Engineering a green built environment: how buildings can support a sustainable future

Abbaspour, Atefeh, Bahadori-Jahromi, Ali ORCID logoORCID: https://orcid.org/0000-0003-0405-7146, Keyhani, Maryam, Seraj, Hamidreza, Maigari, Mukhtar Bello, Fu, Charlie ORCID logoORCID: https://orcid.org/0000-0002-2019-5445, Baladimou, Efcharis, KC, Prapooja, Sudhakaran, Seeja and Sakikhales, Mohammad (2026) Engineering a green built environment: how buildings can support a sustainable future. Cambridge Scholars Publishing. ISBN 978-1-0364-7525-3

CHAPTER 3

INVESTIGATING THE UK EPC DATABASE: CONCEPTUAL FRAMEWORK, LIMITATIONS, AND DATA-DRIVEN APPLICATIONS

HAMIDREZA SERAJ AND
ALI BAHADORI-JAHROMI

Abstract: Energy Performance Certificates (EPC) play a critical role in assessing residential buildings' energy efficiency, energy policy decision-making, and planning retrofit strategies. In the UK, EPC ratings are generated using the Standard Assessment Procedure (SAP), a methodology based on predefined assumptions and quasi steady-state equations. This chapter investigates the conceptual framework behind the UK EPC dataset, exploring how key outputs, such as EPC ratings, CO_2 emissions, and energy consumption, are derived; and assessing the limitations inherent in the SAP based approach. Moreover, a detailed analysis of input/output correlations within the dataset reveals inconsistencies, missing data issues, and class imbalances, which impact its reliability for data-driven applications. To address these challenges, data preprocessing techniques, including synthetic oversampling, are explored to enhance dataset applicability for benchmarking and machine learning based energy modelling. However, even with advanced data science techniques, the inherent limitations of SAP based calculations must be acknowledged. Since the EPC dataset is derived from simplified steady-state models, predictions from data-driven models trained on this data are unlikely to achieve high accuracy in real-world energy performance assessments.

1. Introduction

The assessment of buildings' energy performance is a critical component in achieving global sustainability goals, particularly in reducing carbon emissions and improving energy efficiency. So, Various modelling approaches have been developed to assess energy consumption and environmental impact of the building sector, including physics-based simulations and data-driven machine learning models. These methodologies play a crucial role in regulatory compliance, retrofit, and energy policy planning.

From a broader point of view, energy modelling techniques can be classified into white-box, grey-box, and black-box approaches. White-box modelling approaches, such as dynamic thermal simulation, rely on thermodynamic and heat transfer principles to simulate a building's energy flow. While they provide high interpretability, their extensive data requirements and computational complexity limit their scalability for large-scale applications. Conversely, black-box models leverage historical data and machine learning algorithms to predict energy consumption trends with reduced computational overhead. However, they lack the interpretability of physics-based models which makes them less suitable for detailed performance analysis.

In this context, Yu et al. (2022) conducted a comprehensive review of the methodologies used in white-box, black-box, and grey-box approaches for predicting buildings' energy performance. They also analysed sources of uncertainty related to different building energy prediction approaches in terms of consumption behaviour, building, and weather factors. In addition, Shahcheraghian et al. (2024) focused on the development of tools for building energy modelling, including web-based tools, black-box tools, and white-box tools. They also analysed the scalability of these tools, ranging from single-family homes to districts and cities which highlights the necessity of choosing the appropriate modelling approach based on case study building complexities and retrofit objectives to support decision-making. Furthermore, Pachano et al. (2025) investigated the accuracy of black-box Long Short-Term Memory (LSTM) models using on-site measurement data. The developed LSTM model utilise climate data, building operation parameters, and other HVAC subsystem variables, which emphasise the importance of data quality and quantity for training black-box models.

In the UK, the Standard Assessment Procedure (SAP) serves as the government's primary methodology for evaluating residential building energy performance. It forms the Energy Performance Certificate (EPC) framework, which is widely used for regulatory compliance, policy analysis, and benchmarking energy efficiency. However, its methodology relies on simplified, quasi steady-state calculations and predefined assumptions, which leads to discrepancies between predicted and actual energy use. Furthermore, the dataset extracted from this framework, suffers from issues such as data imbalance and limitations in capturing dynamic building performance.

A study conducted by Hardy et al. (2019) analysed the errors in the EPC database. They categorised potential errors and identified several patterns of data anomaly during EPC lodgement including anomalies based on an analysis of building design and elements. The results of this study highlighted that at least 27% of all EPCs lodged have a discrepancy and the true error percentage for the EPC record is in the range 36–62%. Additionally, Few et al. (2023) compared the differences between SAP-modelled energy consumption and on-site measurements in more than 1,300 case study buildings. The results of this study revealed acceptable performance for EPC bands A and B; however, a significant gap was observed for other EPC ratings, with an 8% error for case studies with an EPC rating of C and a 48% error for bands F and G.

In response, this chapter investigates the conceptual framework behind the development of the UK EPC database, exploring how key energy performance outputs, such as energy ratings, CO_2 emissions, and energy consumption, are calculated. By examining the logic and assumptions embedded in SAP methodology and consequently related tools, this study identifies potential limitations in the dataset, particularly regarding quasi-steady-state calculations, predefined operational conditions, and the exclusion of dynamic energy interactions.

Furthermore, to enhance the usability of the EPC dataset for benchmarking and data-driven applications, this study explores methodologies for addressing data inconsistencies. Techniques such as data imputation, feature engineering, and synthetic oversampling (e.g., SMOTE) are applied to mitigate biases and improve dataset robustness. By refining data preprocessing strategies, this research aims to enhance the dataset's applicability for evaluating retrofit strategies and developing machine learning models for energy performance prediction.

2. Different Approaches for Assessing Buildings' Energy Consumption

There are several approaches to categorise building energy modelling methodologies; one of the fundamental classification frameworks is dividing them into black box and white box models (Olu-Ajayi et al. 2023). White-box models, also known as physics-based models, rely on fundamental physical and engineering principles to model buildings' energy performance. These models provide high interpretability, as the results can be directly related to physical attributes, which makes them valuable for detailed energy performance analysis and system optimisation.

However, they require extensive input data, time consuming calibration, and significant computational resources, which limits their scalability for large-scale applications (Krstić and Teni 2017; Klanatsky, Veynandt, and Heschl 2023). Several well-established tools, such as EnergyPlus, DesignBuilder, and IESVE, have been developed based on this approach, enabling detailed simulations that account for factors like material properties, HVAC systems, occupancy behaviour, and climatic conditions.

On the other hand, black-box models have a data-driven, empirical approach, using statistical methods and machine learning algorithms such as artificial neural networks, XGBoost, and random forests. These models leverage historical data to predict buildings' energy consumption and performance trends. Their main advantage is their relatively fast development time and ability to generate accurate predictions without requiring a deep understanding of the underlying physical processes. However, they lack interpretability, meaning that results cannot be directly explained in physical terms. Moreover, they require extensive and accurate training data, which may not always be readily available.

In terms of calculation techniques, Chartered Institution of Building Services Engineers (CIBSE) in the UK has provided a wide range of methodologies based on a white-box approach that can be applied to estimate buildings' energy demand including quasi steady-state and dynamic thermal simulation (Chartered Institution of Building Services Engineers 2015). One of the simplest approaches is the benchmark method, which estimates energy demand by multiplying a building's floor area by predefined energy intensity values (kWh/m^2) for different building types. This approach can be refined by adjusting for projected occupancy hours or normalising heating benchmarks. Another commonly used technique is the "degree-day" method, which assumes that heating load is linearly related to external air temperature. When the outdoor temperature falls below a defined baseline, heating demand is estimated based on the accumulated degree-days over a given period. More details on these methodologies can be found in (Chartered Institution of Building Services Engineers 2022; 2015).

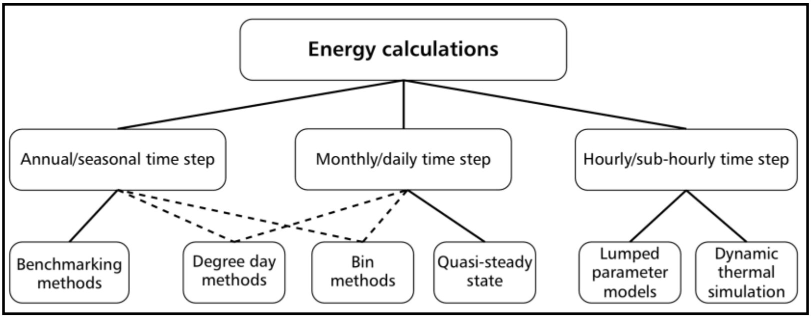


Figure 1. Building performance modelling approaches (Chartered Institution of Building Services Engineers 2015)

While the classification of modelling approaches into white-box and black-box categories provides a fundamental understanding, alternative perspectives also exist for this purpose. Some frameworks classify models as top-down or bottom-up, each with distinct assumptions and applications (Li et al. 2017). However, such categorisations fall outside the scope of this chapter. This research focuses on developing a conceptual framework for a tool aimed to building energy retrofit planning, emphasising scalability and accuracy rather than energy modelling categorisation methodologies.

2.1 Quasi steady-state and dynamic thermal simulations

In general, almost all of well-established modelling tools in the UK for both retrofit planning and code compliance purposes fall into two categories: quasi steady-state based, and dynamic thermal simulation tools. Quasi steady-state models use a simplified approach to simulating HVAC systems based on typical assumptions and rules of thumb on the efficiency of equipment. The error margin on these results would typically be quite large, which reflects the simplifications made to the system types. As an example of the application of the quasi-steady state models, the procedure is currently applied in both the Simplified Building Energy Method (SBEM) and the Standard Assessment Procedure (SAP) as a route to demonstrating compliance with Part L of the Building Regulations for England and Wales and Section 6 of the Building Regulations. These use a simplified approach to simulating HVAC systems based on typical assumptions and rules of thumb on the efficiency of equipment. The error margin on these results would typically be quite large, which reflects the simplifications made to the system types.

For instance, the energy needed to heat (Q_H) or cool (Q_C) a space is given by equations 1 and 2 (Chartered Institution of Building Services Engineers 2021):

$$Q_H = Q_{ht} - \eta_{H,gn}\, Q_{gn} \tag{1}$$

$$Q_C = Q_{gn} - \eta_{C,ls}\, Q_{ht} \tag{2}$$

where Q_H is the energy needed to heat a space, Q_{ht} is the total heat transfer through the fabric and ventilation, Q_{gn} is internal and solar heat gains, $\eta_{H,gn}$ is the heat gain utilisation factor during times of heating, Q_C is the energy needed to cool a space, and $\eta_{C,ls}$ is the heat loss utilisation factor during cooling.

From a quantitative point of view, heat demand is determined by how much heat must be added to offset the heat lost through the building fabric and ventilation after useful heat gains have been subtracted. For cooling, the amount of heat to be removed equals that added by the heat gains, minus the heat lost through the fabric and ventilation.

Furthermore, the total heat transfer Q_{ht} in this approach is given by equation 3; where Q_{tr} is heat transmission through the fabric and Q_{ve} is transmission by ventilation. Similarly, heat gains in a zone may be calculated as equation 4, where Q_{int} is internal heat gains and Q_{sol} is solar heat gain.

$$Q_{ht} = Q_{tr} + Q_{ve} \tag{3}$$

$$Q_{gn} = Q_{int} + Q_{sol} \tag{4}$$

Also, the heat transmission through the fabric Q_{tr} can be obtained using equation 5; where H_{tr} is the overall heat transfer coefficient (w.k^{-1}), θ_{int} is the average internal temperature over the period of calculation, θ_e is the external temperature and t is the duration of the time step.

$$Q_{tr} = H_{tr}(\theta_{int} - \theta_e)t \tag{5}$$

On the other hand, dynamic simulation is a process producing a detailed model of the building, including its fabric, HVAC systems and controls, to more accurately reflect the modelled operational energy usage of the building under expected conditions, hour by hour over a year.

These hourly and sub-hourly dynamic modelling techniques solve the one dimensional conduction of heat through the building fabric. For instance, figure 2 illustrates an incremental volume in a homogeneous slab, which could represent a typical wall or floor element.

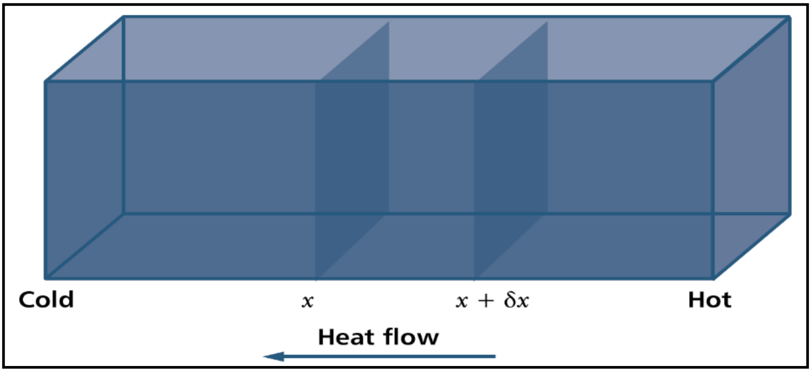


Figure 2. Incremental volume in a slab (Chartered Institution of Building Services Engineers 2015)

It is considered that heat can flow through the material, so the following relationships can be developed. At point $(x + \delta x)$ heat flow rate in, out, and change in heat flow rate over elemental volume can be obtained using equations 6, 7, and 8 respectively. Where λ is the thermal conductivity of the material (W/m·K), A is the face area of the element, T is temperature and x is material length (Chartered Institution of Building Services Engineers 2015).

$$Heat\ flow\ rate\ in = \lambda A \left.\frac{dT}{dx}\right]_{x+\delta x} \tag{6}$$

$$Heat\ flow\ rate\ out = -\lambda A \left.\frac{dT}{dx}\right]_{x} \tag{7}$$

$$Change\ in\ heat\ flow\ rate = \lambda A \frac{d^2T}{dx^2}\delta x \tag{8}$$

Comparing these approaches shows that the main limitation of quasi steady-state models, compared with dynamic thermal simulation, is their inability to represent non-linear behaviour and thermal storage and release effects on a sub-monthly basis. For example, night-purge ventilation, ventilation coupled to structure, demand-controlled ventilation and adaptive set points are all awkward to model. It is therefore recommended that quasi steady-state models only be applied to simply controlled buildings where these features are not present (Chartered Institution of Building Services Engineers 2021).

3. Foundations for Residential Energy Efficiency Assessment in the UK

The Energy Performance Certificate (EPC) database is the most comprehensive and widely utilised dataset for assessing the energy and environmental performance of residential buildings in the UK. It is published by the Department for Levelling Up, Housing and Communities and serves as the primary resource for researchers, policymakers, and industry professionals (Bolton 2024). This database includes every lodged EPC, along with key estimates such as annual energy consumption, CO_2 emissions, and energy efficiency ratings across the UK.

Due to its nationwide coverage, the EPC database has been extensively used in data-driven researches, particularly in training machine learning models to predict energy performance and environmental metrics of buildings in the UK. Researchers utilise this dataset for energy retrofit planning, policy analysis, and benchmarking energy efficiency improvements. Given its importance, the EPC database is considered the most valuable resource for evaluating, monitoring, and improving the energy performance of the UK's residential building stock.

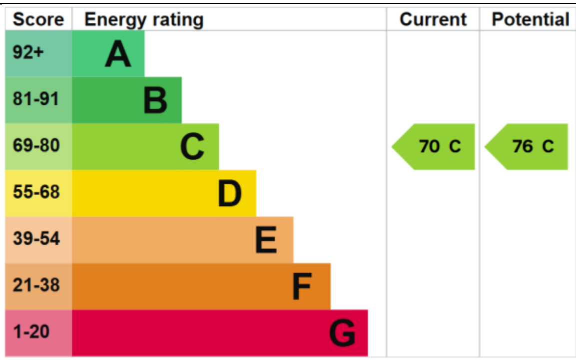
The EPC dataset contains detailed numerical and categorical information on building characteristics, including thermal properties of the building envelope, heating and cooling systems, ventilation, lighting, and renewable energy generation. The data is collected by accredited energy assessors following the SAP or Reduced Data SAP (RdSAP), the UK government's methodology for evaluating energy performance in residential buildings.

SAP is the UK Government's official methodology for assessing the energy performance of residential buildings. First introduced in 1993, SAP calculations are based on quasi steady-state equations and provide a standardised, simplified approach to evaluating a dwelling's energy efficiency (Building research establishment 2012). SAP is primarily designed to ensure compliance with building regulations rather than to serve as a detailed design tool for analysing energy performance. It forms part of the National Calculation Methodology (NCM), which defines the assessment framework for both residential and non-residential buildings (SAP is applied to dwellings, while Simplified Building Energy Model and Dynamic Simulation Models are used for non-domestic buildings).

In general, SAP calculations perform three key functions: 1- Determining the SAP rating, which quantifies a building's energy-related running costs on a scale from 1 to 100 (where 100 represents zero energy cost). A higher SAP rating reflects lower fuel costs and reduced CO_2 emissions. 2- Demonstrating compliance with Part L of the UK Building Regulations, which sets energy efficiency requirements for new buildings and major renovations. 3- Generating EPCs, which provide standardised energy efficiency information to homeowners, tenants, and policymakers. Table 1 shows a brief overview of outputs extracted from a SAP assessment report.

SAP assessments are conducted using either SAP or RdSAP (Reduced Data SAP) approved calculation tools. RdSAP is a simplified version of SAP, used primarily for assessing existing dwellings, where detailed construction data may be unavailable. Both methodologies follow predefined assumptions, which makes them suitable for benchmarking and compliance purposes but less reliable for predicting actual energy consumption.

Table 1. Summary of SAP assessment outputs for a case study building

<table>
<tr><td>Energy rating and score</td><td>
</td></tr>
</table>

Primary energy use	**Primary energy use** The primary energy use for this property per year is 201 kilowatt hours per square metre (kWh/m2).
Impact on the environment	**Carbon emissions** **An average household produces** 6 tonnes of CO2 **This property produces** 2.2 tonnes of CO2 **This property's potential production** 1.6 tonnes of CO2

One of SAP's key limitations is that it does not account for variations in occupant behaviour or adaptive control strategies. It assumes that all homes are occupied under standardised conditions, including:

- A typical occupancy rate derived from floor area.
- Heating schedules, where all rooms are maintained at a set 21°C in living areas and 18°C elsewhere (including bedrooms).
- A fixed standard for hot water usage and no consideration of user adjustments such as changing thermostat settings or varying heating patterns.
- It assumes that buildings are located in middle England, which ignores regional climate variations that can significantly impact actual energy consumption.

Another critical aspect of SAP methodology is that it measures the economic performance of a dwelling rather than its direct energy efficiency. The SAP rating is heavily influenced by fuel type, meaning that a home's SAP value can increase simply by switching to a cheaper fuel source, even if it is more carbon-intensive. For example, a dwelling could achieve a better SAP score by using coal instead of a lower-carbon fuel like wood or biodiesel. This approach makes SAP particularly useful for fuel poverty reduction policies but less effective for strategies aimed at reducing overall energy demand or carbon emissions. Consequently, SAP ratings can fluctuate between successive SAP versions due to changes in fuel prices rather than actual improvements in building performance.

Despite these limitations, SAP remains the backbone of the government policy for assessing the energy performance of UK homes and fulfilling the European Energy Performance of Buildings Directive (EPBD). It is widely used for regulatory compliance, policy analysis, and large-scale energy efficiency assessments. However, its assumptions make it unsuitable for detailed energy demand forecasting or assessing the impact of energysaving measures. Therefore, while SAP plays a crucial role in standardised energy assessments, additional modelling techniques such as parametric simulation or data-driven machine learning approaches are often required for more accurate and flexible energy performance predictions.

4. Developed tools for SAP assessments

In order to conduct SAP assessments in residential properties, accredited energy assessors utilise SAP (or RdSAP) based tools such as "Elmhurst RdSAP Go". These tools can produce EPCs and energy reports based on the government's national calculation methodology for residential buildings and generate outputs as shown in table 1. For instance, figure 3 shows some snapshots of user interface provided in these tools by the "Elmhurst Energy".

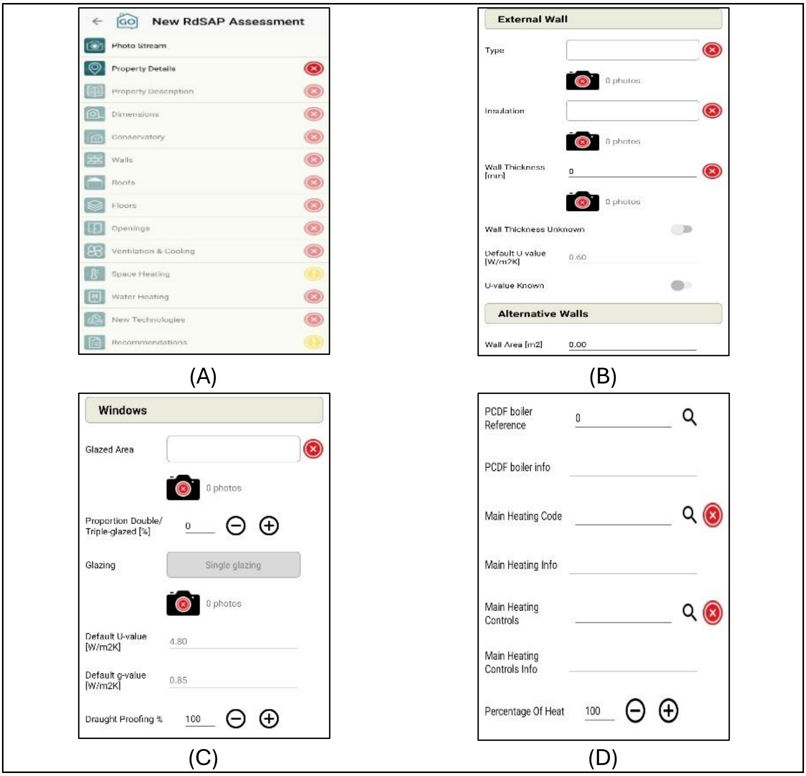


Figure 3. user interface developed for Elmhurst RdSAP Go; (A) overview (B) external wall description (C) glazing system (D) boiler description

These tools require comprehensive building information, including details about the building envelope, heating, ventilation, and air conditioning (HVAC) systems, year of construction, and other relevant parameters. The accuracy of the assessment relies on the completeness and precision of the input data provided by the assessor.

The outputs generated by SAP-based tools serve as the foundation for the UK EPC dataset, which, as previously mentioned, is the most widely used database for policymaking and machine learning model training in the UK. This dataset plays a crucial role in energy policy decision making, guiding retrofit strategies, and supporting research in the built environment.

However, the accuracy of the results produced by these tools is limited, as they are derived using quasi steady-state equations, which assume static conditions and simplified energy balances. These limitations can lead to discrepancies between predicted and actual energy performance, particularly in cases where dynamic factors such as occupancy behaviour, thermal inertia, and varying weather conditions significantly influence energy consumption.

5. Assessing the UK EPC Dataset: Applicability for Benchmarking and Predictive Modelling

After each property is assessed by an accredited EPC assessor and the building features are input into SAP calculation tools, the resulting data is systematically collected in a centralised database. This process has led to the formation of the most extensive dataset on residential buildings in the UK. Given that the advantages and limitations of the SAP methodology have already been discussed, in the following sections the dataset will be thoroughly investigated to identify potential discrepancies. Additionally, strategies to address these issues will be explored to enhance data quality and ensure its robustness for direct use in retrofit decision-making and its effectiveness as a training dataset for black-box models, including machine learning-based predictive frameworks.

To begin with, lodged EPC for more than 40 thousands case study buildings in south west of London between 2021 and 2023 has been extracted from “Energy Performance of Buildings Data: England and Wales” website. The dataset includes following information about each property. Some of the most important variables are presented in table 2 and are divided into input variables and SAP calculation outputs (Department for Levelling Up 2023).

Table 2. description of input and output variables in the UK EPC database

Input variables	
Variable	**Description**
Property type	Describes the type of property; This is the type differentiator for dwellings. (House, Bungalow, Flat, Maisonette)
Built form	The building type of the Property; Together with the Property Type, the Build Form produces a structured description of the property.
Total floor area	The total useful floor area is the total of all enclosed spaces measured to the internal face of the external walls
Floor level	Floor level relative to the lowest level of the property (0 for ground floor)
Glazed type	The type of glazing. From British Fenestration Rating Council or manufacturer declaration, one of; single; secondary; double or triple glazing.
Multi glaze proportion	The estimated banded range of the total glazed area of the Property that is multiple glazed. (0-100 %)
Glazed area	Ranged estimate of the total glazed area of the Habitable Area. (normal, more than typical, less than typical)
Number habitable rooms	Habitable rooms include any living room, sitting room, dining room, bedroom, study and similar
Number heated rooms	The number of heated rooms in the property if more than half of the habitable rooms are not heated
Low energy lighting	The percentage of low energy lighting present in the property as a percentage of the total fixed lights in the property. (0-100%)
Hot water description	Overall description of the hot water system in the property
Floor description	Overall description of the flooring in the property
Walls description	Overall description of the walls material in the property
Roof description	Overall description of the roofing in the property
Main heat description	Overall description of the heating system in the property
Lighting description	Total number of fixed lighting outlets and total number of low-energy fixed lighting outlets
Main fuel	The type of fuel used to power the central heating e.g. Gas, Electricity
Floor height	Average height of the storey in metres

Mechanical ventilation	Identifies the type of mechanical ventilation the property has. (natural, mechanical- supply only, supply and extract)
Solar water heating flag	Indicates whether the hot water in the Property is solar powered
Photo supply	Percentage of photovoltaic area as a percentage of total roof area
Calculation outputs	
Current energy rating	Current energy rating converted into a linear 'A to G' rating (where A is the most energy efficient and G is the least energy efficient)
Potential energy rating	Estimated potential energy rating converted into a linear 'A to G' rating
Current energy efficiency	energy required for space heating, water heating and lighting [in kWh/year] multiplied by fuel costs. (£/m²/year where cost is derived from kWh)
Potential energy efficiency	The potential energy efficiency rating of the property. (0-100)
Current environment impact	A measure of the property's current impact on the environment in terms of carbon dioxide (CO_2) emissions. The higher the rating the lower the CO_2 emissions. (CO_2 emissions in tonnes / year) (0-100)
Potential environment impact	A measure of the property's potential impact on the environment in terms of carbon dioxide (CO_2) emissions after improvements have been carried out. (0-100)
Current energy consumption	Current estimated total energy consumption for the property in a 12 month period (kWh/m2). Displayed on EPC as the current primary energy use per square metre of floor area
Potential energy consumption	Estimated potential total energy consumption for the Property in a 12 month period. Value is Kilowatt Hours per Square Metre (kWh/m²)
Current CO_2 emissions	CO_2 emissions per year in tonnes/year
lighting/heating/hot water cost	Current estimated annual energy costs for lighting/heating/hot water in the property

5.1 Missing Input Values

One of the drawbacks of the UK EPC database is the substantial number of missing values in input parameters, raising concerns about the accuracy and reliability of the SAP-based tool in processing and computing the results. In this regard, Figure 4 illustrates the percentage of missing values for each input parameter. The figure indicates that some input parameters, such as "Mechanical ventilation," "Glazed area," and "Glazed type," have more than 15% missing values. However, it is important to note that missing data in parameters like "Construction age band" and "Main fuel," which significantly affect the energy performance of the system, may reflect potential inconsistencies in data collection. These inconsistencies could impact the reliability of predictive modelling and analysis of SAP-based tools.

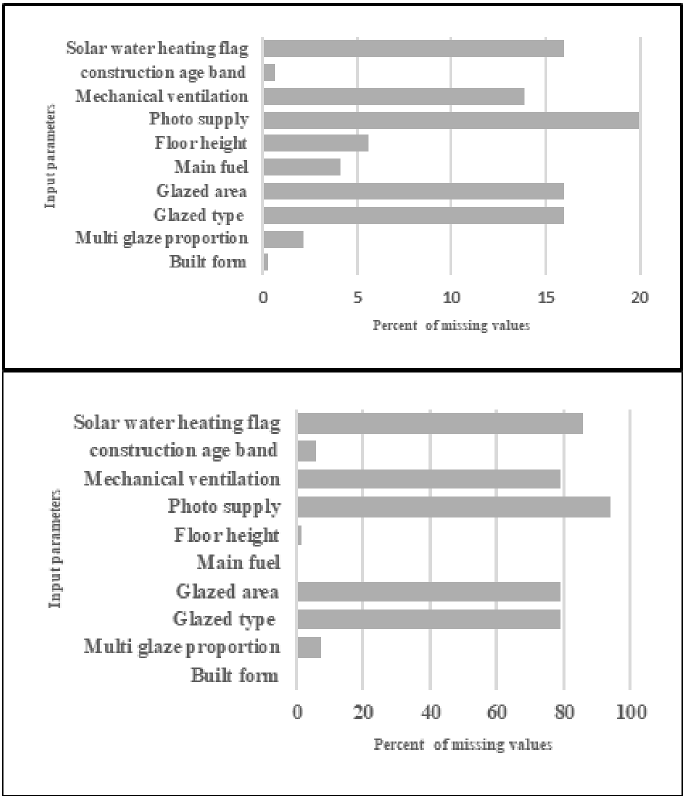
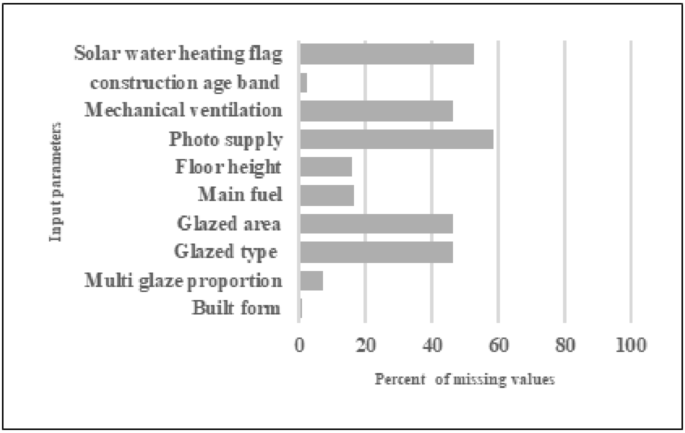


Figure 4. Percentage of missing values for different input features in the dataset

Further analyses show that the issue is more serious in case studies with specific conditions. For example, in case studies with an EPC rating of A, every data point has at least one missing value for input features. Moreover, for five of the input features, as can be observed in figure 5, over 80% of the data is missing. A similar situation occurs in case studies with other EPC ratings, where at least 40% of the data points are missing for five of the input features.

(A)



(B)

Figure 5. percent of missing values for different input features in case studies with EPC rating (A) A and (B) B

A further investigation of the dataset was conducted to assess the number of missing input features for each case study building. This analysis quantifies how many input features are missing for each data point. As shown in Figure 6, nearly 20% of the case study buildings have at least one missing input feature. Furthermore, over 15% of the data points exhibit at least four missing inputs. This raises significant concerns regarding first, the reliability of the SAP-based tool's calculations and, second, the applicability of the dataset for training black-box models, such as machine learning algorithms.

Although various methodologies exist for handling missing inputs, such as imputing missing values using the average or mode of available data, these approaches may introduce bias and reduce the reliability of the analysis. To ensure the accuracy and robustness of the dataset, data points with missing inputs should be excluded from further analyses and model training. This prevents potential distortions in statistical patterns and enhances the predictive performance of data-driven models.

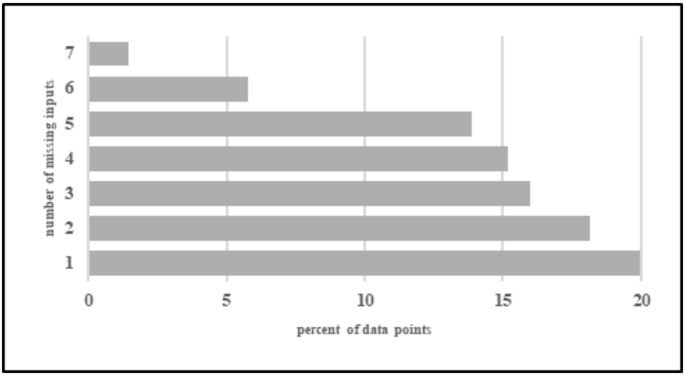


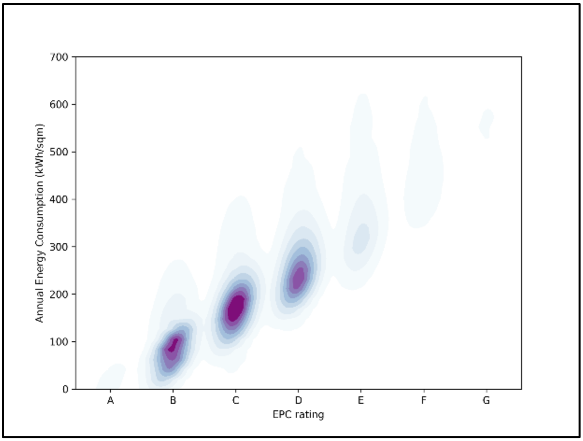
Figure 6. Number of missing inputs in the dataset

5.2 Input/Output Correlation Analysis

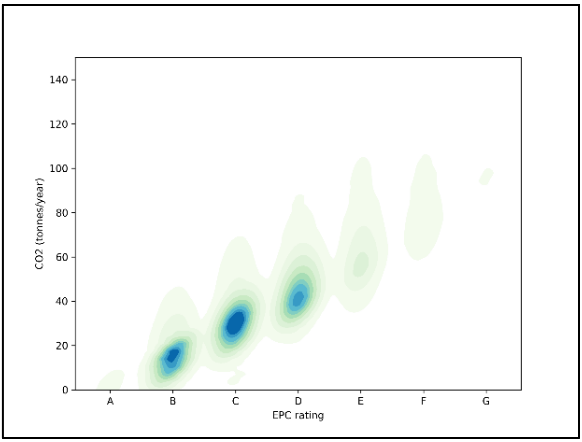
As mentioned earlier, data points in this dataset are directly calculated with SAP based tool; and in the following sections, the calculated results and inputs will be analysed. This section also focuses on the most important outputs that are aligned with the objectives of the research, including the "current energy rating", "current energy consumption", and "current environmental impact". On this context, figure 7 presents the relationships between EPC ratings, annual energy consumption, and CO_2 emissions in the dataset of case study buildings. Part A illustrates how energy consumption varies with EPC ratings, where higher EPC ratings exhibit significantly lower consumption as expected, while poorly rated buildings (E, F, and G) tend to have much higher energy demands. The progressive increase in energy consumption across EPC categories suggests that the rating system effectively captures variations in building energy performance, which validates its use as a benchmarking tool in energy efficiency assessments.

Similarly, part B demonstrates the relationship between EPC ratings and CO_2 emissions, revealing a similar pattern to that of energy consumption. Buildings with high EPC ratings (A and B) produce considerably lower emissions, while those in lower categories (E, F, and G) contribute significantly more CO_2 to the environment. This strong correspondence suggests that buildings with higher energy demand often rely on carbon-intensive energy sources, further amplifying their environmental footprint. Notably, the group of buildings in the C and D categories reinforces the idea that retrofitting efforts in this segment could lead to substantial reductions in emissions.

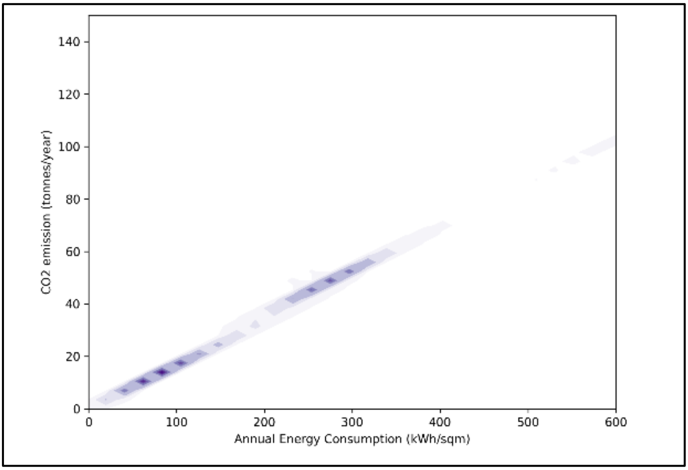
Furthermore, the gradual increase in emissions with declining EPC ratings indicates that the energy efficiency of a building directly translates into its carbon impact, which makes EPC improvements a key strategy for sustainability.



(A)



(B)

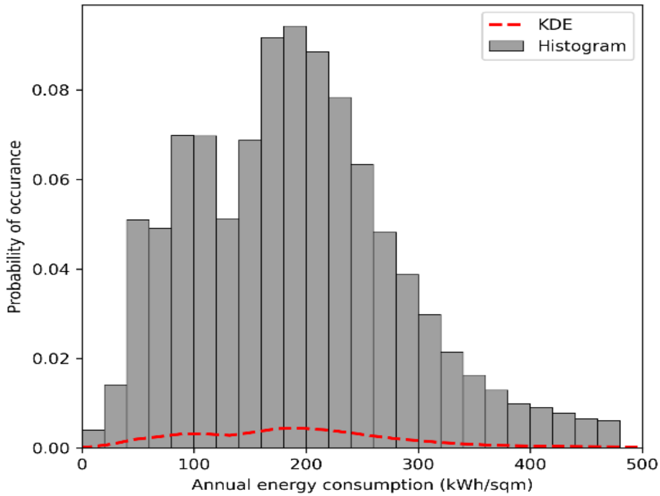


(C)

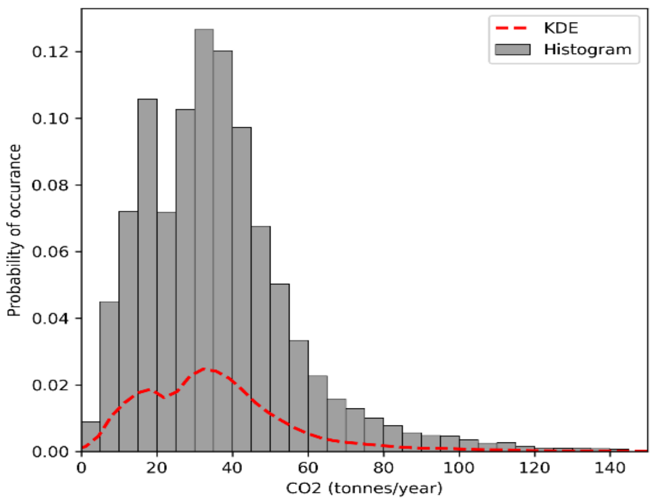
Figure 7. Correlation analyses between outputs (A) EPC - energy consumption (B) EPC - CO_2 emissions (C) energy consumption - CO_2 emissions

Part C explicitly depicts the nearly linear relationship between annual energy consumption and CO_2 emissions, with data points following a strong diagonal trend. This suggests that as energy consumption increases, CO_2 emissions rise proportionally, emphasising the direct dependence of carbon emission on energy use. The density distribution along the trendline indicates that most buildings follow this predictable pattern, with only minor deviations. These deviations could be related to differences in energy sources, such as buildings using renewable energy or more efficient heating systems, resulting in slightly lower emissions for the same energy consumption level. The observed linearity highlights the critical role of reducing energy demand in mitigating carbon emissions, which highlights the necessity for both efficiency improvements and cleaner energy sources in the building sector.

Moreover, Figure 8 presents the probability distributions of annual energy consumption and CO_2 emissions for the case study buildings using histograms and Kernel Density Estimation (KDE). The histogram bars illustrate the frequency of different energy consumption and emissions intervals, while the KDE curve provide continuous, smoothed probability density estimates.



(A)



(B)

Figure 8. Histogram and KDE for (A) annual energy consumption and (B) CO_2 emissions

The distribution of annual energy consumption (kWh/m²) shows that most buildings consume between approximately 160 and 240 kWh/m², while a smaller number of buildings exhibit significantly higher energy consumption, extending up to 400 kWh/m². On the other hand, some records indicate rare cases with net consumption below 60 kWh/m², which could be related to on-site generation covering a significant share of their demand. Furthermore, the KDE curve follows the histogram's general trend but provides a continuous representation of the distribution.

Similarly, the CO_2 emissions distribution (tonnes/year) shows that the majority of buildings emit around 40 tonnes of CO_2 annually. However, the implementation of environmentally friendly measures demonstrates that it is possible for many case study buildings to achieve emissions of around 20 tonnes of CO_2 annually.

5.3 Data Imbalance

Figure 9 shows the distribution of EPC rating assigned to the buildings. It can be observed that the majority of the buildings fall within the middle rating categories. Specifically, the highest proportion of buildings (34%) have been assigned a C rating, which indicates a moderate level of energy efficiency. This is followed by 29% of buildings with a D rating, which suggests a slightly lower efficiency but still within an acceptable range.

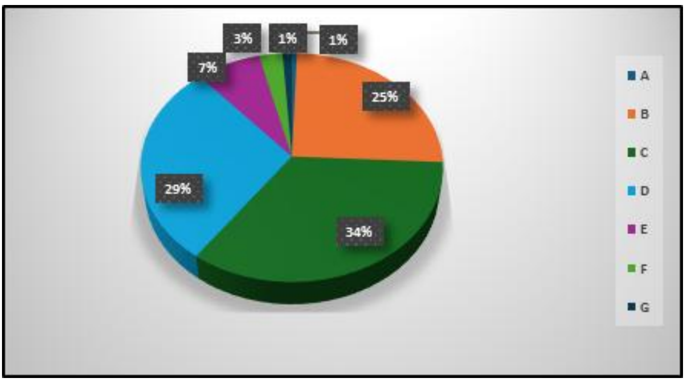


Figure 9. EPC rating distribution (percentage) in case study buildings

On the other hand, a smaller percentage of buildings fall into the very low or high efficiency categories. Around 7% of buildings are rated E, while 3%

are classified as F, and only 1% of buildings have the worst rating of G. Similarly, only less than 1% of the buildings have achieved the highest efficiency rating of A, suggesting that very few buildings meet the most strict energy efficiency criteria. This distribution highlights class imbalance in the dataset which will be discussed further in the study. When using this dataset for training machine learning models, it is important to address this issue with methodologies such as SMOTE (Synthetic Minority Over-sampling Technique) and SMOTE-Tomek to avoid model bias; which will be further discussed in the following sections.

One of the issues that can affect the accuracy of a model is the imbalanced distribution of classes in the dataset; meaning that one class has significantly more instances than others. This can negatively impact model training procedure, as the model tends to fit toward majority classes which leads to biased prediction and unreliable accuracy metrics. On this context, Zhang et al. (2021) conducted a research to highlight the effect of data imbalance in building energy performance prediction. They showed that addressing data imbalance issue in a machine learning model can decrease building energy load prediction mean absolute error by 12% and enhance R^2 score up to 14%.

SMOTE and SMOTE-Tomek are popular methods used to address the problem of imbalanced datasets by generating synthetic examples for the minority classes. In these techniques new instances will be created using interpolation between existing datapoints and it will populate the minority classes. Swana et al. (2022) investigated integrating SMOTE technique with Naïve Bayes (NB), SVM, and K nearest neighbours (KNN) machine learning algorithms to improve fault detection accuracy in a wound-rotor induction generator. The results of this study revealed that SMOTE integrated with KNN outperformed other models that could increase the accuracy score almost 30 points.

As mentioned earlier, imbalanced datasets can negatively affect the performance of data-driven models (particularly classifiers) in a way that the classifier system tends to be biased in favour of Majority classes. The classifier also tends to ignore the minority instances and detects them as noise (Napierala and Stefanowski 2016). To address this, three types of data resampling techniques including over-sampling, under-sampling, and combine-sampling have been developed to balance the dataset. over-sampling is a technique to duplicate or produce new synthetic data within the minority class, whereas under-sampling is a technique to delete or merge data within the majority category. Combine-sampling is one method that combines over-sampling and under-sampling (Jonathan, Panca Hadi, and Ruldeviyani 2020).

On this context, an over-sampling method SMOTE and a combine-sampling method, SMOTE-Tomek, have been utilised in this section to help balance the dataset, and give the model more diverse training data for minority classes to improve its performance.

SMOTE is based on the k-nearest neighbour to generate new synthetic sampling in the feature space based on a certain percentage for the minority classes. SMOTE generates new synthetic samples for the minority class by creating data points between a minority instance and its nearest neighbours to help balance the dataset. This synthetically generated data can be formulated as shown in equation 9 (Swana, Doorsamy, and Bokoro 2022).

$$S_{syn} = r\left(S_{knn} - S_f\right) + S_f \tag{9}$$

Where S_{syn} is generated synthetic samples; S_f feature samples; S_{knn} considered feature sample k-nearest neighbour; and r is a random number between 0 and 1. Moreover, SMOTE-Tomek is a hybrid method that combines SMOTE's synthetic oversampling with Tomek Links, a technique for cleaning data. Tomek Links are applied to remove noisy examples, specifically those are difficult to classify because they are too close to instances of the opposite class (Batista, Prati, and Monard, 2004.).

As a result, Figure 10 illustrates the impact of removing data points with missing inputs from the dataset, followed by the application of SMOTE and SMOTE-Tomek on the distribution of EPC ratings; as one of the key output variables in the dataset. In part (A), the imbalanced dataset shows certain EPC labels, such as label C, dominating the distribution, while other labels like F and G are significantly underrepresented. Parts (B) and (C) illustrate the application of SMOTE and SMOTE-Tomek, where synthetic samples have been added to the minority classes (F, G), resulting in a more balanced class distribution. For instance share of case studies with EPC label F has increased from 2% to 16.66% and 17.35% using SMOTE and SMOTE-Tomek, respectively.

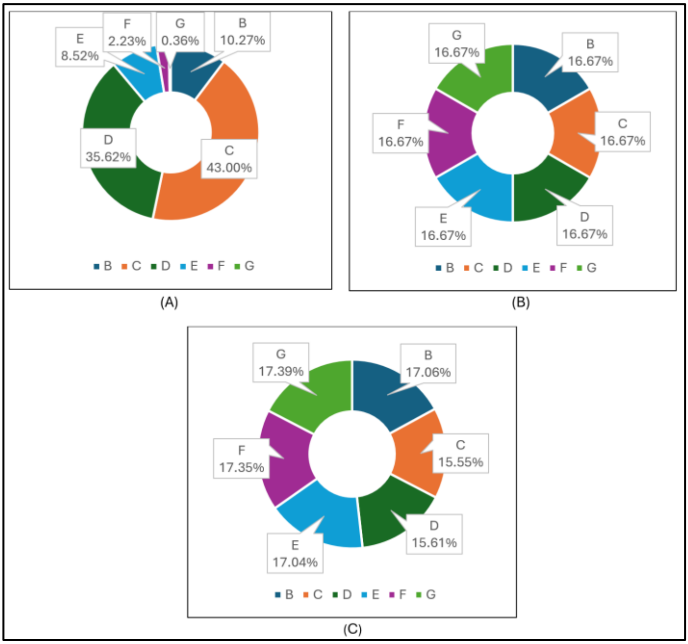
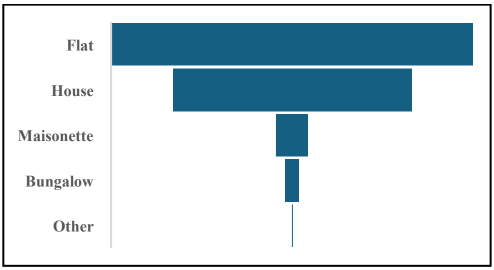


Figure 10. Effect of resampling techniques on target variable distribution: (A) Imbalanced dataset (B) SMOTE applied (C) SMOTE-Tomek applied

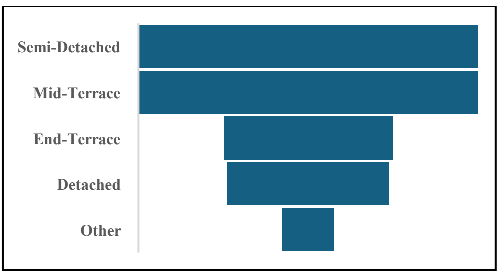
This class imbalance issue extends beyond the target variable to the input features as well, as evident in Figure 11, where certain categories are underrepresented while others appear more frequent. Figure 11 presents the distribution of categorical input variables within the dataset, which highlights the frequency of different categories across six key features: property type, built form, glazing type, heating system, ventilation system, and hot water system. The distributions reveal a noticeable imbalance, with certain categories being significantly overrepresented while others appear infrequently. Such bias in categorical distributions may affect the performance of predictive models by increasing their bias toward common property characteristics and reducing their ability to provide accurate predictions for underrepresented categories.

In terms of property type, flats dominate the dataset, followed by houses, while maisonettes, bungalows, and other property types are far less frequent. A similar trend is observed in the built form, where semi-detached and mid-terrace buildings are the most common building types. The distribution of other categorical variables follows a similar pattern of imbalance. Double glazing is by far the most common glazing type, while secondary, triple, and single glazing options appear much less frequently. Heating systems are dominated by boiler systems, with electric storage systems, room heaters, and alternative heating solutions being significantly underrepresented.

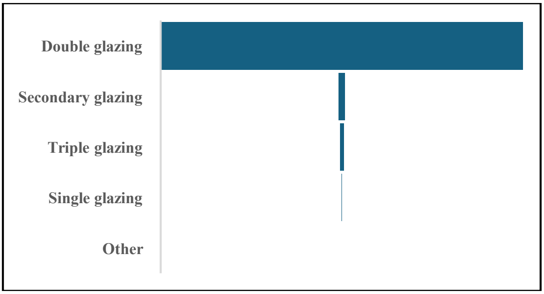
Furthermore, a strong imbalance is evident in the ventilation and hot water system distributions. Natural ventilation is the most common method, while mechanical ventilation systems are rarely present. Similarly, hot water systems are mostly sourced from heating systems, with electric immersion and other types appearing at much lower frequencies. The underrepresentation of particular buildings' features such as air source heat pumps, triple glazing, and advanced ventilation systems may limit the model's ability to generalise predictions for case studies featuring these features. Consequently, predictive models trained on this dataset will primarily reflect conventional residential heating and ventilation configurations, with limited capacity to extend insights to buildings employing less common, energy-efficient, or emerging technologies.



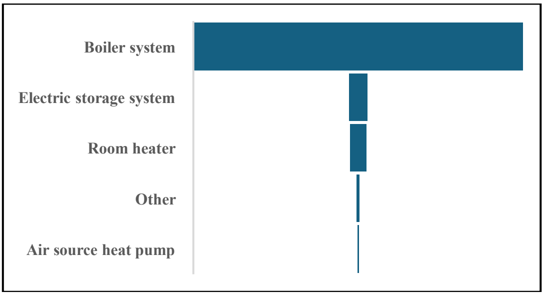
(A)



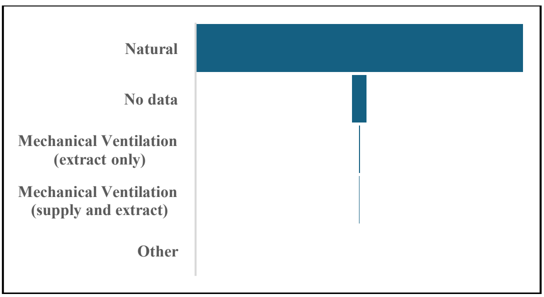
(B)


Double glazing
Secondary glazing
Triple glazing
Single glazing
Other

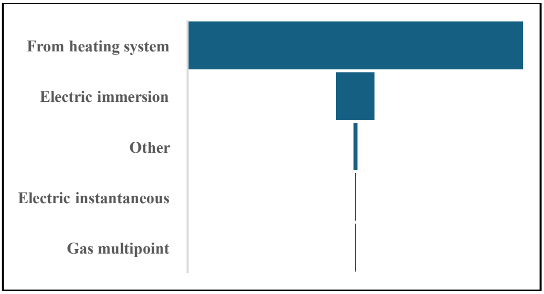
(C)


Boiler system
Electric storage system
Room heater
Other
Air source heat pump

(D)


Natural
No data
Mechanical Ventilation (extract only)
Mechanical Ventilation (supply and extract)
Other

(E)



(F)

Figure 11. Distribution of categorical input variables in the dataset; (A) property type (B) built form (C) glazing type (D) heating system (E) ventilation system (F) hot water system

5.4 Data Inconsistencies in Input Variables

Another discrepancy observed in the dataset's input features is that building envelope characteristics, including floor description, roof description, and walls description, are represented using both numerical and categorical values. Specifically, for some case studies, the average thermal transmittance of the floor, roof, and walls has been provided as a numerical value (U-value), whereas for the remaining cases, descriptive information regarding wall type and insulation has been included. As shown in Figure 12, descriptive entries account for more than 85% of the case studies, while numerical values are less frequently reported.

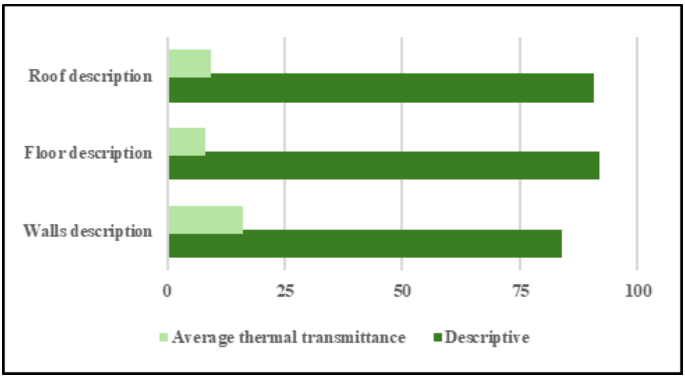


Figure 12. Representation of building envelope features in the dataset

In order to address this discrepancy, the study has utilised SAP methodology documentation which assigns a U-value to building envelope features, including walls, roof and floor based on their construction year, material and insulation used according to table 3. This approach quantifies the quality of insulation across all case study buildings in the dataset and facilitate comparison of different data points in terms of effect of building envelope material on energy performance. Same tables can be found in (Building research establishment 2012) which quantify the insulation quality of different types of roofs and floors in terms of a U-value.

Table 3. External walls U-value in England based on construction age band, wall type, and insulation

External wall type	Building construction year											
	before 1900	1900 - 1929	1930 - 1949	1950 - 1966	1967 - 1975	1976 - 1982	1983 - 1990	1991 - 1995	1996 - 2002	2003 - 2006	2007 - 2011	2012 onwards
Solid brick- as built	2.1	2.1	2.1	2.1	1.7	1	0.6	0.6	0.45	0.35	0.3	0.28
Solid brick- 50mm insulation	0.6	0.6	0.6	0.6	0.55	0.45	0.35	0.35	0.3	0.25	0.21	0.21
Solid brick- 100mm insulation	0.35	0.35	0.35	0.35	0.35	0.32	0.24	0.24	0.21	0.19	0.17	0.16
Solid brick- 150mm insulation	0.25	0.25	0.25	0.25	0.25	0.21	0.18	0.18	0.17	0.15	0.14	0.14
Solid brick- 200mm insulation	0.18	0.18	0.18	0.18	0.18	0.17	0.15	0.15	0.14	0.13	0.12	0.12
Cavity- as built	2.1	1.6	1.6	1.6	1.6	1	0.6	0.6	0.45	0.35	0.3	0.28
Unfilled cavity- 50mm insulation	0.6	0.53	0.53	0.53	0.53	0.45	0.35	0.35	0.3	0.25	0.21	0.21
Unfilled cavity- 100mm insulation	0.35	0.32	0.32	0.32	0.32	0.3	0.24	0.24	0.21	0.19	0.17	0.16
Unfilled cavity- 150mm insulation	0.25	0.23	0.23	0.23	0.23	0.21	0.18	0.18	0.17	0.15	0.14	0.14
Unfilled cavity- 200mm insulation	0.18	0.18	0.18	0.18	0.18	0.17	0.15	0.15	0.14	0.13	0.12	0.12
Filled cavity	0.5	0.5	0.5	0.5	0.5	0.4	0.35	0.35	0.45	0.35	0.3	0.28
Filled cavity- 50mm insulation	0.31	0.31	0.31	0.31	0.31	0.27	0.25	0.25	0.25	0.25	0.21	0.21
Filled cavity- 100mm insulation	0.22	0.22	0.22	0.22	0.22	0.2	0.19	0.19	0.19	0.19	0.17	0.16
Filled cavity- 150mm insulation	0.17	0.17	0.17	0.17	0.17	0.16	0.15	0.15	0.15	0.15	0.14	0.14
Filled cavity- 200mm insulation	0.14	0.14	0.14	0.14	0.14	0.13	0.13	0.13	0.13	0.13	0.12	0.12
Timber frame- as built	2.5	1.9	1.9	1	0.8	0.45	0.4	0.4	0.4	0.35	0.3	0.28

Timber frame-internal insulation	0.6	0.55	0.55	0.4	0.4	0.4	0.4	0.4	0.4	0.35	0.3	0.28

Figure 13 illustrates the distribution of annual energy consumption in relation to the U-value of external walls for the case study buildings. This highlights two primary regions of high data density, indicating distinct clusters of buildings with similar thermal performance characteristics. The first cluster, centred around a lower U-value of 0.5 W/m²K, corresponds to recent constructed buildings with higher insulation levels, which tend to have energy consumption levels ranging between 150 and 200 kWh/sqm. In contrast, the second concentration of data points appears at a significantly higher U-value around 2.0 W/m²K, where the energy demand is noticeably greater, often exceeding 250 kWh/sqm; which highlights the fact that buildings with lower insulation quality experience higher energy consumption due to increased heat transfer through the external envelope.

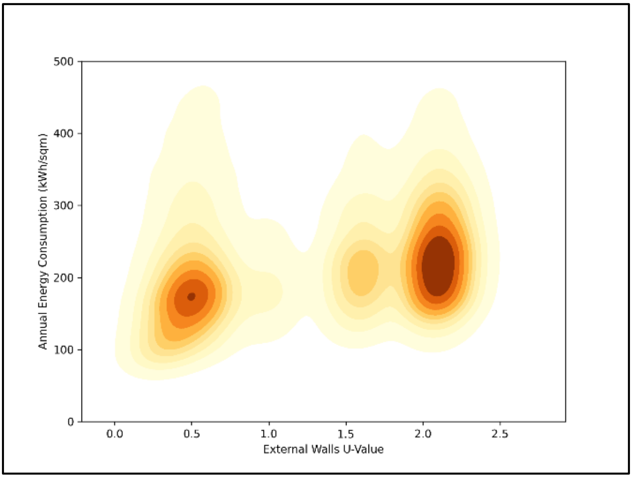


Figure 13. Distribution of external wall u-values in dataset and their relation with annual energy consumption

Furthermore, the figure reveals a non-linear relationship between the U-value of external walls and annual energy consumption. While a general trend suggests that higher U-values lead to greater energy consumption, the distribution is not strictly uniform, as evidenced by the presence of secondary peaks and variations across the range. Notably, some buildings with moderate U-values (around 1.5 W/m²K) exhibit lower energy consumption than those with both lower and higher U-values, indicating that factors beyond insulation such as HVAC efficiency, glazing system, and building type play a significant role. This non-linearity underscores the complexity of building energy performance, where improvements in insulation alone do not always result in proportionate energy savings, and the interaction of multiple design parameters must be considered for an accurate assessment of energy efficiency.

6. Conclusion

The UK EPC dataset is a crucial resource for assessing residential buildings energy performance, guiding policy decisions, and supporting retrofit planning. However, the reliability and applicability of the dataset depend on the accuracy of its underlying SAP methodology. This study has investigated the conceptual framework behind energy performance calculations to highlight its key limitations such as utilising steady-state equations, missing data, and inconsistencies in EPC predictions. These issues present challenges for direct use in benchmarking and predictive modelling, necessitating further utilisation of data preprocessing techniques.

The analyses conducted in the study underscore the impact of data gaps and class imbalances, which can negatively affect the effectiveness of EPC-based benchmarking and machine learning applications. Addressing these discrepancies is essential to enhance the dataset's reliability for evaluating retrofit strategies and supporting data-driven energy modelling.

To improve the dataset's quality, advanced preprocessing techniques, such as data imputation, feature engineering, and synthetic oversampling, can be employed to mitigate data inconsistencies. These enhancements enable the EPC dataset to serve as a more robust benchmarking tool and training dataset for machine learning models. However, it is important to consider that, even with the application of data science techniques to address dataset limitations, the EPC dataset remains the output of SAP-based tools, which rely on quasi steady-state calculations and simplified rule-of-thumb equations. As a result, predictions derived from models trained on the UK EPC dataset should not be expected to achieve high accuracy in real-world scenarios. At best, these models may serve as tools for regulatory compliance checks rather than precise energy performance assessments.

By refining dataset quality and improving predictive capabilities, this study contributes to the broader goal of leveraging data-driven approaches for accurate and scalable building energy performance assessment. Future research should focus on integrating real-world energy consumption data with EPC records to further validate and enhance predictive modelling frameworks.

References

Batista, Gustavo E A P A, Ronaldo C Prati, and Maria Carolina Monard. 2004. “A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data.”

Bolton, Paul. 2024. “Energy Efficiency of UK Homes.”

Building research establishment. 2012. *The Government’s Standard Assessment Procedure for Energy Rating of Dwellings*.

Chartered Institution of Building Services Engineers. 2015. *Building Performance Modelling - CIBSE AM11*. Chartered Institution of Building Services Engineers.

Chartered Institution of Building Services Engineers .2021. *Environmental Design - CIBSE Guide A*. Chartered Institution of Building Services Engineers.

Chartered Institution of Building Services Engineers .2022. *Evaluating Operational Energy Use at the Design Stage - CIBSE TM54*. Chartered Institution of Building Services Engineers.

Department for Levelling Up, Housing & Communities. 2023. “Energy Performance Certificates and Display Energy Certificates Data for Buildings in England and Wales.” Department for Levelling Up, Housing & Communities. 2023. https://epc.opendatacommunities.org/.

Few, Jessica, Despina Manouseli, Eoghan McKenna, Martin Pullinger, Ellen Zapata-Webborn, Simon Elam, David Shipworth, and Tadj Oreszczyn. 2023. “The Over-Prediction of Energy Use by EPCs in Great Britain: A Comparison of EPC-Modelled and Metered Primary Energy Use Intensity.” *Energy and Buildings* 288 (June). https://doi.org/10.1016/j.enbuild.2023.113024.

Hardy, A., and D. Glew. 2019. “An Analysis of Errors in the Energy Performance Certificate Database.” *Energy Policy* 129 (June):1168–78. https://doi.org/10.1016/j.enpol.2019.03.022.

Jonathan, Bern, Putra Panca Hadi, and Yova Ruldeviyani. 2020. *Proceedings, the 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology : July 7-8, 2020, Bali, Indonesia*. IEEE.

Klanatsky, Peter, François Veynandt, and Christian Heschl. 2023. “Grey-Box Model for Model Predictive Control of Buildings.” *Energy and Buildings* 300 (December). https://doi.org/10.1016/j.enbuild.2023.113624.

Krstić, Hrvoje, and Mihaela Teni. 2017. “Review of Methods for Buildings Energy Performance Modelling.” In *IOP Conference Series: Materials Science and Engineering*. Vol. 245. Institute of Physics Publishing. https://doi.org/10.1088/1757-899X/245/4/042049.

Li, Wenliang, Yuyu Zhou, Kristen Cetin, Jiyong Eom, Yu Wang, Gang Chen, and Xuesong Zhang. 2017. “Modeling Urban Building Energy Use: A Review of Modeling Approaches and Procedures.” *Energy*. Elsevier Ltd. https://doi.org/10.1016/j.energy.2017.11.071.

Napierala, Krystyna, and Jerzy Stefanowski. 2016. “Types of Minority Class Examples and Their Influence on Learning Classifiers from Imbalanced Data.” *Journal of Intelligent Information Systems* 46 (3): 563–97. https://doi.org/10.1007/s10844-015-0368-1.

Olu-Ajayi, Razak, Hafiz Alaka, Hakeem Owolabi, Lukman Akanbi, and Sikiru Ganiyu. 2023. “Data-Driven Tools for Building Energy Consumption Prediction: A Review.” *Energies*. MDPI. https://doi.org/10.3390/en16062574.

Pachano, José Eduardo, Cristina Nuevo-Gallardo, and Carlos Fernández Bandera. 2025. “An Empirical Comparison of a Calibrated White-Box versus Multiple LSTM Black-Box Building Energy Models.” *Energy and Buildings* 333 (April). https://doi.org/10.1016/j.enbuild.2025.115485.

Shahcheraghian, Amir, Hatef Madani, and Adrian Ilinca. 2024. “From White to Black-Box Models: A Review of Simulation Tools for Building Energy Management and Their Application in Consulting Practices.” *Energies*. Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/en17020376.

Swana, Elsie Fezeka, Wesley Doorsamy, and Pitshou Bokoro. 2022. “Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset.” *Sensors* 22 (9). https://doi.org/10.3390/s22093246.

Yu, Jiaqi, Wen Shao Chang, and Yu Dong. 2022. “Building Energy Prediction Models and Related Uncertainties: A Review.” *Buildings*. MDPI. https://doi.org/10.3390/buildings12081284.

Zhang, Chaobo, Junyang Li, Yang Zhao, Tingting Li, Qi Chen, Xuejun Zhang, and Weikang Qiu. 2021. “Problem of Data Imbalance in Building Energy Load Prediction: Concept, Influence, and Solution.” *Applied Energy* 297 (September). https://doi.org/10.1016/j.apenergy.2021.117139.