Residential Archetypes Dataset

Methodology

Prepared for the Climate Change Committee

kammaclimate.com

Contents

Introduction

In 2025, the Climate Change Committee (CCC) published its advice for the Seventh Carbon Budget (CB7), establishing a legally binding emissions cap for the UK between 2038 and 2042. The advice forms a pathway for emissions reduction from now to 2050, informed by economic modelling of feasible abatement strategies across different sectors.

Buildings, which account for 17% of the UK's direct emissions, play a critical role in the effectiveness of these strategies. To support this work, the CCC has developed a model that generates pathways for decarbonising the heating of homes. A crucial component of this model is a comprehensive dataset of UK residential housing archetypes, which will ensure that the UK's 28 million existing homes are accurately represented.

To develop this dataset, the CCC partnered with Kamma, a geospatial data and technology company specialising in risk, emissions and retrofit opportunity analysis for property-related assets.

The dataset captures key attributes influencing the decarbonisation of heat in existing residential buildings. The data is categorised into archetypes - groups of homes with near-identical characteristics. It deploys a mix of categorical (e.g., typology, insulation levels, heating systems) and continuous variables (e.g., floor area, energy use). It also covers geographic characteristics, including proximity to hydrogen sources and heat networks.

The final dataset includes over 11,000 archetypes. This reflects the diversity of the UK housing stock while maintaining statistical validity. It draws on several baseline datasets, including the Housing Surveys and EPC datasets, while also supplemented with other reliable data sources. The dataset is representative across the UK, shaping future decarbonisation strategies for the residential sector.

1. Process overview

Over three years, Kamma has built Energy Performance Certificate (EPC) data ingestion, normalisation, and validation technology to amass a datastore of over 27.5 million active and expired EPCs. UPRNs[1] have been attached to 27.1 million properties, building an authoritative view of the EPC database, while extensive data modelling has been carried out of properties without an EPC to fill in the gaps using predicted data.

The process detailed here outlines, in the first instance, an approach to refining property-level datasets, starting with the EPC dataset, and then leveraging data from the UK Housing Survey in order to enhance the dataset's coverage and accuracy, with the ultimate goal of generating a set of archetypes that is representative of the entire UK housing stock.

In isolation, the EPC and Housing Survey datasets have data gaps related to the coverage of variables. Therefore, through a series of steps – beginning with data collection, followed by data transformation and cleaning, and culminating in the integration of related variables – the process sought to validate EPC data accuracy with the Housing Survey datasets, alongside other supplementary datasets, including from ONS. By

[1] A Unique Property Reference Number (UPRN) is a unique identifier assigned to every spatial address in the UK, so that every property, from residential homes to commercial buildings and public spaces, can be precisely identified.

categorising properties into distinct archetypes based on a blend of EPC and Housing Survey data, and subsequently validating and recalibrating these archetypes, this methodology provided a robust framework for analysing the decarbonisation pathways of UK housing.

2. Data preparation and normalisation

Data preparation was focused on three primary sources:

Dataset	Description	Source
Energy Performance Certificate (EPC)	EPC data provides detailed information on the energy efficiency of buildings, rating them from A (most efficient) to G (least efficient) based on factors like insulation, heating, and lighting. This data is crucial for homeowners, buyers, and tenants, offering insights into a property's energy consumption and potential costs, as well as identifying opportunities for energy saving improvements. EPC data is also invaluable for policymakers and researchers, aiding in the development of targeted energy conservation strategies and policies to enhance the sustainability of the housing stock.	Open Data Communities: The Energy Performance of Buildings Register
Housing Surveys	UK Housing Surveys, such as the English Housing Survey (EHS), are comprehensive assessments conducted to gather detailed information on the housing conditions and characteristics of residential properties across the United Kingdom. These surveys collect data on a wide range of topics including property types, sizes, conditions, energy efficiency, occupancy patterns, and the socio-economic characteristics of households. The survey datasets used included: • English Housing Survey, 2019: Housing Stock Data: Special Licence Access • English Housing Survey, 2020-2021: Household Data: Special Licence Access • English Housing Survey: Fuel Poverty Dataset, 2019: Special Licence Access • Welsh Housing Conditions Survey, 2017-2018 • Scottish House Condition Survey, 2012-2019 • Northern Ireland House Condition Survey, 2016	UK Data Service (including special licence access)

	These datasets were chosen because they represent the most recent period of 'normal' conditions, excluding abnormal years such as those impacted by the COVID-19 pandemic.	
Office for National Statistics (ONS)	The Office for National Statistics (ONS) is the UK's largest independent producer of official statistics and its recognised national statistical institute. The ONS is responsible for collecting and publishing statistics related to the economy, population, and society at national, regional, and local levels.	ONS

2.1. Housing Survey data preparation

The methodology described below outlines a process aimed at using UK Housing Survey data to bolster the integrity of EPC datasets. This approach enhances the accuracy of Kamma's datastore of EPC data, allowing for the integration of several variables missing from the dataset. This process encompasses several key stages, from the initial data collection to the transformation and cleaning of the data, ultimately leading to the combination of related variables.

Housing Survey data is collected from the UK Data Service in order to validate EPC data accuracy and join certain survey variables onto archetype datasets.

1. **Housing Survey data collection:** The initial step involved the collection of raw survey data from the UK Data Service. This stage is crucial for two primary reasons: it provides a means to validate the accuracy of EPC data and allows for the incorporation of specific survey variables into archetype datasets.
2. **Data transformation and standardisation:** Once collected, the raw survey data undergoes a transformation process. Using the UK Data Service's Data Dictionary, the data is converted back into its original survey response format. This transformation is critical for several reasons. Firstly, it ensures that the data is in a consistent format, making it easier to analyse and compare. Secondly, it reinstates the data in a form that aligns with the original responses, preserving the integrity and context of the information provided by respondents.
3. **Data cleaning:** The next stage involves cleaning the data, which includes the removal of survey responses categorised as 'Not applicable' or 'Answer not provided'. These are converted to nulls to streamline the dataset. This cleaning process is essential for eliminating ambiguities and enhancing the dataset's clarity and usability. By converting irrelevant or incomplete responses to nulls, it is ensured that the dataset only contains valid, actionable data points.
4. **Combine related variables:** Finally, the methodology includes combining variables that, although related, were initially separated by survey questions. An example application of this approach is the combination of 'Mainheat fuel' and 'Mainheat type'. This combination is performed to normalise terms as agreed upon with the Climate Change Committee (CCC). By merging these related variables, the data becomes more cohesive and reflective of actual energy use and heating practices.

2.2. Unify datasets

The aim of unifying Housing Survey datasets across the UK is to create a single, streamlined 'UK Survey' dataset that can be used for more efficient data normalisation and validation. By combining the data into a single repository, it reduces the complexity involved in cross-referencing and calibrating information from various sources, helping to make sure that the dataset represents the entire UK.

1. **Merging survey data:** The initial step involves consolidating survey data from England, Wales, Northern Ireland, and Scotland into a single dataset. This consolidation facilitates easier management and analysis of the data, setting a foundation for uniform normalisation and validation processes.
2. **Data integrity checks:** Following the merge, a rigorous examination of the dataset for completeness and accuracy is conducted. This includes identifying missing data and verifying the consistency of categories for each variable. Such checks are essential for maintaining the dataset's reliability and ensuring that it accurately reflects the diverse housing landscapes across the UK.
3. **Creation of a unified UK table:** The final step is to combine the individual national survey datasets into one UK-wide table. This unified table simplifies the normalisation and validation of terms, making the dataset more user-friendly for subsequent analyses. It serves as a critical resource for accurately assessing housing conditions and trends across the United Kingdom.

2.3. Data normalisation

2.3.1 Housing Surveys

The next step is to normalise Housing Survey terms to a reduced set that is aligned with variable terms agreed upon with the CCC. The process standardised data types, made regional adjustments, cleaned and normalised data, and addressed data gaps. These steps had a dual purpose: to build consistency and clarity in the data, and to ensure that the resulting analyses would be as representative of the UK housing landscape as possible.

1. **Data verification and preparation:** The first step involved an inspection and verification of the regional datasets to ensure their integrity. This included identifying any missing data and examining the distinct categories present. Such verification was essential for establishing a reliable foundation for subsequent normalisation and analysis.
2. **Standardising data types:** Numerical values across the datasets were standardised, especially for factors like grossing factors (adjustments made to survey data to make it representative of a population), income levels, and energy efficiency ratings. Standardisation ensured that these values were consistent across all datasets, facilitating accurate aggregation and comparison.
3. **Regional adjustments:** Adjustments were made to align regional names across datasets, ensuring that each region was consistently represented. This step was crucial for enabling a unified analysis across the UK, preventing discrepancies that could arise from varied naming conventions.
4. **Combining regional data:** Individual regional datasets were merged into a comprehensive UK-wide table, incorporating key variables such as property type, fuel poverty status, heating systems, and energy efficiency ratings. This unified dataset was pivotal for facilitating broad analyses that reflected the entire UK housing landscape.

5. **Cleaning and normalising data:** Data was cleaned and normalised to standardise terms for property types, fuel poverty status, and other categorical data, enhancing the dataset's readability and analytical utility. The normalisation involved adjusting data to a common standard, making it easier to analyse and interpret.
6. **Addressing data gaps:** Efforts were made to identify and rectify any missing or null values, especially those resulting from the integration of different datasets. In cases where data could not be accurately placed within the UK framework due to missing regional information, records were deleted to maintain the dataset's accuracy.
7. **Recalibrating weights:** Grossing factors are multipliers used to scale up sample data to represent the entire population. The grossing factors were adjusted based on UK Gov dwelling stock data[2] to more accurately reflect the distribution of households across different regions. This step involves analysing discrepancies between the sampled data and the real-world figures, and identifying areas where adjustments are necessary. This recalibration ensured that the dataset provided estimates that were representative of the actual population distribution.
8. **Normalising survey variables for joining and validation of EPC variables:** Finally, survey variables related to property characteristics, heating systems, and energy ratings were refined and normalised. This included converting technical descriptions of heating systems into standardised categories and adjusting property sizes based on floor area metrics. The purpose was to ensure that these variables could be accurately joined with and validated against EPC variables.

   Survey variables that were normalised:

- Region
- Property type
- Fuel poverty
- Main heat type
- Location type
- Tenure
- Age
- Walls
- Glazing
- Hot water cylinder
- Roof

2.3.2 EPC

Similarly, the EPC terms were standardised with CCC's required variables. This means that attributes across EPC datasets from different nations that have similar or identical uses are put in the same name format to be easily comparable. By normalising these terms and property attributes, it enabled a detailed analysis of property size distribution, space constraints, outdoor space availability, regional consistency, and housing stock trends, among other factors. This was critical for ensuring consistency in data analysis and interpretation.

[2] Statistical data set: Live tables on dwelling stock (including vacants). GOV UK, December 2023

2.3.2.1 Property type

Property types were normalised into a set of types provided by the CCC. High rise flats were identified by using Ordnance Survey's building heights dataset, imposing a minimum height of 18 metres to determine when a building should be considered a high rise. This follows guidance given by the Health and Safety Executive in February 2024[3] . An alternative approach would have been to look at the number of storeys in each building, but that data was too sparse and was only available for a small subset of the housing stock.

Property types	Description
Flat	An individual housing unit within a larger building, typically with one floor, although this category includes "maisonettes" spread across multiple floors.
High rise flat	A flat located in a building with six or more storeys, usually in an urban setting. A filter is applied to dwellings within a building at least 18m in total height.
Detached	A standalone house not joined to any other house.
Bungalow	A single-storey house, which may be detached, semi-detached, or occasionally terraced.
Terrace	A house connected in a row, sharing side walls with neighbours.
End terrace	The last house in a row of terrace houses, sharing only one common wall.
Semi-detached	A house that shares one common wall with another house.
Park home	A single-storey, detached, pre-fabricated home located in a designated area for such homes.

2.3.2.2 Property size

Properties were classified based on their floor area into three categories: small (<70% relational size for data extrapolation), medium (70-130%), or large (>130%). This classification used University College London's 'Analysis Work to Refine Fabric Energy Efficiency Assumptions for use in Developing the Sixth Carbon Budget (2020)' research paper methodology to classify properties based on their total floor area and property type. The average property sizes were taken from the study. The methodology is outlined from the paper below:

Standard assumptions for extrapolation[4]

		Average dwelling total floor area by type (over a floors m2)		
Property type	**Size**	**England**	**Scotland**	**Average (weighted)**
Flat	Small	55	67	55

[3] Guidance: Find a high-rise residential building. GOV, February 2024

[4] Average dwelling total floor area by type derived from Donaldson, L., 2018. Floor Space in English Homes - main report; Scottish Government, 2017., *Scottish House Condition Survey: 2017 Key Findings*. Scottish Government; Energy Saving Trust (EST), 2019., Determining the Costs of Insulating Non-Standard Cavity Walls and Lofts. Department for Business, Energy & Industrial Strategy (BEIS).

High rise flat	Small	55	67	55
Park home	Small	65	75	69
Bungalow	Medium	77	N/A	77
Semi-detached	Medium	93	101	93
Terrace	Medium	81	94	81
End terrace	Medium	86	94	86
Detached	Large	152	146	151

The average total floor area in square metres to two decimal places was also supplied for each archetype in the dataset.

2.3.2.3 Region

UK regions are first grouped using the SAP 10.2[5] region groupings, which are outlined in the diagram below. An additional grouping was then used with Temperature Region mappings provided by the CCC in order to reduce the total number of regions. Using mathematical approximation, a 1.0°C change in average temperature over the heating period gives a change in Heating Degree Days (and therefore space heating demand) of about 12%. On that basis SAP regions were grouped where the variance in average temperature for the heating period is less than 0.5°C. This resulted in the 21 SAP regions being assigned to 9 larger regions. The analysis supporting these groupings is as follows (underneath the map image):

Group	SAP region	Min[6]	Max	Range
A	South West England	8.55	8.55	0
B	South England	7.8125	8.2875	0.475
C	Midlands/Northern England	7.275	7.5625	0.2875
D	North England	6.55	6.8125	0.2625
E	Wales	7.525	7.8125	0.2875
F	Scottish Borders/Western Isles	6.55	6.925	0.375
G	Central Scotland/Orkney/Shetland	6.0125	6.4875	0.475
H	North Scotland	5.3125	5.7125	0.4
I	Northern Ireland	7.2	7.2	0

[5] The Government's Standard Assessment Procedure for Energy Rating of Dwellings, Version 10.2, pg. 124
[6] Minimum and maximum of the average temperature during the heating period for the SAP regions.

Figure: *UK's 21 SAP regions*[5]

2.3.2.4 Age bands

Property ages are identified in EPC data using the letters A to L. The ages which these letters relate to vary depending on which UK nation the home is located in as they relate to default assumptions about the fabric of homes. In collaboration with the CCC and Kamma agreed to condense the SAP age bands into four categories, as shown below:

SAP letter band	Archetype age band
A and B	Pre-1930
C, D and E	1930 to mid 1970s
F, G and H	Mid 1970s to mid 1990s
I, J, K and L	Mid 1990s to present

Below is a reference of the SAP age banding system for each of the UK nations[7]:

	Years of construction			
Age band	England & Wales	Scotland	Northern Ireland	Park home (UK)
A	Before 1900	Before 1919	Before 1919	–
B	1900-1929	1919-1929	1919-1929	–
C	1930-1949	1930-1949	1930-1949	–
D	1950-1966	1950-1964	1950-1973	–
E	1967-1975	1965-1975	1974-1977	–
F	1976-1982	1976-1983	1978-1985	Before 1983
G	1983-1990	1984-1991	1986-1991	1983-1995
H	1991-1995	1992-1998	1992-1999	(not applicable)
I	1996-2002	1999-2002	2000-2006	1996-2005
J	2003-2006	2003-2007	(not applicable)	(not applicable)
K	2007-2011	2008-2011	2007-2013	2006 onwards
L	2011 onwards	2012 onwards	2014 onwards	(not applicable)

2.3.2.5 Heating systems

The main heating systems data had numerous unstandardised naming conventions, sometimes with different names for the same system. Trying to generate archetypes based on these would have resulted in an unmanageable number of archetypes. These main heating systems were condensed into 14 key systems that considered both the space and water heating technology. The condensed descriptions are indistinguishable when assessed by differences in energy or emissions to their original descriptions.

These 14 systems were agreed upon through consultation with the CCC. The mappings are outlined below:

[7] Appendix S: Reduced Data SAP for existing dwellings, RdSAP 2012 version 9.94 (20th September 2019). BRE, 2019, pg. 3

Archetype heating system	EPC descriptions included
Air Source Heat Pump (ASHP)	- Air Source Heat Pump - ASHP with electric immersion water heating - "Heat pump - Other" - ASHP with oil boiler as secondary system - Heat pump with community gas as secondary system - ASHP with solid fuel burner as secondary system - ASHP with gas boiler as secondary system - ASHP with LPG boiler as secondary system
Biofuel boiler	- Biomass boiler - Biomass boiler with community electric as secondary system - Biomass boiler with electric immersion water heating - Biomass boiler with gas boiler as secondary system
Solid fuel boiler	- Solid fuel boiler - Solid fuel room heaters with electric immersion water heating - Solid fuel boiler with solid fuel room heaters - Solid fuel room heaters with oil boiler as secondary system - Solid fuel room heaters with gas community scheme as secondary - Solid fuel boiler with gas boiler as secondary system
Communal gas	- Gas community scheme - Gas community scheme with electric immersion water heating - Gas community scheme with gas boiler as secondary system - Oil community scheme - LPG community scheme - Oil community scheme with electric immersion water heating - Solid fuel community scheme - Oil community scheme with gas boiler as secondary system - Solid fuel community scheme with gas boiler as secondary system - LPG community scheme with electric immersion water heating - LPG community scheme with oil boiler as secondary system - Oil community scheme with oil boiler as secondary system - Solid fuel community scheme with electric immersion water heating - Gas community scheme with solid fuel as secondary system
Communal low carbon	- Biomass community scheme - Electric community scheme - Waste heat community scheme - Electric community scheme with electric immersion water heating - Biomass community scheme with electric immersion water heating - Electric community scheme with gas boiler as secondary system - Biomass community scheme with gas boiler as secondary system - Community heat pump with electric immersion water heating - Waste heat community scheme with electric immersion water heating - Community heat pump - Community heat pump with community gas as secondary system - Community heat pump with gas boiler as secondary system

	- Electric community scheme with oil boiler as secondary system - Biomass community scheme with oil boiler as secondary system - Biomass community scheme with solid fuel as secondary system
Electric boiler	- Electric boiler - Electric boiler with electric immersion water heating - Electric boiler with gas community scheme as secondary system - Electric boiler with solid fuel as secondary system - Electric boiler with oil boiler as secondary system - Electric boiler with LPG boiler as secondary system
Electric resistive	- Electric room heaters with electric immersion water heating - Electric underfloor heating with electric immersion water heating - Electric room heaters with gas boiler as secondary system - Electric room heaters with gas community scheme as secondary system - Electric underfloor heating with gas community scheme as secondary system - Electric underfloor heating with gas boiler as secondary system - Electric room heaters with solid fuel as secondary system - Electric room heaters with oil boiler as secondary system - Electric underfloor heating with oil boiler as secondary system - Electric radiators with electric immersion water heating - Electric underfloor heating with solid fuel as secondary system - Electric room heaters with LPG boiler as secondary system
Electric storage	- Electric storage heaters with electric immersion water heating - Electric storage heaters with gas boiler as secondary system - Electric storage heaters with gas community scheme as secondary system - Electric storage heaters with solid fuel as secondary system - Electric storage heaters with oil boiler as secondary system
Ground Source Heat Pump (GSHP)	- Ground source heat pump - GSHP with electric immersion water heating - GSHP with gas boiler as secondary system - GSHP with gas community scheme as secondary system - GSHP with oil boiler as secondary system - GSHP with solid fuel as secondary system
Gas boiler	- Gas boiler - Gas boiler with gas community scheme as secondary system - Solid fuel room heaters with gas boiler as secondary system - Solid fuel boiler with gas boiler as secondary system - Warm air boiler with gas community scheme as secondary system - Gas room heaters with gas community scheme as secondary system - Gas boiler with solid fuel as secondary system
Gas boiler (electric hot water)	- Gas boiler with electric immersion water heating
Gas other	- Warm air boiler with gas boiler as secondary system

	- Gas room heaters with gas boiler as secondary system - LPG room heaters with gas boiler as secondary system - Warm air boiler with oil boiler as secondary system - Gas room heaters with solid fuel as secondary system - Warm air boiler with solid fuel as secondary system - Gas room heaters with oil boiler as secondary system - Gas room heaters with LPG boiler as secondary system - Warm air boiler with LPG as secondary system
Gas other (electric hot water)	- Gas room heaters with electric immersion water heating - Warm air boiler with electric immersion water heating
Oil boiler	- Oil boiler - LPG boiler - Oil boiler with electric immersion water heating - LPG boiler with electric immersion water heating - Oil room heater with electric immersion water heating - LPG room heaters with electric immersion water heating - Oil boiler with solid fuel as secondary system - Oil room heaters with gas boiler as secondary system - LPG boiler with solid fuel as secondary system - LPG room heaters - LPG boiler with oil boiler as secondary system - Oil boiler with gas community scheme as secondary system - Oil room heaters with oil boiler as secondary system - LPG boiler with gas community scheme as secondary system - LPG room heaters with solid fuel as secondary system - Oil room heaters with gas community scheme as secondary system - Biomass boiler with oil boiler as secondary system

2.3.2.6 Other variables

Variable	**Description**
Space constrained	The properties with limited living space were identified by examining the floor area per habitable room. The assessment was made using the method outlined in the 'Development of trajectories for residential heat decarbonisation to inform the Sixth Carbon Budget (Element Energy)' by the CCC which states: "*Homes where the available dwelling floor area per habitable room is less than 16 m² were considered space constrained.*"[8] This determination allowed for the assessment of living space adequacy and its correlation with energy efficiency, using specific studies as references for benchmarks.

[8] Development of trajectories for residential heat decarbonisation to inform the Sixth Carbon Budget (Element Energy). 2021, pg. 62

Proportion with outdoor space	The percentage of outdoor space available for each property was estimated using data from the Office for National Statistics (ONS)[9]. Averages were used at the Local Authority level and were split by 'Flat' and 'House' property types. **Outdoor space property type normalisation** Property types were split into either House or Flat. The list below details all house types that were listed as 'Flat' and then all other terms were considered 'House': - Basement flat - Flat - Ground-floor flat - Ground-floor maisonette - Mid-floor flat - Mid-floor maisonette - Top floor flat - Top-floor flat - Top-floor maisonette - Maisonette - Semi-detached flat - Mid-terrace flat - Mid-terrace maisonette - End-terrace flat - Basement maisonette The average percentages for outdoor space were calculated by joining this analysis to property types, tenure and property age data in the archetype dataset.
Wall type	Walls were normalised into the following categories: - Solid uninsulated - Solid insulated - Cavity uninsulated - Cavity uninsulated HTT (Hard To Treat) - Cavity insulated To discern which walls were Hard to Treat (HTT), statistics from the English Housing Survey 2020 to 2021 Energy section, Chapter 4 - Annex Table 4.1: Hard to treat cavity walls, by dwelling characteristics, 2020 were used to attribute whether a home had 'standard fillable' or 'hard to treat cavity' based on the following variables: - Tenure - Dwelling age - Dwelling type

9 Access to gardens and public green space in Great Britain. ONS, 2020

The data from this table is provided in the appendix.

Data for England was used for other regions as no other similar data was available.

Additional rules applied to reflect typical HTT conditions:

1. All high rise flats with uninsulated cavities have been identified as 'Cavity uninsulated HTT'.
2. 'Solid uninsulated (flat)' applied to all solid walled uninsulated flats and high-rise flats.

Wall type Hard to Treat data

Hard-to-treat cavity walls, by dwelling characteristics, 2020
uninsulated cavity wall homes

	Standard fillable	Hart-to-treat cavity	All dwellings	Sample sizes
Tenure				
Owner occupied	1,738	1,359	3,097	*1,363*
Private rented	450	519	969	*431*
Local authority	124	189	313	*216*
Housing association	248	194	443	*301*
Dwelling age				
Pre 1919	141	225	365	*177*
1919 to 1944	412	413	825	*371*
1945 to 1964	408	440	848	*417*
1965 to 1980	716	680	1,395	*649*
1981-1990	533	295	828	*431*
Post 1990	352	209	560	*266*
Dwelling type				
Small terraced house	305	125	429	*192*
Medium/large terraced house	320	369	690	*300*
Semi-detached house	747	419	1,167	*509*
Detached house	558	273	831	*449*
Bungalow	239	102	341	*190*

<table>
<tr><td></td><td>
<table>
<tr><td>Converted flat</td><td>26</td><td>80</td><td>106</td><td>53</td></tr>
<tr><td>Purpose built flat, low rise</td><td>365</td><td>799</td><td>1,164</td><td>567</td></tr>
<tr><td>Purpose built flat, high rise</td><td>0</td><td>94</td><td>94</td><td>51</td></tr>
</table>
</td></tr>
<tr><td>Glazing type</td><td>Windows were normalised into either single or double glazing type. The double glazing category included all forms of glazing that are better than single including high performance glazing and triple glazing.</td></tr>
<tr><td>Roof type</td><td>Roofs were classified based on insulation levels and normalised into the following categories:

- Insulated
- Uninsulated
- Underinsulated
- Underinsulated HTT (Hard To Treat)
- No roof.

To discern which roofs were Hard to Treat (HTT), statistics from the English Housing Survey 2020 to 2021 Energy section, Annex Table 4.11: Loft barriers, by dwelling characteristics, 2020 were used to attribute whether a home had ‘non-problematic’ or ‘hard to treat lofts’ based on the following variables:

- Tenure
- Dwelling age
- Dwelling type

The data from this table is provided in the appendix.

Data for England was used for other regions as no other similar data was available.</td></tr>
<tr><td>Floor type</td><td>Floors were normalised into:

- Solid uninsulated
- Solid insulated
- Suspended uninsulated
- Suspended insulated
- No ground floor.

Floor descriptions that only contained thermal transmittance or U-values were classified based on typical averages for different UK floor types.</td></tr>
<tr><td>Tenure</td><td>Different tenure types were normalised into:

- Private rented
- Social rented
- Owner occupied</td></tr>
</table>

	This was derived from the tenure information captured in the EPC register and then later calibrated against both survey data and published totals for the UK.
Fuel poverty	Boolean variable populated with either "yes" or "no". This variable was populated from the Housing Surveys and the logic is described in section 3.3.
Hot water tank	Boolean variable populated with either "yes" or "no". This variable was populated from the Housing Surveys and the logic is described in section 3.3.

Heritage status	Heritage status was populated with either "Heritage", "Non-heritage" or "Traditional".
	
	Heritage status includes:
	
	- Listed buildings and conservation areas, sourced from:
	  ○ England: Historic England
	  ○ Scotland: Historic Environment Scotland
	  ○ Wales: DataMapWales
	  ○ Northern Ireland: OpenData NI
	
	Traditional status was defined as any property constructed before 1930.

Location type	The location type is populated with either "Rural" or "Urban". These values come from the UK Government's RUC2011 census classification, and are based largely on OAs (Output Areas) and LSOAs (Lower Super Output Areas). The logic is as follows: OAs are treated as 'urban' if they were allocated to a 2011 built-up area with a population of 10,000 people or more, while all remaining OAs are classed as 'rural'. The urban and rural domains are then subdivided into six broad settlement types. The classification also categorised OAs based on context – whether the wider surrounding area of a given OA is sparsely populated or less sparsely populated. RUC2011 of OAs contains 10 classes. RUC2011 of LSOAs, MSOAs (Middle-layer Super Output Areas), and wards is built up from the OA level classification, with assignment to urban or rural made by reference to the category to which the majority of their constituent OAs is assigned. At the SOA/ward scale, settlement form is less homogenous than at OA level and so there are only five settlement categories.
Solar PV	As agreed with the CCC, the "Photo Supply" field from the EPC register was used to flag any properties that have solar PV, or not ("Yes", "No"). Many more properties with solar PV haven't had an EPC assessment done subsequently, and so the total number of installations will be grossly understated.

2.3.2.7 Heat networks

Heat network data was provided to Kamma from Arup, via the CCC. The dataset received by Kamma consisted of a CSV file populated with TOIDs (Topographical IDs) for England, Scotland, and Wales, and BU_FUSIONIDs (Northern Irish equivalent to a TOID) for Northern Ireland. Each entry in the dataset represented a UPRN for a residential property, and a flag determined whether each property would be eligible for future heat network connection.

Kamma mapped TOIDs to UPRNs using Ordnance Survey's AddressBase Plus product, and BU_FUSIONIDs to UPRNs using OSNI's Pointer product. In both cases the match rate was only slightly short of 100% and it was agreed with the CCC that it was sufficient to proceed, ignoring the unmatched records.

Finally, a "Yes", "No" boolean field was added to the dataset where any UPRNs flagged as eligible in the Arup dataset were set to "Yes" – as well as any child UPRNs where the parent UPRN was flagged as eligible.

2.3.2.8 Hydrogen networks

The CCC produced analysis identifying local authority areas that sit within 25km either side of the future hydrogen transmission pipeline. They provided a list of these local authorities in March 2024 with the instruction to flag any "urban" properties within the local authority boundaries as "Yes" for hydrogen network eligibility, and all other properties set to "No".

The list of local authorities were first compared to those within Kamma's analysis environment and discovered that two of the authorities, Scarborough and Selby, had been absorbed into the North Yorkshire local authority in 2023. Rather than incorporating the whole of the North Yorkshire local authority into the scope for eligibility, which would have resulted in non-eligible properties being flagged as eligible, so the local authority data was backtracked for those areas to pre-2023 authorities and were able to match on the legacy authorities.

By leveraging Kamma's local authority data and rural/urban flags (described in "Location type" above), it was possible to flag eligible properties as "Yes" and all other properties as "No".

2.4. Modelling to fill EPC data gaps

Kamma has developed a "shared building" model to infer certain EPC values where no EPC certificate exists for a property. This model applies where a dwelling without an EPC shares the same building as at least one dwelling with an EPC. This is determined by overlaying UPRN geometric points with geometric polygons representing building boundaries, captured by satellite imagery. It makes sense that multiple dwellings in the same building are likely to share the same fabric, and often the same heating systems. This model was applied to the dataset before establishing a join between the EPC dataset and the Housing Survey dataset.

3. EPC to Housing Survey data join

The EPC dataset and the Housing Survey dataset were joined to serve two purposes:

1. To append variables absent from the EPC dataset that exist only in the Housing Survey dataset (explained in this section)
2. To recalibrate the distribution of attributes within each variable of the dataset so that they are an accurate representation of the UK housing stock (explained in section 7).

Housing Survey data was used to append absent variables from the EPC dataset. The two datasets were chosen for being the most comprehensive sources of collected data for UK properties. Joining the pair contributes significantly toward filling data and testing for reliability.

This was executed by calculating the mode of missing categorical variables from the Housing Survey data across common variables between the EPC and Housing Survey datasets. The mode was used as an alternative to adding variables proportionately. Adding Housing Survey variables proportionately to the EPC dataset caused an unmanageable increase in the number of archetypes. Using the mode provided accuracy without significantly increasing the number of final archetypes.

The Housing Survey mode for each specific set of housing attributes was then used to update homes in the EPC database with equivalent attributes. The rationale behind this step was to fill the gaps in the EPC database with valuable data from the Housing Surveys.

For variables that do not exist in the EPC dataset, a mode is calculated across the maximum number of common variables with the Housing Survey dataset, and then joined to the EPC dataset to the maximum number of available variables.

3.1. Common variables

3.1.1 Joining common variables

1. **Identifying common variables between datasets:** Variables that both the EPC and Housing Survey datasets shared were identified:

Variables that do not exist in the EPC data

- Fuel poverty
- Hot water cylinder

Common variables between EPC and Housing Survey

- Region
- Property type
- Main heating system
- Tenure
- Location type
- Dwelling age
- EPC SAP Score

2. **Finding the mode for missing variables:** A mode of the missing variable was calculated across sets of shared variables for the information missing from the EPC dataset but present in the Housing Survey.
3. **Joining the datasets together:** After identifying the most common values for the missing variables, this was then joined to the EPC dataset, thereby aligning as many shared characteristics as possible

when adding the new information to ensure variables were joined on as similar a housing profile as possible.

3.1.2 Recalibration of attribute counts for categorical variables

The EPC register covers around 70% of residential dwellings, and is often out-of-date. If the total counts for each attribute from the EPC data were simply extrapolated to meet the size of the entire housing stock, then the resulting archetypes would be inaccurate.

By comparing the percentage split of dwellings between attributes of each categorical variable in the EPC dataset with those in the more recent Housing Survey dataset, the final distribution of dwelling stock across attributes is adjusted to become reflective of real-world totals.

3.2. Normalising datasets

Two datasets are normalised to ensure they are on a comparable scale, allowing for meaningful comparisons and analysis when combining or comparing data. This is particularly important when the original datasets have different ranges or units, which can skew the results of any analysis if not standardised

Prior to merging these datasets, the raw data from both sources had to be standardised to a uniform set of terms. This process is covered in an earlier section, but to give a little more detail into the process, here are the specific steps:

1. **Normalisation terms defined:** The list of terms to be used for normalising was defined and agreed upon with correspondence with the CCC.
2. **Matching to normalisation tables:** Each variable in the datasets was matched with its corresponding entry in the normalisation table.
3. **Integration of datasets:** After normalisation, the EPC dataset was enriched with the Housing Survey Data. This involved joining the Housing Survey and EPC datasets on multiple common attributes e.g. property type, heating type etc.
4. **Verification of normalisation:** Following the appendage of data, a final review was conducted to confirm that all information had been correctly matched and that normalisation was consistently applied across the combined dataset.
5. **Sharing of normalisation terms:** For transparency and further analysis, the normalisation terms used for each column were documented and made accessible, as outlined in 3.3.

3.3. Joining datasets for the purpose of adding missing variables

Once normalised, the ‘join’ between the two datasets is performed by calculating a mode, the most frequently occurring binary value for each missing variable, for fuel poverty and hot water cylinder variables for each distinct combination of property characteristics.

Using the mode as the most representative value ensured the data was accurate without resulting in an unmanageable number of final archetypes.

Incremental calculation approach: If there is a distinct set of variables that exists in the EPC dataset but not in the Housing Surveys, meaning that a join is not possible, then the calculation is rerun on a reduced set of variables and repeated until a successful join is established. Below is a list of the variables and their prioritisation in this incremental calculation approach:

Fuel poverty mode calculation

- Region
- EPC SAP Score
- Property type
- Age band
- Heating type
- Tenure
- Location type

Hot water cylinder mode calculation

- Heating type
- Region
- Property type
- Age band
- Location type
- EPC SAP score

This multi-variable approach helps ensure that data assigned to each EPC in the Kamma EPC database reflects a similar context or set of characteristics as found in the UK data service housing surveys.

4. Archetype generation

Archetype generation is the process of organising a dataset of separate entries for every property in the UK to a set of archetypes which group properties with shared characteristics.

All variables in the dataset were cleaned and normalised, making it possible to generate 3 archetype datasets that summarised the housing stock in different levels of detail, from "Final Archetypes" being the most detailed, through "Energy Archetypes" representing the energy profiles of the stock, to "Basic Archetypes" being the most condensed and smallest dataset in terms of number of archetypes.

The process used to generate the archetype datasets initially was simply to group properties based on them sharing the same set of attributes across each of the categorical variables in the dataset. This initial process resulted in over 2 million archetypes.

A clustering process was employed to reduce the number of archetypes (see section 6). The resulting archetypes are summarised below.

4.1. Final archetypes

The most detailed dataset was created first and then condensed to produce the less detailed datasets by removing variables in turn.

Total number of archetypes: 11,657
Minimum stock per archetype: 18

Maximum stock per archetype: 131,785
Average stock per archetype: 2,515

The variables included in the final dataset are as follows:

Variable name	**Variable type**	**Possible values / Data type**
Final archetype number	Categorical	Integer from 1 to 11,657
Energy archetype number	Categorical	Integer from 1 to 5,260
Basic archetype number	Categorical	Integer from 1 to 1,184
Sector	Categorical	Residential buildings
Subsector	Categorical	Existing residential buildings
Property type	Categorical	Detached, Semi-detached, Terrace, End terrace, Bungalow, Flat, High rise flat
Age band	Categorical	Pre 1930, 1930 to mid 1970s, Mid 1970s to mid 1990s, Mid 1990s to present
Wall type	Categorical	Solid uninsulated, Solid insulated, Cavity uninsulated, Cavity uninsulated HTT, Cavity insulated
Roof type	Categorical	Insulated, Uninsulated, Underinsulated, Underinsulated HTT, No roof
Floor type	Categorical	Solid uninsulated, Solid insulated, Suspended uninsulated, Suspended insulated, No ground floor
Glazing type	Categorical	Single glazed, Double glazed
Size	Categorical	S, M, L
Space constrained	Categorical	Yes, No
Region	Categorical	Central Scotland/Orkney/Shetland, Midlands/Northern England, North England, North Scotland, Northern Ireland, Scottish Borders/Western Isles, South England, South West England, Wales
Country	Categorical	England, Scotland, Wales, Northern Ireland
Heating system	Categorical	ASHP, Biofuel boiler, Communal gas, Communal low carbon, Electric boiler, Electric resistive, Electric storage, Gas boiler, Gas boiler (electric hot water), Gas other, Gas other (electric hot water), GSHP, Oil boiler, Solid fuel boiler
Location type	Categorical	Urban, Rural
Solar PV	Categorical	Yes, No

Tenure	Categorical	Owner occupied, Private rented, Social rented
Heritage status	Categorical	Heritage, Traditional, Non-heritage
DH eligibility	Categorical	Yes, No
H2 eligibility	Categorical	Yes, No
Fuel poverty	Categorical	Yes, No
Hot water tank	Categorical	Yes, No
Stock	Continuous	Integer
Floor area	Continuous	Decimal
Proportion with outdoor space	Continuous	Percentage expressed as a decimal from 0 to 1
Annual space heating energy demand (excluding electric)	Continuous	Integer (avg kWh/y per dwelling)
Annual space heating demand	Continuous	Integer (avg kWh/y per dwelling)
Annual hot water demand (non electric)	Continuous	Integer (avg kWh/y per dwelling)
Annual hot water demand	Continuous	Integer (avg kWh/y per dwelling)
Space heating fuel consumption (non-electric)	Continuous	Integer (avg kWh/y per dwelling)
Space heating fuel consumption	Continuous	Integer (avg kWh/y per dwelling)
Hot water fuel consumption (non-electric)	Continuous	Integer (avg kWh/y per dwelling)
Hot water fuel consumption	Continuous	Integer (avg kWh/y per dwelling)
Lighting fuel consumption	Continuous	Integer (avg kWh/y per dwelling)
Appliances fuel consumption	Continuous	Integer (avg kWh/y per dwelling)
Cooking fuel consumption	Continuous	Integer (avg kWh/y per dwelling)
Cooking fuel consumption (non-electric)	Continuous	Integer (avg kWh/y per dwelling)
Space heating CO_2 (direct)	Continuous	Decimal (avg kg/y per dwelling)
Hot water CO_2 (direct)	Continuous	Decimal (avg kg/y per dwelling)
Space heating N_2O (direct)	Continuous	Decimal (avg kg/y per dwelling)

Hot water N_2O (direct)	Continuous	Decimal (avg kg/y per dwelling)
Space heating CH_4 (direct)	Continuous	Decimal (avg kg/y per dwelling)
Hot water CH_4 (direct)	Continuous	Decimal (avg kg/y per dwelling)

4.2. Energy archetypes

Subsequent divisions of the basic archetypes were made according to different combinations of heating system and regional location, which together define the property's energy demand. These energy variables comprised heating type and hot water tank details from the survey data, alongside nation, region, and location type from the EPC.

Total number of archetypes: 5,260
Minimum stock per archetype: 18
Maximum stock per archetype: 300,908
Average stock per archetype: 5,573

The variables included in the energy dataset are as follows. This is a subset of the same variables in the final archetypes dataset::

Variable name	**Variable type**	**Possible values / Data type**
Energy archetype number	Categorical	Integer from 1 to 5,260
Basic archetype number	Categorical	Integer from 1 to 1,184
Sector	Categorical	Residential buildings
Subsector	Categorical	Existing residential buildings
Property type	Categorical	Detached, Semi-detached, Terrace, End terrace, Bungalow, Flat, High rise flat
Age band	Categorical	Pre 1930, 1930 to mid 1970s, Mid 1970s to mid 1990s, Mid 1990s to present
Wall type	Categorical	Solid uninsulated, Solid insulated, Cavity uninsulated, Cavity uninsulated HTT, Cavity insulated
Roof type	Categorical	Insulated, Uninsulated, Underinsulated, Underinsulated HTT, No roof
Floor type	Categorical	Solid uninsulated, Solid insulated, Suspended uninsulated, Suspended insulated, No ground floor
Glazing type	Categorical	Single glazed, Double glazed
Size	Categorical	S, M, L

Space constrained	Categorical	Yes, No
Region	Categorical	Central Scotland/Orkney/Shetland, Midlands/Northern England, North England, North Scotland, Northern Ireland, Scottish Borders/Western Isles, South England, South West England, Wales
Country	Categorical	England, Scotland, Wales, Northern Ireland
Heating system	Categorical	ASHP, Biofuel boiler, Communal gas, Communal low carbon, Electric boiler, Electric resistive, Electric storage, Gas boiler, Gas boiler (electric hot water), Gas other, Gas other (electric hot water), GSHP, Oil boiler, Solid fuel boiler
Location type	Categorical	Urban, Rural
Solar PV	Categorical	Yes, No
Stock	Continuous	Integer
Floor area	Continuous	Decimal
Proportion with outdoor space	Continuous	Percentage expressed as a decimal from 0 to 1
Annual space heating energy demand (excluding electric)	Continuous	Integer (avg kWh/y per dwelling)
Annual space heating demand	Continuous	Integer (avg kWh/y per dwelling)
Annual hot water demand (non electric)	Continuous	Integer (avg kWh/y per dwelling)
Annual hot water demand	Continuous	Integer (avg kWh/y per dwelling)
Space heating fuel consumption (non-electric)	Continuous	Integer (avg kWh/y per dwelling)
Space heating fuel consumption	Continuous	Integer (avg kWh/y per dwelling)
Hot water fuel consumption (non-electric)	Continuous	Integer (avg kWh/y per dwelling)
Hot water fuel consumption	Continuous	Integer (avg kWh/y per dwelling)
Lighting fuel consumption	Continuous	Integer (avg kWh/y per dwelling)
Appliances fuel consumption	Continuous	Integer (avg kWh/y per dwelling)
Cooking fuel consumption	Continuous	Integer (avg kWh/y per dwelling)
Cooking fuel consumption (non-electric)	Continuous	Integer (avg kWh/y per dwelling)

Space heating CO_2 (direct)	Continuous	Decimal (avg kg/y per dwelling)
Hot water CO_2 (direct)	Continuous	Decimal (avg kg/y per dwelling)
Space heating N_2O (direct)	Continuous	Decimal (avg kg/y per dwelling)
Hot water N_2O (direct)	Continuous	Decimal (avg kg/y per dwelling)
Space heating CH_4 (direct)	Continuous	Decimal (avg kg/y per dwelling)
Hot water CH_4 (direct)	Continuous	Decimal (avg kg/y per dwelling)

4.3. Basic archetypes

This is the smallest archetype dataset and organises the housing stock around fewer groupings, focusing on age, type and fabric of dwellings.

Total number of archetypes: 1,184
Minimum stock per archetype: 18
Maximum stock per archetype: 990,692
Average stock per archetype: 24,760

The variables included in the basic dataset are as follows. This is a subset of the same variables in the final and energy archetype datasets:

Variable name	**Variable type**	**Possible values / Data type**
Basic archetype number	Categorical	Integer from 1 to 1,184
Sector	Categorical	Residential buildings
Subsector	Categorical	Existing residential buildings
Property type	Categorical	Detached, Semi-detached, Terrace, End terrace, Bungalow, Flat, High rise flat
Age band	Categorical	Pre 1930, 1930 to mid 1970s, Mid 1970s to mid 1990s, Mid 1990s to present
Wall type	Categorical	Solid uninsulated, Solid insulated, Cavity uninsulated, Cavity uninsulated HTT, Cavity insulated
Roof type	Categorical	Insulated, Uninsulated, Underinsulated, Underinsulated HTT, No roof
Floor type	Categorical	Solid uninsulated, Solid insulated, Suspended uninsulated, Suspended insulated, No ground floor
Glazing type	Categorical	Single glazed, Double glazed

Size	Categorical	S, M, L
Space constrained	Categorical	Yes, No
Stock	Continuous	Integer
Floor area	Continuous	Decimal
Proportion with outdoor space	Continuous	Percentage expressed as a decimal from 0 to 1
Annual space heating energy demand (excluding electric)	Continuous	Integer (avg kWh/y per dwelling)
Annual space heating demand	Continuous	Integer (avg kWh/y per dwelling)
Annual hot water demand (non electric)	Continuous	Integer (avg kWh/y per dwelling)
Annual hot water demand	Continuous	Integer (avg kWh/y per dwelling)
Space heating fuel consumption (non-electric)	Continuous	Integer (avg kWh/y per dwelling)
Space heating fuel consumption	Continuous	Integer (avg kWh/y per dwelling)
Hot water fuel consumption (non-electric)	Continuous	Integer (avg kWh/y per dwelling)
Hot water fuel consumption	Continuous	Integer (avg kWh/y per dwelling)
Lighting fuel consumption	Continuous	Integer (avg kWh/y per dwelling)
Appliances fuel consumption	Continuous	Integer (avg kWh/y per dwelling)
Cooking fuel consumption	Continuous	Integer (avg kWh/y per dwelling)
Cooking fuel consumption (non-electric)	Continuous	Integer (avg kWh/y per dwelling)
Space heating CO_2 (direct)	Continuous	Decimal (avg kg/y per dwelling)
Hot water CO_2 (direct)	Continuous	Decimal (avg kg/y per dwelling)
Space heating N_2O (direct)	Continuous	Decimal (avg kg/y per dwelling)
Hot water N_2O (direct)	Continuous	Decimal (avg kg/y per dwelling)
Space heating CH_4 (direct)	Continuous	Decimal (avg kg/y per dwelling)
Hot water CH_4 (direct)	Continuous	Decimal (avg kg/y per dwelling)

5. Archetype design conditions

Archetype design conditions refers to a set of predefined criteria or rules that guide the grouping or clustering of properties into meaningful categories of archetypes.

This approach is categorised into Basic Archetypes, Energy Archetypes, and Final Archetypes, however first it worth defining a few key terms:

- **Minimum population threshold:** This refers to a predefined criterion used to determine the smallest size of a property group, or archetype, that is considered significant enough to be analysed independently. It sets the lower limit for the number of properties that constitute a 'meaningful category', meaning that it represents a significant proportion of homes.

  Useful energy is defined as the energy consumed after accounting for losses due to heating system inefficiencies. The archetypes were ordered in a list by total useful energy descending. The top 70% of archetypes in that list were considered meaningful, while the bottom 30% were clustered together using hierarchical clustering. The project aimed to create as few categories as possible without impacting accuracy.

  This ensures that the focus remains on those property groups that have the most substantial impact on overall emissions. This threshold is flexible and can be changed depending on the desired number of archetypes.

- **Clustering:** This refers to the process of organising properties within the joined database into groups, or "archetypes," based on shared characteristics that are relevant to their energy efficiency, emissions profile, or other related criteria. It's a method used to analyse and segment the housing stock into meaningful categories, making it easier to understand and address the diverse energy challenges and opportunities across different types of properties.

In the methodology, 'design conditions' are applied to guide the clustering process by defining specific criteria that dictate how properties are grouped, ensuring distinct characteristics always result in separate clusters. The variables with design conditions are listed below, meaning that merging these variables to create clusters is either prohibited or restricted. If a variable does not appear below then there are no restrictions for that variable.

Category	Design condition	Allowable merged variables	Explanation
Property type	Flat	-	Properties are to be categorised according to their property type, acknowledging that each type exhibits unique patterns of energy use and potential for emission reduction. The rationale behind this is to address the distinct retrofit needs inherent to different property types. For example, flats, particularly those in high rise buildings, face specific challenges
	High rise flat	-	
	Detached	Detached, Bungalow	
	Park home	-	

	Semi-detached	Terrace, Semi-detached	that make certain energy efficient solutions, like heat pumps, impractical, unlike detached homes which may accommodate a wider range of retrofit options.
Wall type	Cavity insulated	-	Additionally, properties are clustered based on the construction materials and insulation status of their walls, recognising that these materials' thermal properties directly influence the suitability of various retrofit interventions. For instance, homes with solid walls offer a different set of retrofit challenges and opportunities compared to those with cavity walls. A property with solid walls may be more suited for external wall insulation projects, whereas insulating a cavity wall might be simpler and involve injecting insulation material into the cavity space.
	Cavity uninsulated	-	
	Solid insulated	-	
	Solid uninsulated	-	

Nation	England	-	The decision to cluster properties based on their nation and SAP region is informed by the understanding that climate conditions, energy policies, and retrofit programmes are often specific to particular geographic areas. By organising properties within these national and regional frameworks, the clustering acknowledges the unique energy needs and retrofit opportunities that vary across different parts of the UK. This approach ensures that any analysis or recommendations for energy efficiency improvements are relevant and applicable to the specific conditions of each region, thereby enhancing the effectiveness of targeted retrofit programmes.
	Scotland	-	
	Wales	-	
	Northern Ireland	-	
SAP region	Central Scotland, Orkney, Shetland	-	
	Midlands, Northern England	-	
	North England	-	
	North Scotland	-	
	Northern Ireland		
	Scottish Borders/Western Isles	-	
	South England	-	
	South West England	-	
	Wales	-	

Heating type	ASHP	-	Heating systems play a pivotal role in determining a property's energy consumption and environmental footprint. Recognising this, properties are further clustered according to the type of heating system they employ. This categorisation allows for a detailed examination of energy use patterns and the identification of retrofit solutions that are most appropriate for reducing emissions from different heating systems.
	GSHP	-	
	Electric boiler	-	
	Electric resistive	-	
	Electric storage	-	
	Gas boiler	-	
	Gas boiler (electric hot water)	-	
	Gas other	-	
	Gas other (electric hot water)	-	
	Biofuel boiler	-	
	Solid fuel boiler	-	
	Communal gas	-	
	Communal low carbon	-	
	Oil boiler	-	

Fuel poverty	Fuel poor	-	The clustering process incorporated the categorisation of homes based on their fuel poverty status. Recognising that fuel-poor homes often required specific retrofit solutions that are cost-effective yet impactful, these properties were grouped to identify retrofit pathways that could be feasibly implemented, possibly with the aid of government grants or subsidies. This focused approach ensured that interventions were not only economically viable for households in fuel poverty but also contributed significantly to reducing their energy bills and improving their living conditions.
	Not fuel poor	-	
Heritage status	Yes	-	The clustering process included listed buildings, conservation areas, and World Heritage Sites. This helps identify common characteristics affecting energy efficiency, retrofit potential, and regulatory constraints.
	No	-	

			Clustering by heritage status ensures that analysis accounts for specific planning restrictions, construction methods, and permitted interventions.

6. Archetype clustering

Hierarchical clustering was utilised to systematically organise properties into archetypes, starting with each archetype as its own cluster. This applied to the bottom 30% of archetypes when organised by useful energy, as mentioned above. However, where an archetype within the bottom 30% exceeded a minimum threshold of 500 properties, it was omitted from the clustering process.

Hierarchical clustering begins with each object in its own cluster and merges them into increasingly larger clusters based on their similarity, ultimately resulting in a single cluster that contains all objects.

1. **Establish design conditions:** As described in previous sections, this ensured that specific variables were not mixed together within clusters.
2. **Similarity measurement:** The process initiated by calculating the similarity, also known as 'distance', between properties in the dataset. The specific variables used for clustering were:
    a. Construction age
    b. Roof type
    c. Floor type
    d. Glazed type
    e. Property size
    f. Space constrained
    g. Location type
    h. Tenure

    These variables were chosen based on their relevance and importance to the clustering objective, informed by domain expertise, and their ability to capture significant characteristics of housing types. By focusing on these variables, it addressed issues related to dimensionality reduction, computational efficiency, and interpretability, while minimising noise and redundancy that would have resulted from adding additional variables like solar PV. Other variables were not as critical for identifying housing archetypes.

3. **Creating clusters:** With established similarities, the hierarchical clustering treated each property as an individual cluster. Iteratively, clusters with the highest similarity were merged, ensuring that combined properties shared significant characteristics.
4. **Building the hierarchy:** This merging continued, forming larger groups and developing a hierarchical tree structure, or dendrogram. This visual representation highlighted the similarity levels between properties, marking each merge point where clusters joined.
5. **Determining archetypes:** The hierarchical tree was then 'cut' at a specified level to identify distinct archetypes, chosen based on the analysis's desired detail and a minimum population threshold for meaningful contribution. This selection process ensured that each archetype represented a property

group with shared energy characteristics important enough for separate analysis or targeted retrofit interventions.

6. **Grouping remaining archetypes:** Archetypes falling below the population threshold were grouped using hierarchical clustering, with the cluster size determined by a silhouette score to measure similarity within a cluster versus other clusters. A maximum of three clusters per design condition was set to keep the total number of archetypes reasonable, although this maximum could be adjusted downward for fewer clusters.
7. **Cluster labelling:** Finally, clusters were assigned a 'cluster label', facilitating the creation of a complete archetype table. These labels allowed for the easy viewing, filtering, and grouping of archetypes according to their specific label, enhancing the manageability and interpretability of the data.

Through this hierarchical clustering process, properties were meaningfully categorised based on shared energy characteristics, balancing the depth of analysis with the manageability of the total number of archetypes by employing silhouette scores and cluster labels for efficient data organisation and analysis.

6.1 Further manual clustering

Once the clustering process was complete, the number of archetypes in total was still too high to ensure a manageable dataset that could be used by the CCC for their own modelling initiatives. The aim was to reduce the total number of final archetypes down from over 20,000 to around half that number, if possible. Applying a threshold stock level of 500 properties and analysing the makeup of the archetypes that fell below this threshold, it was determined that a large number of archetypes had less than 100 properties in their stock count. Based on the design conditions outlined earlier, these small archetypes were prohibited from being merged together with other archetypes. Further consolidation meant removing design conditions one by one in order to further group these small archetypes, until they breached the 500 property threshold. This process was reiterated many times on a trial-and-error basis to establish which design conditions to remove in which order of priority in order to both reduce the total number of archetypes and preserve overall data integrity. The result was a dataset of fewer than 12,000 archetypes. The total number of archetypes could not have been reduced any further without damaging data integrity.

6.2 Preservation of less common variable attributes

The QA processes highlighted that the clustering process (both machine learning and manual steps) resulted in some attributes dropping out of the dataset altogether.

Some attributes within the archetype dataset apply to a relatively small number of dwellings in the overall housing stock. A good example would be park homes, which account for a very small proportion of the total housing stock. Where the property type was park home, combined with location details and other variables, then the number of dwellings in each archetype would be so small that all archetypes were at risk of being merged with others having a more dominant property type. The result was that park homes in this example were missing from the final dataset for Northern Ireland, Scotland, and Wales. Aside from park homes, other affected attributes were ground source heat pumps, biofuel boilers, communal low carbon, electric boilers, gas boiler (electric hot water), gas other, gas other (electric hot water), solid insulated floors, and suspended insulated floors.

The workaround to this problem was to insert a preliminary clustering step into the process before any other clustering took place. By identifying archetypes with “at risk” attributes and manually merging them together with other archetypes having the same “at risk” attribute, the attributes inherited a more dominant presence in the dataset. In a similar way to the approach set out in 5.1, several iterations were tested by removing one design condition at a time, but only for archetypes that contained “at risk” attributes. The order in which each of the “at risk” archetypes were manually clustered also had an impact on the resulting data, which was totalled and compared against totals for each attribute in the Housing Surveys to ensure that data integrity was maintained.

The overall aim was to have archetypes for “at risk” attributes that exceeded the 500 stock threshold. This was achieved for some of the “at risk” attributes, which then became protected by their size from the rest of the clustering process, but not for all. This is simply because the stock levels for some attributes were too sparse regardless of manual clustering. For these, the threshold was lowered below 500 specifically for the “at risk” archetypes that would otherwise be merged out of existence. This explains why a small number of archetypes still exist that have very low stock counts, with the lowest being 18.

7. Validation and recalibration of archetypes against Housing Surveys

Validation: The generated archetypes from the EPC dataset were first validated by comparing them to survey data, providing a more accurate representation of the UK housing stock. Archetypes were created from the EPC data without regard to the Housing Survey data (see sections on archetype creation above). Both the archetype data and the Housing Survey data were then extrapolated so that the total dwellings match the UK’s dwelling stock, at a regional and national level. This comparison aimed to ensure that the distribution of normalised variables in the EPC dataset closely matched that in the Housing Survey data, thereby validating the archetype generation process.

Recalibration: Upon validating the archetypes, the project moved to recalibrate the archetype counts to better reflect the UK housing stock. This involved:

1. Assessing differences in counts between EPC and Housing Survey data to identify disparities in the distribution of normalised variables.
2. Establishing the survey's regional grossing factor totals as benchmarks for the total UK house count within each region. Working through each categorical variable in turn, the percentage split of stock between each attribute was compared between the archetype dataset and the Housing Survey dataset.
3. Creating scaling factors based on the Housing Survey counts by nation, property type, and tenure. These factors were used to adjust the count of properties within the EPC dataset to more accurately reflect the distribution found in survey data, providing a finer resolution of scaling factors.

On iterating through each categorical variable there will be cases where the stock count of an archetype is being adjusted multiple times, where multiple categorical variables need to be recalibrated. The split of dwelling stock across categorical variable attributes can therefore not yield an exact match against the totals in the Housing Surveys. The order in which categorical variables are adjusted has an impact on the closeness of the match between archetype totals and Housing Survey totals, and therefore the optimal sequence was established based on trial and error. The resulting dataset yielded a close match to survey totals while preserving the integrity of EPC data.

As an additional verification step, the percentages for each attribute were compared against published percentages such as those in the BRE's report on The Housing Stock of The United Kingdom, 2017.

8. Energy and emissions calculations

CO2 emissions were calculated from property energy demand. As energy demand data is not available in the EPC database, this needed to be calculated from EPC bill estimates. Bill estimates were converted to energy consumption based on Building Research Establishment (BRE Group) fuel specific cost per kWh data from the date of the EPC certificate. Total energy consumption was then converted into CO_2 using emissions intensity factors from that date. This process was used for each heating system.

Energy calculations were made twice, once before the granular heating systems in the EPC dataset were condensed into the final 14 heating system categories, and again afterwards. The reasoning for this approach was discussed and agreed with the CCC with benefits as follows:

- The precise fuel types would feed into the initial energy calculations, which was required to ensure that bill cost is accurately converted into kWh of energy consumption
- The precise heating system efficiencies would feed into the initial energy demand calculations, maintaining overall accuracy
- Recalculation against the condensed heating systems would ensure that useful energy would map directly to delivered energy and emissions based on the blended efficiency factors of the 14 condensed heating systems. This gives the CCC the flexibility to change efficiency and emission factors in their copy of the dataset without the need for Kamma to recompute the whole model.

The steps in the calculation sequence were as follows:

1. Calculate delivered energy from EPC bill estimates (section 8.1)
2. Calculate useful energy from granular EPC heating systems (section 8.2)
3. Condense granular heating systems into 14 condensed heating categories (section 8.3)
4. Generate archetypes (sections 4, 5, and 6)
5. Recalculate delivered energy against the 14 condensed heating categories (section 8.4)
6. Recalibrate delivered energy against ECUK data (section 8.5)
7. Recalculate useful energy against the 14 condensed heating categories using the ECUK recalibrated delivered energy numbers (section 8.6)
8. Derive direct delivered and direct useful energy from total numbers (section 8.7)
9. Calculate direct CO_2, N_2O and CH_4 emissions from the ECUK recalibrated delivered energy numbers (section 8.8)

8.1 Delivered energy (energy consumption) for detailed heating systems

The methodology for calculating energy consumption started with the extraction of estimated energy bill costs from EPC data. Energy prices utilised in SAP calculations, which were originally sourced from the BRE's Product Characteristics Database (PCDB) based on a rolling 3-year average, were updated biannually. On 15th February 2023, this pricing model was revised to more accurately reflect contemporary energy prices. Kamma had documented all SAP pricing adjustments, including those from the PCDB and the subsequent updates, ensuring that estimates of EPC energy bills could be accurately converted into kWh of delivered energy demand for individual properties, contingent on the lodgement date of the EPC.

In calculating energy demand, Kamma took into account the heating systems recorded in each EPC assessment. If an EPC outlined both a primary and a secondary system for space heating, Kamma adhered to the division of energy demand between the systems, as stipulated in the SAP 2012 guidance. If a standing fee was applicable, it was allocated proportionally among the primary, secondary, and hot water systems before the remaining cost was divided by the cost per kWh value relevant at the time the EPC was lodged.

The outcome was an annual kWh estimate for space heating and hot water heating, taking into consideration each heating system. This approach also allowed for the classification of energy demand by the type of fuel used, facilitating queries of energy demand by both fuel type (e.g., electric, gas, oil) and usage type (space heating, water heating, lighting).

8.2 Useful energy (energy demand) for detailed heating systems

Useful energy, defined as the energy produced after accounting for losses due to heating system inefficiencies, was calculated. Efficiency factors for each heating system were established in collaboration with the CCC, employing the efficiencies listed in SAP 10.2, as shown in the table below. These efficiency factors were multiplied by the delivered energy estimates from the previous step.

Heating system	**Space heating (%)**	**Hot water (%)**	**Hot water only (%)**
Gas boiler	84	84	74
Electric heating (excluding heat pumps)	100	100	100
ASHP	219	190	175
GSHP	260	230	175
Oil boiler	84	84	72
LPG boiler	84	84	74
Biomass boiler	65	65	65
Solid fuel boiler	65	65	65
Gas room heaters	58	n/a	n/a
Oil room heaters	60	n/a	n/a
Solid fuel room heaters	65	n/a	n/a
LPG room heaters	60	n/a	n/a
Gas boiler warm air	76	76	76
Communal gas	80	80	80
Communal electric	300	300	300
Communal biofuel	65	65	65
Communal waste heat	100	100	100
Electric with storage	n/a	n/a	100
Electric instantaneous	n/a	n/a	100

8.3. Condense granular heating systems

The granular EPC heating systems were mapped to 14 condensed heating systems in collaboration with CCC. The mappings are fully documented in section 2.3.2.4.

8.4. Recalculate delivered energy for condensed heating categories

At this point, the archetypes have been generated as per sections 4, 5, and 6. Each archetype has a condensed heating category and a total value for useful energy. This useful energy was converted back to delivered energy by dividing the useful energy by the efficiency factor for each of the 14 condensed heating categories. Because some of these categories (e.g. "Gas other" and "Communal low carbon") contained a mix of different heating systems and fuel types, a weighted average efficiency factor was calculated for each of the mixed categories and agreed with the CCC, based on the percentage split of heating systems and fuel types within the category. The final efficiency factors that were used were as follows:

Condensed heating category	**Space heating (%)**	**Water heating (%)**
Gas boiler	84	84
Oil boiler	76	76
Electric boiler	100	100
ASHP	219	190
GSHP	260	230
Communal gas	80	80
Electric resistive	100	100
Electric storage	100	100
Gas boiler (electric hot water)	84	100
Gas other	65	84
Gas other (electric hot water)	65	100
Biofuel boiler	65	65
Solid fuel boiler	65	65
Communal low carbon	168.7	168.7

8.5. Recalibrate delivered energy against ECUK data

The CCC needed the totals for delivered energy across the archetypes datasets to equate to the totals published in the ECUK (Energy Consumption in the UK) report for 2019. The "End Use Tables" published in September 2023 were used for this purpose, using the totals for 2019 from that report. The end use tables split energy consumption into end use (e.g. space heating, water heating, cooking, appliances) and also into fuel types (e.g. natural gas, oil, electricity). For each permutation of end use and fuel type within the ECUK report (Table U2), the total energy was converted from ktoe (kilotonnes of oil equivalent) into kWh and then compared against the total kWh across the archetype dataset. This yielded an extrapolation factor to apply against each of these end use/fuel type permutations in order to make the totals match. In general there was a less than 10% variance between archetype totals and ECUK totals, with a couple of exceptions, namely biofuel and to a lesser degree solid fuel, where the totals varied more between datasets.

For solid fuel boilers the amount of energy allocated to water heating in the ECUK report formed a very small percentage of the overall delivered energy for the fuel type (around 3.8%, whereas the EPC data would suggest anywhere between 15% and 25%). For this reason it was agreed to recalibrate the total delivered energy for solid fuel based on the ECUK total for space plus water heating, but then split the total into space and water heating values based on the average ratio in the EPC data.

All archetype totals were recalibrated against ECUK totals except for biofuel heating where the ECUK total was some 2.4 times that in the archetype dataset. This is understood to be because ECUK includes estimates of domestic wood burning which are not captured in the EPC data. In this case the archetype totals were preserved, as agreed with the CCC.

8.6. Recalculate useful energy for the condensed heating systems

Once the ECUK recalibration was complete and the delivered energy estimates were effectively finalised, it was a straightforward task to recalculate useful energy demand from the recalibrated totals for each archetype by multiplying with the efficiency factors from the table in section 8.4.

8.7. Derive direct delivered and direct useful energy

Direct energy refers to the energy utilised for heating spaces and water within a property, generated on-site. This category excludes electrical energy generated off-premises. Direct energy numbers were derived by referring to the fuel type in place for space heating and hot water, using the 14 condensed heating categories. Where the fuel type was electricity, the energy demand was filtered out of the “direct” totals.

8.8. Calculate direct CO_2, N_2O and CH_4 emissions

Emissions were calculated from the direct delivered energy totals, referring to the fuel type(s) used for space heating and hot water heating, using the 14 condensed heating system categories. This was done by multiplying the direct delivered energy total by the relevant emissions factor from the DESNZ “Greenhouse gas reporting: conversion factors 2023” dataset. The resulting emissions numbers are expressed in kg per property per year.

8.9. Unregulated energy

The considerations and calculations previously discussed were applicable solely to regulated energy. Regulated energy encompasses the building's energy consumption attributed to specified, controlled building services and installations. This includes space heating and cooling, hot water, ventilation, as well as the operation of fans, pumps, and lighting – all of which were inherent to the building's design.

In contrast, unregulated energy pertains to the energy consumption by the property's occupants for the operation of electrical appliances, along with cooking and catering activities. This type of energy usage is distinct from the building's inherent energy needs.

8.9.1 Energy from electrical appliances

For estimating the energy consumption associated with electrical appliances, a formula from the SAP 2012 methodology (Appendix L2) was employed. This formula provided a standardised approach to quantifying the energy used by electrical appliances within a property:

$$EA = 207.8 \times (TFA \times N)\ 0.4714$$

...where TFA is the total floor area in m² and N is the assumed number of occupants.

The first step was to estimate the number of occupants of the property, which comes from a formula for assumed occupants provided in SAP 2012 Table 1b, and is as follows:

$$\text{if } TFA > 13.9: N = 1 + 1.76 \times [1-\exp(-0.000349 \times (TFA-13.9)^2)] + 0.0013 \times (TFA-13.9)$$
$$\text{if } TFA <= 13.9: N = 1$$

...where N is the assumed number of occupants, TFA is the total floor area of the dwelling

Energy for appliances was recalibrated against the ECUK totals using the same approach as in section 8.5.

8.9.2 Energy from cooking and catering

SAP 2012 gives a formula for estimating the CO_2 emissions that come from cooking and catering (Appendix L3). Again, these are based on an assumed number of occupants (see above). The formula is as follows:

$$(119 + 24\ N) / TFA\ (L16)$$

...where TFA is the total floor area in m² and N is the assumed number of occupants.

From CO_2 emissions, estimates of the kWh of energy responsible for generating those emissions were derived. This estimation process was based on broad assumptions, due to the inability to discern from the EPC data the types of fuel utilised for cooking and catering. It was presumed that properties not connected to the gas grid relied exclusively on electricity for cooking and catering. Conversely, properties with gas grid access were assumed to utilise 70% electricity and 30% gas, based on the percentage split between homeowners with gas ovens versus electric ovens in the BRE's "Report 9: Domestic appliances, cooking & cooling equipment". Utilising these general assumptions, current carbon intensity factors were applied to electricity and/or gas to estimate the CO_2 volumes and determine the kWh of energy consumption, distinguishing between direct and indirect energy demand. Here, energy consumption from cooking is the total of direct and indirect energy, whereas energy consumption from cooking using non-electric sources is just direct energy consumption, calculated using the ratio above.

Energy for cooking and catering was recalibrated against the ECUK totals using the same approach as in section 8.5.

---

kammaclimate.com

KAMMA CLIMATE