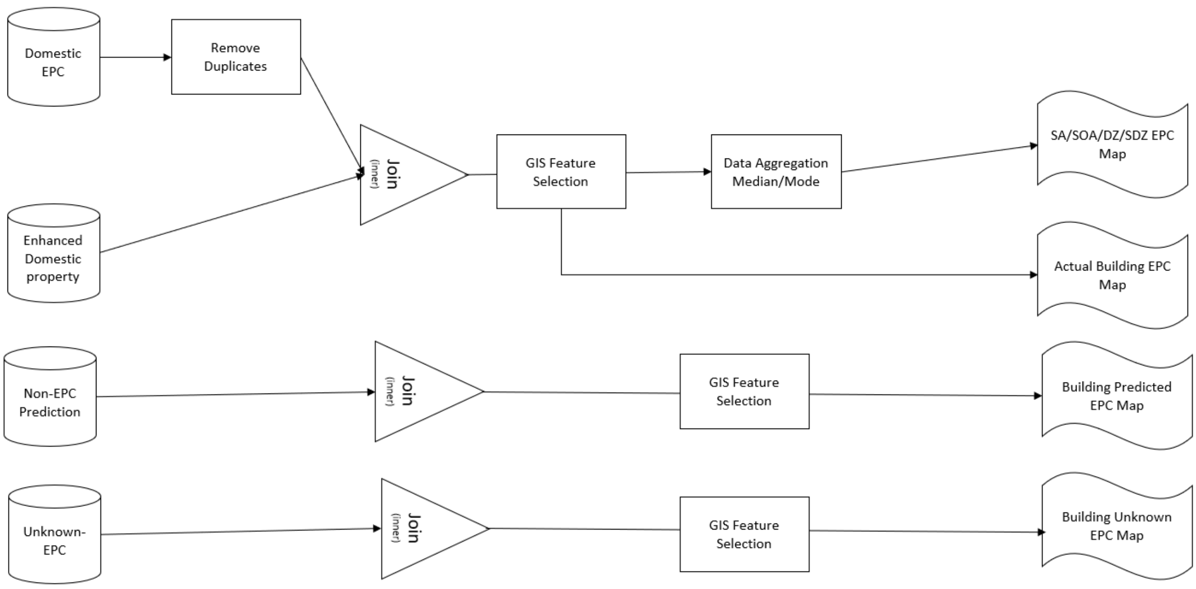

Figure 11 Methodology for GIS mapping of EPC domestic buildings

Figure 12 Domestic Building GIS Mapping of Actual EPC

Figure 13 Domestic Building GIS Mapping of Predicted EPC

Figure 14 Domestic Building GIS Mapping of the combined (Actual, Estimated, and No) EPC

Figure 12, Figure 13, and Figure 14 are GIS mappings of domestic buildings, showing the actual EPC, predicted EPC, and combined EPC results, respectively. These figures help policymakers and other stakeholders understand the energy performance of individual buildings and identify areas that require energy-efficient measures.

Figure 15 is a GIS mapping of small areas with actual EPC domestic buildings, displaying the most frequent energy rating for each specific small area. This figure allows stakeholders to identify areas with poor energy performance and prioritize energy-efficient measures accordingly.

Figure 16 is a similar GIS mapping, but it shows the actual and predicted EPC results for domestic buildings in each small area. By comparing the actual and predicted results, stakeholders can identify areas where the predicted EPC values are significantly different from the actual EPC values, suggesting the need for further investigation and potential improvement in the prediction model.

This mapping process helps with energy planning, and the generated maps can aid stakeholders in analysing and identifying priority areas for implementing energy-efficient strategies. The results can also be used to identify areas where energy policymakers can run targeted community-based events to increase retrofitting activity. The developed multiple-scale map helps decision-makers to identify areas with a large number of energy-inefficient buildings, which can be used to conduct targeted community-based social marketing to increase retrofit rates. The map also identifies clusters of residential buildings with poor energy performance, suggesting areas with poor insulation and heating system performance. By providing a clear picture of the energy performance of buildings, these maps can inform policymakers and aid in the development of effective energy efficiency strategies.

Furthermore, this approach allows energy performance to be interpreted more consistently across different spatial scales, from individual neighbourhoods to wider administrative areas. The comparison between actual-only and combined datasets shows that fuller coverage reduces the risk of misleading patterns caused by missing EPC records and gives a more balanced picture of where lower-performing housing is concentrated. This is especially important for area-based planning, as decisions on retrofit support, funding allocation and community engagement may be less effective if they are based on incomplete evidence. Using the combined maps therefore strengthens the reliability of spatial targeting and improves the value of the analysis for practical policy use.

Overall, GIS mapping provides a comprehensive visualization of the energy performance of domestic buildings at both individual and small area scales. The results can aid policymakers in identifying areas that require energy-efficient measures and developing effective energy policy decisions. By linking the figures together and cohesively presenting them, the report can provide a complete understanding of the energy performance of domestic buildings and inform the development of effective energy-efficient strategies.

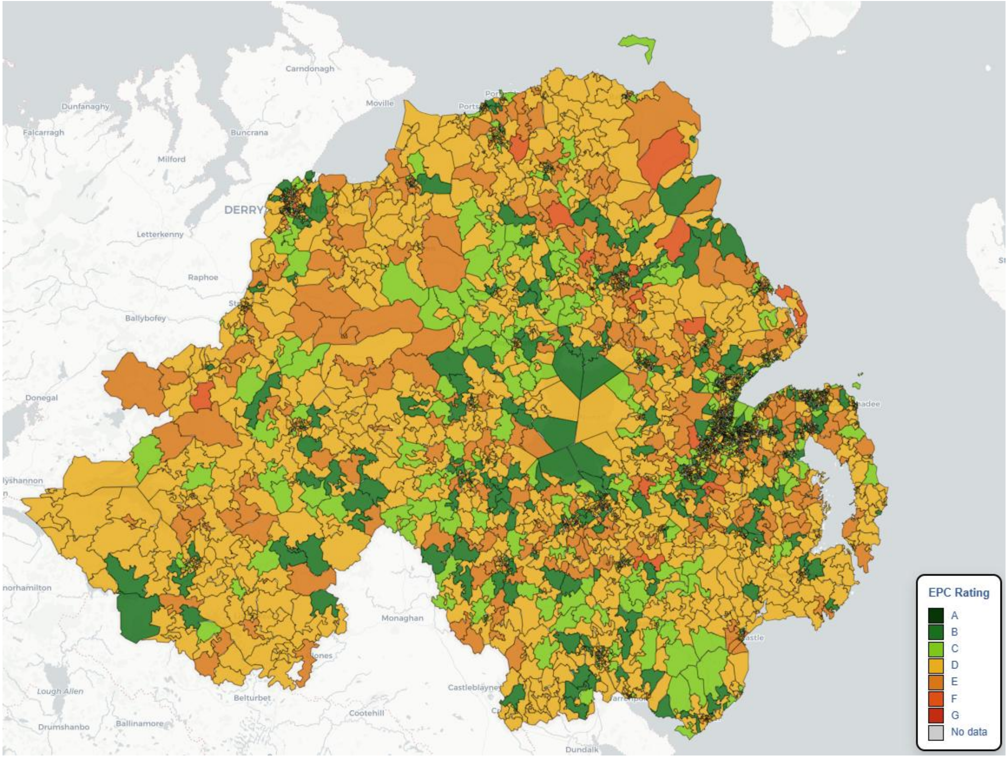

Figure 15 GIS Mapping of Small Area Actual EPC Domestic Buildings and Results Show the Most Frequent Energy Rating in Each Specific Small Area

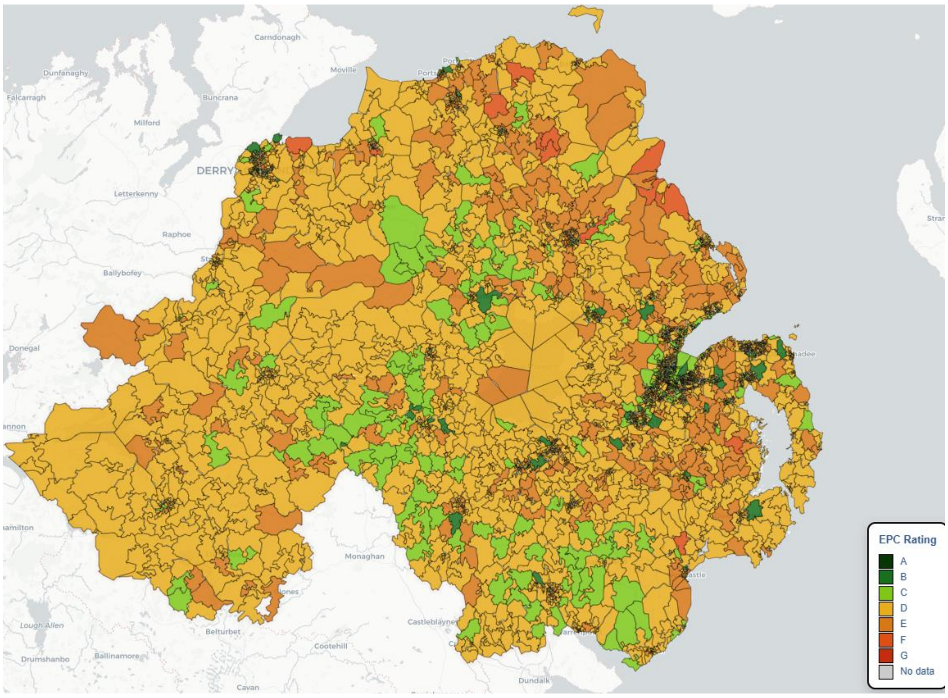

Figure 16 GIS Mapping of Small Area Actual & Predicted EPC domestic buildings and results show most frequent energy rating in each specific small area

4. Conclusion and Future Work

This project presents a data-driven methodology based on a bottom-up approach for mapping residential building energy performance in Northern Ireland. Urban planners and energy policymakers often face significant challenges when implementing sustainable energy analysis and planning at a large scale due to the complexity of the energy system. This research formulates a methodology to effectively integrate various data resources using machine learning algorithms to facilitate energy planning. The proposed method can estimate building energy performance with limited knowledge of building dynamics, and modelling results help develop an energy efficiency footprint of buildings at the urban scale.

In this project, 51% of EPC data are used to predict 43% of the building stock, and only 6% of the domestic building stock was unable to predict due to limited input data. The overall EPC coverage, including actual and predicted buildings, is 94% of the entire domestic building stock of Northern Ireland. These results demonstrate the effectiveness of the data-driven approach coupled with spatial data, which can enhance the quality of existing data and extract meaningful knowledge for decision-making.

Moreover, the bottom-up approach used in this study allows for detailed qualitative and quantitative analysis, enabling the development of energy efficiency footprints at the urban scale. By linking various data resources through machine learning algorithms, this methodology can provide a comprehensive understanding of the energy performance of buildings and aid in developing effective energy-efficient strategies.

Overall, this project provides a valuable resource for policymakers and urban planners seeking to improve the energy performance of buildings. This methodology can facilitate energy planning and enhance the quality of existing data, leading to more informed decision-making and effective energy-efficient strategies using a data-driven approach.

This project has identified a data-driven methodology for mapping residential building energy performance in Northern Ireland. However, several potential areas for future work could enhance the accuracy and effectiveness of the methodology. The following are potential future work topics arising from this project:

- Obtain more detailed EPC data based on Standard Assessment Procedure (SAP) inputs to improve the data-driven model.
- Update the modelling results annually using updated EPC or Census data.
- Improve the quality of census data by making raw format data available.
- Extend GIS modelling by incorporating 3D modelling, which requires building height data for the entire NI building stock.
- Replicate the platform for non-domestic buildings.
- Extend the project to an automated multi-criteria tool.
- Validate or improve results using gas or electric metered data.
- The EPC modelling result could further improve by aligning with dynamic building energy modelling using off-the-shelf building energy simulation software.

Overall, these potential areas for future work could enhance the effectiveness of the data-driven methodology and aid policymakers and urban planners in developing effective energy-efficient strategies.

5. Appendix

There are 442,298 properties in Northern Ireland with actual EPC records, representing 51.0% of the total housing stock. Estimated EPC records account for a further 368,651 properties, or 42.5%. When combined, the dataset covers 810,949 properties, representing 93.5% of the Northern Ireland housing stock, leaving 56,565 properties, or 6.5%, without coverage (Table 5). These results show that actual EPC data alone provides only partial coverage and may not fully represent spatial patterns across all district areas. The inclusion of estimated EPCs substantially improves completeness across LGDs and provides a stronger basis for comparing district-level patterns across Northern Ireland. However, the combined dataset should still be interpreted with caution, as estimated values are less certain than actual EPC records and are therefore more appropriate for area-wide analysis than for property-level assessment.

Table 5 Summary of EPC coverage across the Northern Ireland housing stock, showing the number of properties with actual, estimated and combined EPC records

Category	Properties	Coverage of NI stock
Actual EPC records	442,298	51.0%
Estimated EPC records	368,651	42.5%
Combined EPC records	810,949	93.5%
Uncovered stock	56,565	6.5%

Table 6 EPC Ratings (Actual) by Local Government District

Local Government District	A	B	C	D	E	F	G
Antrim and Newtownabbey	61	4,977	8,212	12,703	6,426	2,175	335
Ards and North Down	45	5,736	10,156	14,655	7,987	3,230	613
Armagh City, Banbridge and Craigavon	36	5,785	9,213	16,843	9,166	3,379	554
Belfast	118	10,022	34,415	34,545	15,476	6,690	1,194
Causeway Coast and Glens	90	3,633	5,835	13,818	7,460	3,119	652
Derry City and Strabane	39	4,388	7,830	13,719	6,211	2,298	402
Fermanagh and Omagh	41	2,190	4,059	9,153	4,481	1,448	223
Lisburn and Castlereagh	147	6,383	8,380	11,611	6,228	2,330	377

Mid and East Antrim	29	3,629	6,707	11,484	6,575	3,071	656
Mid Ulster	40	4,243	5,118	9,497	5,040	1,731	357
Newry, Mourne and Down	63	4,267	6,011	13,129	6,417	2,393	549
TOTAL	**709**	**55,253**	**105,936**	**161,157**	**81,467**	**31,864**	**5,912**

Table 7 EPC Ratings (Estimated) by Local Government District

Local Government District	A	B	C	D	E	F	G
Antrim and Newtownabbey	0	452	4,675	14,025	6,695	1,107	103
Ards and North Down	0	471	5,100	15,685	9,048	1,652	106
Armagh City, Banbridge and Craigavon	0	751	6,349	21,098	11,117	2,488	263
Belfast	0	1,285	15,258	27,264	11,189	1,676	188
Causeway Coast and Glens	0	531	3,744	14,267	8,760	2,756	295
Derry City and Strabane	0	469	4,570	14,753	6,838	1,642	88
Fermanagh and Omagh	0	420	4,797	13,322	6,860	2,203	187
Lisburn and Castlereagh	0	541	3,921	14,478	6,993	1,300	120
Mid and East Antrim	0	411	4,330	12,612	8,614	2,399	315
Mid Ulster	0	631	5,670	13,649	7,523	2,266	257
Newry, Mourne and Down	0	1,008	6,159	18,167	9,652	2,845	243
TOTAL	**0**	**6,970**	**64,573**	**179,320**	**93,289**	**22,334**	**2,165**

Table 8 EPC Ratings (Combined) by Local Government District

Local Government District	A	B	C	D	E	F	G
Antrim and Newtownabbey	61	5,429	12,887	26,728	13,121	3,282	438
Ards and North Down	45	6,207	15,256	30,340	17,035	4,882	719
Armagh City, Banbridge and Craigavon	36	6,536	15,562	37,941	20,283	5,867	817
Belfast	118	11,307	49,673	61,809	26,665	8,366	1,382
Causeway Coast and Glens	90	4,164	9,579	28,085	16,220	5,875	947
Derry City and Strabane	39	4,857	12,400	28,472	13,049	3,940	490
Fermanagh and Omagh	41	2,610	8,856	22,475	11,341	3,651	410
Lisburn and Castlereagh	147	6,924	12,301	26,089	13,221	3,630	497
Mid and East Antrim	29	4,040	11,037	24,096	15,189	5,470	971
Mid Ulster	40	4,874	10,788	23,146	12,563	3,997	614
Newry, Mourne and Down	63	5,275	12,170	31,296	16,069	5,238	792
TOTAL	**709**	**62,223**	**170,509**	**340,477**	**174,756**	**54,198**	**8,077**

Figure 17 and Figure 18 show how data coverage affects interpretation across Northern Ireland's 890 Super Output Areas (SOAs). Figure 17, based only on actual EPC records, gives a partial and potentially misleading picture because areas with fewer observed records may appear less representative of true local energy performance. Figure 18, which combines actual and estimated EPCs, provides much fuller coverage and a more spatially consistent pattern across all SOAs.

This comparison is important for policy analysis. If policymakers rely only on actual EPC data, they may misjudge conditions in areas with weaker coverage, overemphasise clusters where certificates are more common, and overlook broader spatial inequalities. The combined dataset reduces these gaps and gives a stronger basis for comparing areas across Northern Ireland.

The maps also suggest that the overall pattern remains dominated by mid-range EPC bands, but the combined map presents this more clearly and continuously. This makes it better suited to strategic assessment and resource targeting at area level. However, the estimated component should still be interpreted with caution, as it strengthens completeness without fully replacing the reliability of observed EPC records.

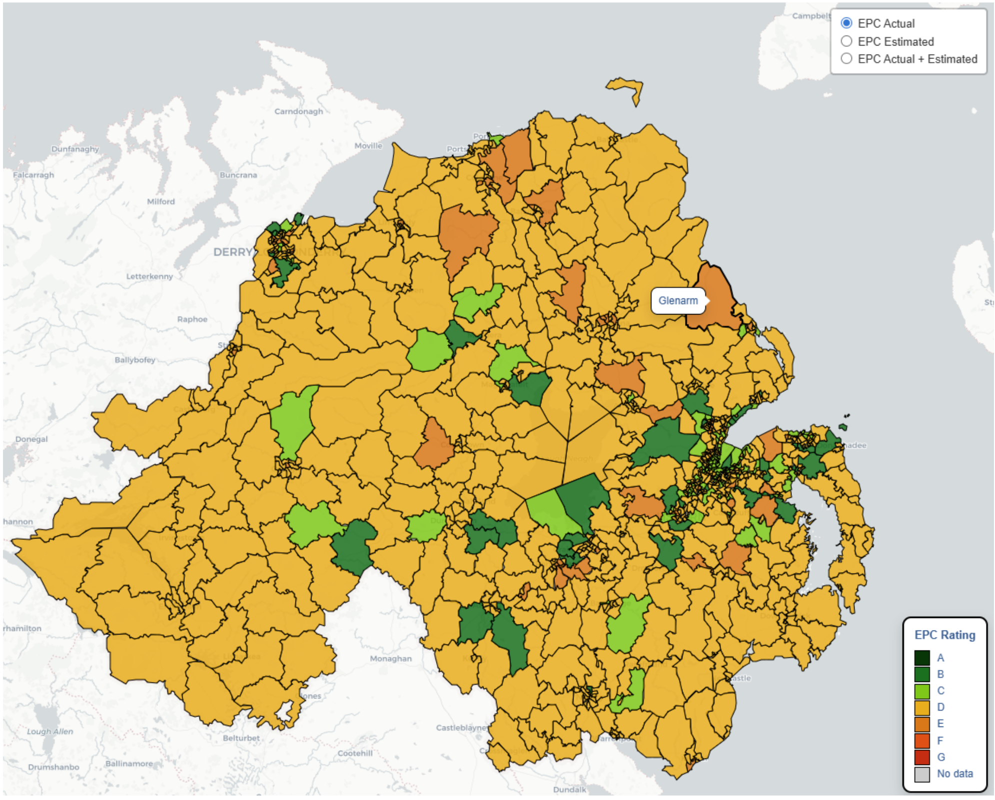

Figure 17 GIS Mapping of Super Output Area Actual EPC Domestic Buildings and Results Show the Most Frequent Energy Rating in Each Specific SOA

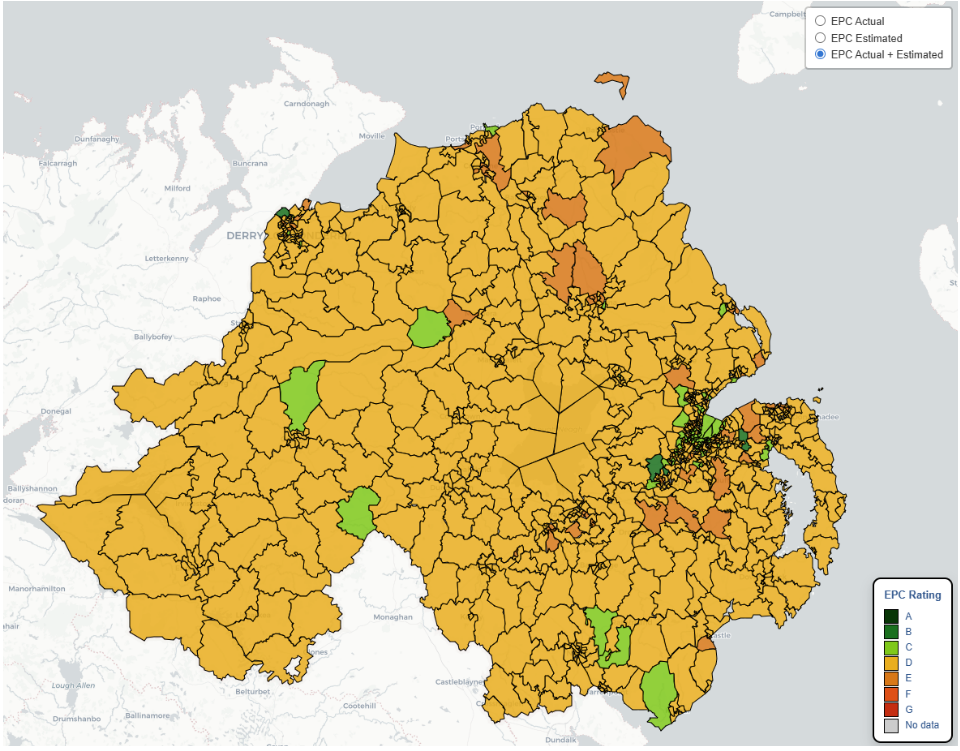

Figure 18 GIS Mapping of Super Output Area Actual & Predicted EPC domestic buildings and results shows most frequent energy rating in each specific SOA

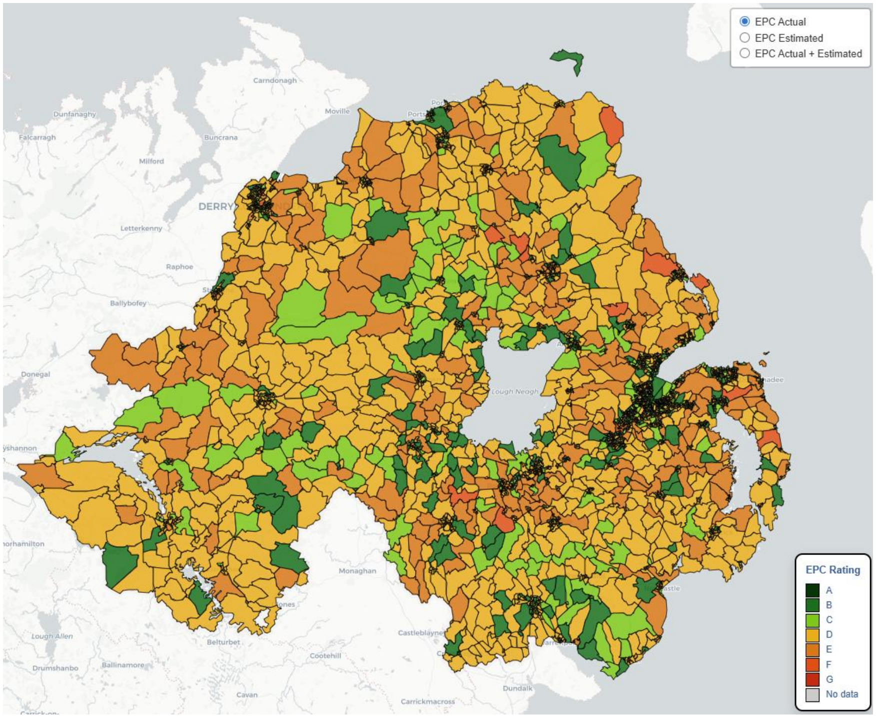

Figure 19 GIS Mapping of Data Zone Area Actual EPC Domestic Buildings and Results Show the Most Frequent Energy Rating in Each Specific DZ

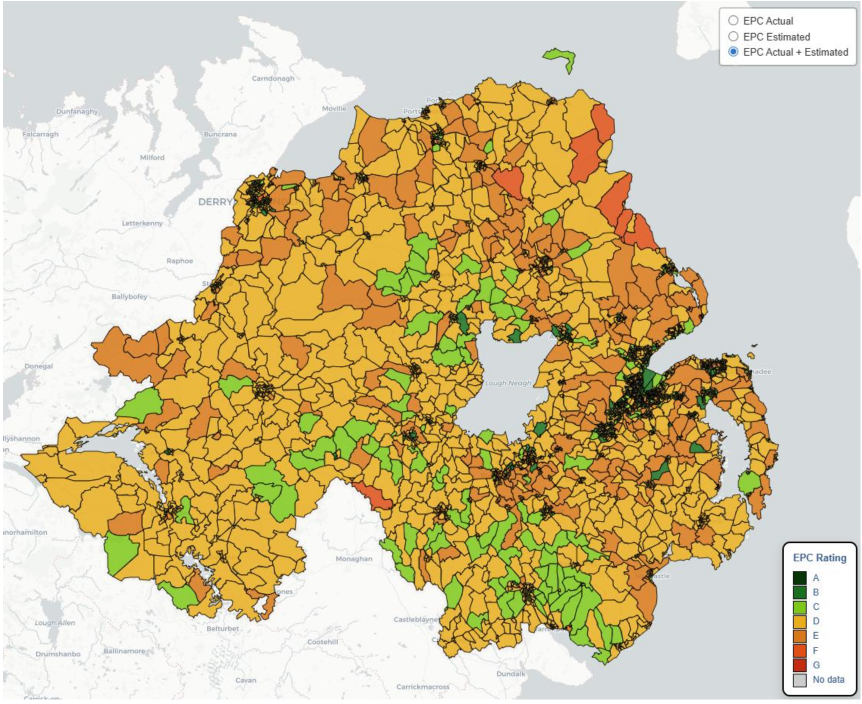

Figure 20 GIS Mapping of Data Zone Area Actual & Predicted EPC domestic buildings and results shows most frequent energy rating in each specific DZ

Figure 19 and Figure 20 show a similar pattern at Northern Ireland Data Zone(DZ) level. Figure 19, based only on actual EPC records, appears to show a greater concentration of higher-performing A and B rated buildings in some areas. However, this pattern is influenced by incomplete coverage and may give a misleading impression of overall energy performance. Figure 20, which uses the combined actual and estimated dataset, provides fuller coverage and shows that these higher ratings are less widespread when all areas are represented more consistently.

This comparison is important for policy analysis because reliance on actual EPC data alone could overstate the presence of better-performing buildings and distort understanding of local conditions. The combined dataset gives a more balanced and representative picture across Data Zones, making it more suitable for identifying broad spatial patterns and supporting area-based decision-making. However, the estimated component should still be interpreted cautiously, as it improves completeness but does not have the same certainty as observed EPC records.

At Super Data Zone level, Figures 21 and 22, show that the actual-only map suggests a stronger presence of higher-performing B ratings in several areas. However, this pattern is likely influenced by where observed EPC records are available, rather than reflecting complete underlying housing conditions. As a result, some places may appear to perform better simply because coverage is stronger there.

When actual and estimated EPCs are combined, the pattern becomes more even and representative across Northern Ireland. Higher-rated areas are still visible, but they are less extensive, while mid-range ratings become more dominant across Super Data Zones. This indicates that relying only on actual EPC data could give an overly positive view in some neighbourhoods, whereas the combined dataset provides a more robust basis for local comparison and intervention targeting.

Overall, the mapping results support the wider aim of this project, which is to develop a data-driven, bottom-up methodology for assessing residential building energy performance across Northern Ireland. By integrating multiple data sources and combining observed EPC records with machine learning-based estimates, the study provides a more complete urban-scale energy efficiency footprint than would be possible from actual EPC data alone. The comparison across SOAs, Data Zones and Super Data Zones shows that fuller coverage is essential for reliable spatial analysis, as relying only on observed records may distort local patterns and misinform policy decisions. In this context, the combined dataset offers a stronger evidence base for large-scale energy planning, while still requiring careful interpretation because estimated values do not fully replace the certainty of measured EPC records.

This is particularly relevant for urban planners and energy policymakers, who need consistent and spatially complete information to identify priority areas, compare local energy performance, and design targeted interventions. The methodology developed in this project helps address the limitations of fragmented building data by estimating energy performance where direct EPC information is missing. As a result, it improves the practical value of the analysis for strategic decision-making and demonstrates how machine learning can support sustainable energy planning at regional and neighbourhood scales.

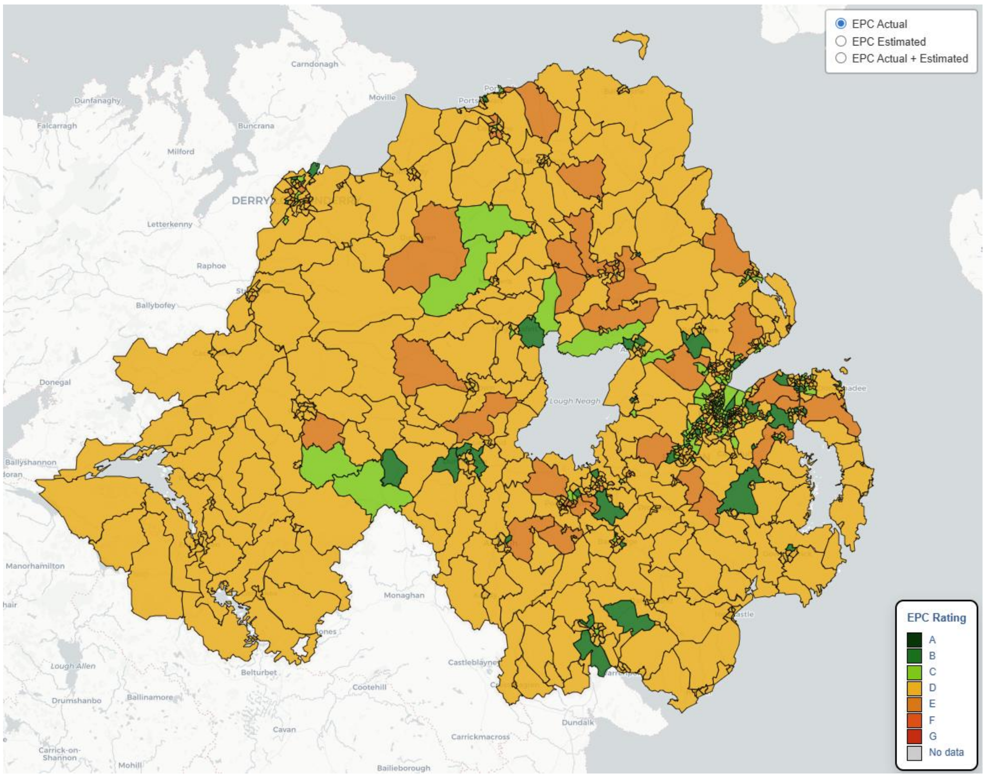

Figure 21 GIS Mapping of Super Data Zone Area Actual EPC Domestic Buildings and Results Show the Most Frequent Energy Rating in Each Specific SDZ

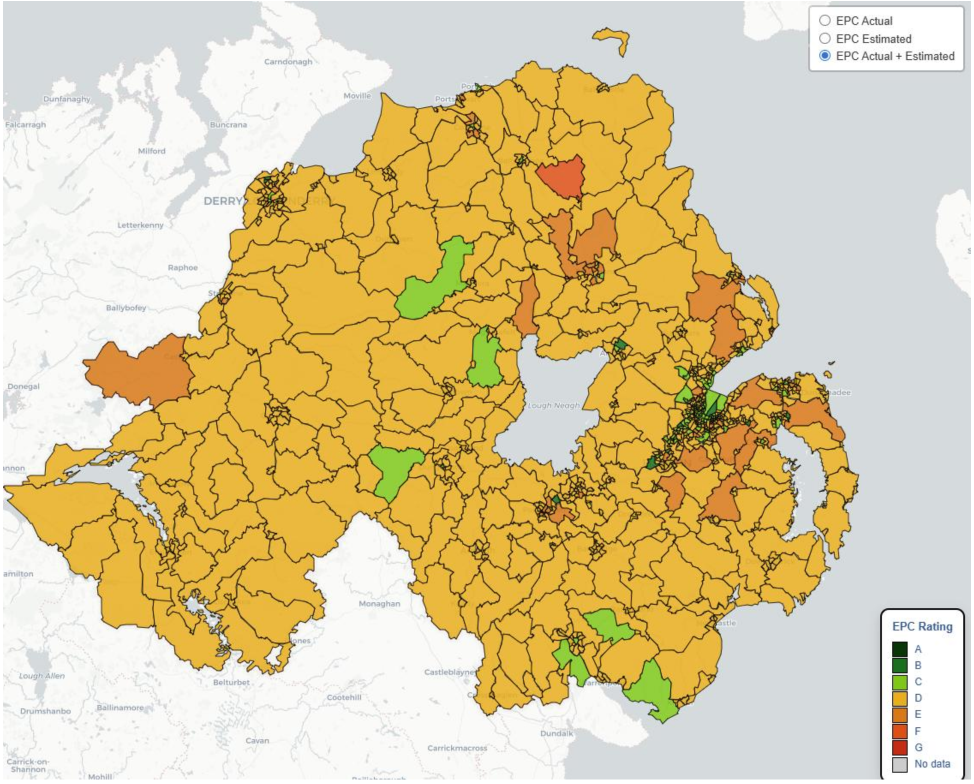

Figure 22 GIS Mapping of Super Data Zone Area Actual & Predicted EPC domestic buildings and results shows most frequent energy rating in each specific SDZ