Technical appendix

Using smart meters to identify energy use profiles

Sofia Pinto and Roisín Gorman

July 2025

Contents

1. Motivation

Nesta's data science team, embedded within the sustainable future mission, has analysed smart meter data to identify energy-use profiles, *i.e.* groups of households who consume energy in similar ways, and how they differ in household characteristics, appliances, physical property characteristics and demographic factors. This technical appendix provides detailed information about the methodology used in this project. To read about the results and next steps, read our report, explore the energy-use profiles explorer, and share your thoughts via our feedback form.

2. Data source

The analyses presented in this document were conducted using Smart Energy Research Lab (SERL) observatory data[1], containing longitudinal smart meter electricity and gas data for over 13,000 households in Great Britain. The data is accessible through the UK Data Service SecureLab by accredited researchers.

3. Smart meter data features

To create energy usage profiles, we need to find similarities between energy usage patterns across households. To make this possible, we created features from smart meter data that will enable the methods we use to identify these underlying patterns.

Where possible, features should map to behaviours/lifestyles or property characteristics, as this allows for interpretability of results. In Table 1 below, you can find features we've developed and the characteristics they map to.

Table 1. Smart meter features and corresponding behaviours/lifestyles

Smart meter features	**Behaviours and lifestyles and property characteristics**
Average daily electricity usage Average daily gas usage	Household demographics Affordability of energy for a household
Ratio between winter and summer electricity usage Ratio between winter and summer gas usage	Type of heating system installed in the home Property features (e.g. quality of insulation, house size)
Average daily electricity usage from 9am to 5pm Average daily gas usage from 9am to 5pm	Households with occupants doing home working Household composition (e.g. parents on parental leave, presence of retired/older or disabled occupants) Working status of occupants
Average (and proportion of) daily electricity overnight Average (and proportion of) daily gas overnight	Electric vehicle charging or storage heating Woking patterns
Average (and proportion of) daily electricity usage at specific times of day: morning, lunch and dinner	Household composition Occupancy behaviour

Average (and proportion of) daily gas usage at specific times of day: morning, lunch and dinner	
Average daily usage during weekends (or weekends and bank holidays) for electricity and gas	Working patterns Presence of high-energy assets
Standard deviation of daily electricity and gas usage	Presence of appliances and high-energy assets
Average daily magnitude of electricity and gas usage	Presence of appliances and high-energy assets
Inter-day variability of energy usage	Occupancy behaviour Affordability of energy for a household Tariff type of household Presence of solar panels and /or battery storage

4. Contextual features about households and properties

Energy-use profiles are identified using energy consumption data only. After the profiles are created, contextual information about household demographics/behaviours, as well as characteristics about their properties, is brought in to find more about the households in each profile.

Table 2 briefly describes the contextual features currently used in our prototype. For more information on these contextual features and how they're computed, visit methodology subsubsection 5.8.

Table 2. Contextual features information

Contextual features	Description
Heat pump	Whether or not the household has a heat pump.
Tenure	Whether the house is a social or private rental, or owner-occupied.
Solar panels	Whether or not the household has solar panels.
Electric vehicle charger	Whether or not the household has an electric vehicle charger.
Battery storage	Whether or not the household has battery storage.
Air conditioning unit	Whether or not the household has an air-conditioning unit.
Gross household income	Gross annual income bracket of household.
Plug-in electric vehicle ownership	Whether or not the household has a plug-in electric vehicle.
Central heating fuel	Main fuel type of central heating system.
Smart central heating controls	Whether or not the household has smart central heating controls in the property.
Property type	Type of property selected from the following options: a house or bungalow that is detached; semi-detached; or terraced; or a flat, maisonette, or apartment.

Property age band	Build year band of property.
Number of occupants	Total number of occupants reported in the property.
Household composition	Number of occupants in each age band and gender.
Working status of occupants	Whether or not the household has children, or adults aged 65 years or over only
Region	Region in Great Britain.
Index of Multiple Deprivation (IMD) quintile	IMD quintile for the LSOA in which the household is located.
EPC rating	Whether property has an EPC rating of C and above
Time-of-use tariff	Whether household is on a time-of-use tariff

5. Methodology

5.1. Pipeline overview

Figure 1 provides an overview of the data science pipeline developed to identify energy-use profiles and relevant information about households and homes in each profile.

Each of the steps in this pipeline is described in more detail in the remaining subsections of section five.

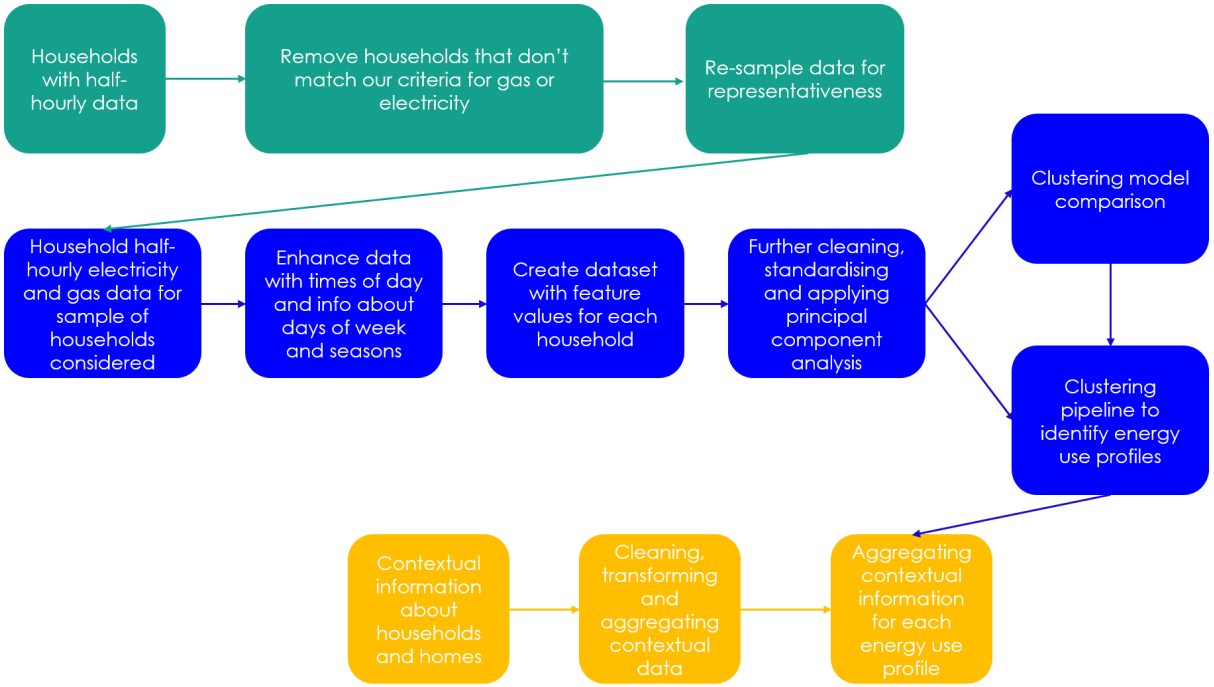

Figure 1. Overview of the data science pipeline

5.2. Households considered in the analysis

Step 1: Households with sufficient valid reads

We take valid half-hourly electricity consumption data in the 12 month period from July 2023 to June 2024, and remove households that don't match the following criteria:

- Households with valid reads in every month considered;
- Households with 15 or more valid half-hourly reads for every half-hour in a month (out of 27, 28, 30 or 31 per month);
- Households with 40 or more valid half-hourly reads each day (out of a maximum 48, one per each half-hour).

Separately, we do the same for gas. The results are two lists of PUPRNs – one for households with electricity smart meter data that matches our criteria and another for households with gas smart meter data that matches our criteria.

Step 2: On- and off- gas grid households

Not all households with electricity smart meter data will have gas meter data available due to a variety of reasons including 1) the house is off-gas grid and 2) the house is on the gas grid, but no gas smart meter data is available.

Because we want to cluster based on both electricity and gas data, we need to have a set of PUPRNs representing households that:

- are on the gas grid and for which we have both electricity and gas data available and with sufficient reads.
- are off the gas grid, but for which electricity data is available and with sufficient reads.

Using participant data and EPC, and our lists of PUPRNs from the previous set we identify households that match one of the above criteria.

Step 3: Preserving representativeness

After identifying one list of PUPRNs in step 2, the sample is corrected for IMD quintiles while preserving regional proportions. This is done to maintain the representativeness of the original sample. Note that this step might lead to different results if run multiple times, due to the stochastic nature of the process.

Step 4: Households with infinite values

Smart meter data for the households identified in step 3 is aggregated into features (read more in the next section). These smart meter features dataset is then put through clustering, to identify groups of households using energy in similar ways. Some of the features created are ratios, for example the ratio of electricity usage between winter and summer. If any of these ratio values is infinity (due to having 0 in

the denominator) then we need to remove the respective households) as the clustering models used can't deal with infinity.

The SERL dataset contains data on 13,209 households. After removing households not following the criteria above, we get 5,994 households. These are the households used in our analyses.

5.3. Data cleaning and data preparation before clustering

The half-hourly data needs to go through cleaning and enhancements before applying the clustering pipeline:

1. Importing the valid half-hourly electricity readings for the households considered in the analysis.
2. Enhancing the half-hourly electricity data with information about days of the week, weekend vs weekdays, seasons and times of day ('overnight', 'morning', 'lunch', 'dinner' and '9:00 to 5:00' - see below for more information on these times of day).
3. Creating a dataset with features for the households considered in the analysis by aggregating the half-hourly data appropriately (this dataset consists of one line per household and as many columns as there features).
4. Removing households with infinity values, resulting from computing ratios (eg, when computing the ratio between winter and summer usage, if average summer usage is zero).
5. Data standardisation.
6. Applying principal component analysis to create a new dataset of features which are uncorrelated with each other. We choose the number of components required to explain at least 99% of the variance in the data.

The times of day periods we're considering are as follows:

- Overnight: from midnight to 4:00am, both inclusive.
- Morning: from 6:00am to 10:00am, both inclusive.
- Lunch: from 12:00pm to 2:00pm, both inclusive.
- Dinner: from 5:00pm to 9:00pm, both inclusive.
- 9:00 to 5:00: from 9:00am to 5:00pm, both inclusive.

You can read more about the features created from electricity smart meter data in the following section.

5.4. Smart meter data features

We've computed smart meter features described in Table 5 below, where `meter_type` can take electricity and gas as values.

Table 5. Feature names and how they are computed

Smart meter features	**Feature name**	**How the feature was computed**
Average daily gas / electricity usage	avg_daily_{meter_type}_usage	Mean of average daily gas / electricity usage for each PUPRN.
Ratio between winter and summer gas / electricity usage	ratio_winter_summer_{meter_type}_usage	Mean of average daily gas / electricity usage in winter and summer are computed for each PUPRN. The ratio between the winter and summer values is then computed.
Average (and proportion of) daily gas / electricity usage from 9am to 5pm	avg_daily_{meter_type}_9_to_5 avg_prop_daily_{meter_type}_9_to_5	Mean of average daily gas / electricity usage for each PUPRN, during the period between 9am and 5pm. The proportion consists of the average of the daily gas / electricity usage used between 9am and 5pm

		divided by the total daily gas / electricity usage.
Average (and proportion of) daily gas / electricity usage overnight	avg_daily_{meter_type}_overnight prop_daily_{meter_type}_overnight	Same as above but for the overnight period.
Average (and proportion of) daily gas / electricity usage at specific times of day: morning, lunch and dinner	avg_daily_{meter_type}_morning prop_daily_{meter_type}_morning avg_daily_{meter_type}_lunch prop_daily_{meter_type}_lunch avg_daily_{meter_type}_dinner prop_daily_{meter_type}_dinner	Same as above for the morning, lunch and dinner periods.
Average (and proportion of) daily gas / electricity usage during weekends	avg_daily_{meter_type}_weekend prop_daily_{meter_type}_weekend	Same as above but for the weekends.
Standard deviation of daily gas / electricity usage	std_daily_{meter_type}_usage	Standard deviation of average daily gas / electricity usage for each PUPRN.

Average daily magnitude of gas / electricity usage	avg_daily_magnitude_{meter_type}	Daily magnitude is computed for each day per PUPRN. We then take the average daily magnitude per PUPRN.
Inter-day variability of gas / electricity usage	{meter_type}_avg_mean_abs_diffs_over_year {meter_type}_avg_mean_pc_diffs_over_year	Mean of the absolute differences in gas / electricity usage between consecutive daily average usage. For `pc_diffs`, we calculate the mean of the percentage differences between consecutive days.

5.5. Model comparison

We've compared two clustering methodologies, k-means and gaussian mixture models (GMMs), using implementations from the Scikit-learn Python package. We have also compared results for all numbers between 2 to 20 clusters. Results were compared using the silhouette metric, by computing the average silhouette value across 10 runs of the same hyper-parameters (the default hyper-parameters in combination with the number of clusters being tested).

5.6. Clustering pipeline

The final results were produced using k-means clustering, as per the results from our model comparison pipeline.

5.7. Creating contextual information about profiles

Contextual information is brought in for each household in the sample. This is done through the unique pseudo-anonymised household identifier available both in the feature dataset, as well as in all the sources of contextual information. However, not all households have contextual information available, and some have this but not for

all possible contextual features. When unavailable, the contextual information appears as 'Unknown'.

After having contextual information matched to households, it can then be aggregated by profile to identify trends. You can find more about the contextual information in the following section.

5.8. Contextual features about households and properties

This subsection provides extensive information about the contextual information used to identify trends about households and property characteristics of households in each energy usage profile.

The following datasets are available for contextualising the resulting clusters:

- participant summary data
- participant survey data
- follow-up survey data
- EPC data
- tariff data
- smart meter consumption data.

Table 6 summarises the contextual features computed as part of the prototype resulting from the first phase of work and respective source.

Table 6. Summary of contextual features

Contextual feature	Description	Source
Plug-in electric vehicle ownership	Whether or not the household has a plug-in electric vehicle. Options: • Yes • No • Unknown	Participant survey
Electric vehicle charger	Whether or not the household has an electric vehicle charger.	Follow-up survey

Battery storage	Whether or not the household has battery storage.	Follow-up survey
Air conditioning unit	Whether or not the household has an air-conditioning unit.	Participant survey
Heat pump	Whether or not the household has a heat pump.	Follow-up survey and EPC data
Tenure	Whether the house is a social or private rental, or owner-occupied.	Follow-up survey and EPC data
Gross household income	Gross annual income bracket of the household (not adjusted for inflation).	Follow-up survey
Central heating fuel	Main fuel type of central heating system. This feature was produced by combining EPC and participant survey data. If valid heating fuel data is available from the EPC dataset for a household, this is used preferentially over heating fuel self-reported by the household in the participant survey.	EPC data and participant survey
Central heating controls	Type of central heating controls in the property.	Participant survey
Property type	Type of property selected from the following options: a house or bungalow that is detached; semi-detached; terraced; a flat, a maisonette, or an apartment.	Participant survey
Property age band	Property build year band.	Participant survey
Number of occupants	Total number of occupants reported in the property. Number of occupants are reported as integers and we have grouped them where relevant.	Participant survey
Household composition	Participants report the number of occupants in each age band and gender. We have processed this	Participant survey

	information to group households into four categories: adult(s) only (not all aged 65+); adult(s) aged 65+ only; adult(s) and child(ren); and unknown.	
Working status of occupants	Participants report the number of adult occupants with each working status. We have processed this information to group households into four categories: all working and/or students; all not working; mix of working and/or students and not working; unknown.	Participant survey
Presence of solar panels	Presence of solar panels. When data on solar panels is unavailable, it is assumed using electricity export data. Households with total annual active electricity export greater than 0kWh are assumed to have electricity exports.	Participant survey, EPC and smart meter data
Region	Region in Great Britain.	Participant summary
IMD quintile	IMD quintile for the LSOA in which the household is located. Quintiles 1 (most deprived) to 5 (least deprived).	Participant summary
EPC rating	Whether EPC rating is C or above	EPC data
Tariff type	Whether households are on a time-of-use tariff	Tariff data

6. Analysis

In this section, we provide additional information about the analyses conducted using the methodologies highlighted in section five for the purpose of the first phase prototype. The analysis presented in this file was done using the SERL's seventh edition data, covering half-hourly electricity data from January 2019 to June 2024. **For the purpose of this analysis, we focused on the period between July 2023 and June 2024.**

6.1 Households considered in the analysis: sample representation

SERL data covers 13,209 households in Great Britain, constituting a representative sample of households with respect to region and income decile. Our analysis includes 5,994 of those households (45% of households).

Figure 2 shows the distribution of households in the SERL study and in our sample, by IMD quintile. The highest difference is found for IMD quintile 5 (the least deprived) with 17.95% households in the SERL study and 20.29% in the sample analysed. This is followed by IMD quintile 1 (the most deprived), with 20.8% of households in SERL's study and 19.6% in the sample analysed. In summary, our sample has less households in the most deprived quintile and more households in the least deprived quintile. Note: we haven't assessed if differences are statistically significant.

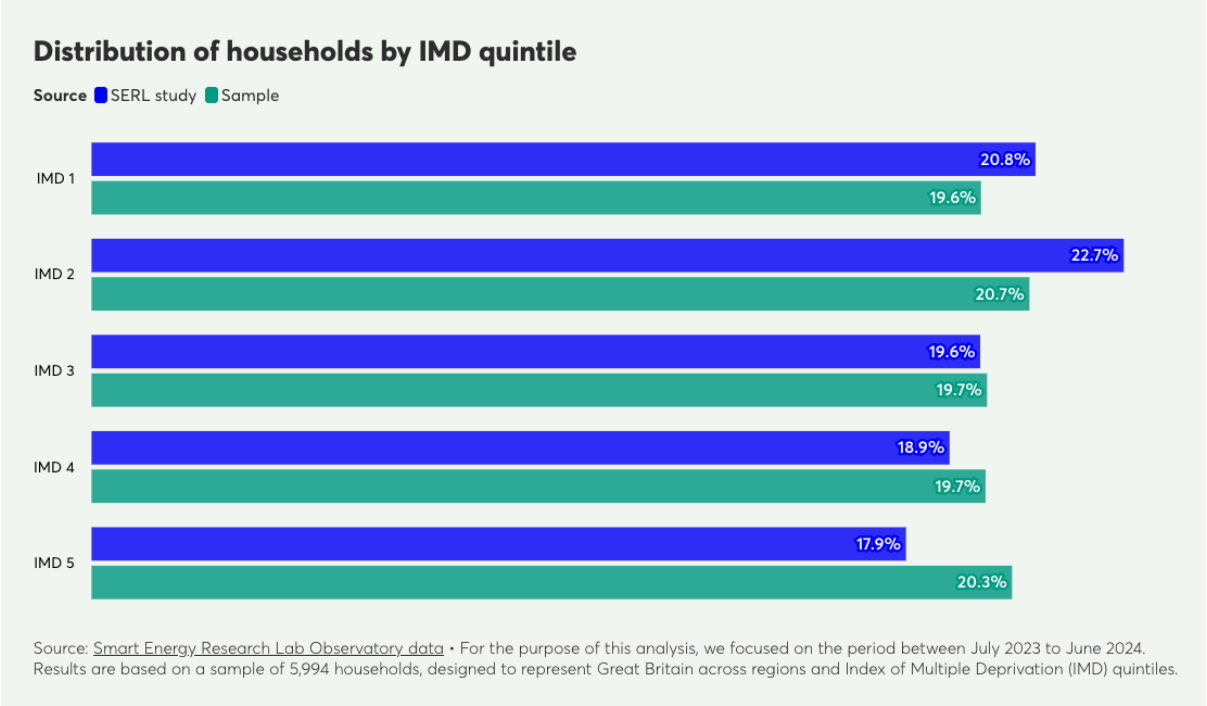

Figure 2. Distribution of households by IMD quintile in SERL households and sample used in analysis

Figure 3 shows the distribution of households in the SERL study and in our sample, by region. The highest difference is found for Scotland, with 9.7% households in SERL's study and 7.3% in the sample analysed. This is followed by the South East with 12.9% of households in SERL's study and 14.0% in the sample analysed. In summary, the highest differences show that SERL's study proportionally has more households in Scotland and less houses in the South East, when compared to the sample of

households analysed. Note: we haven't assessed if differences are statistically significant.

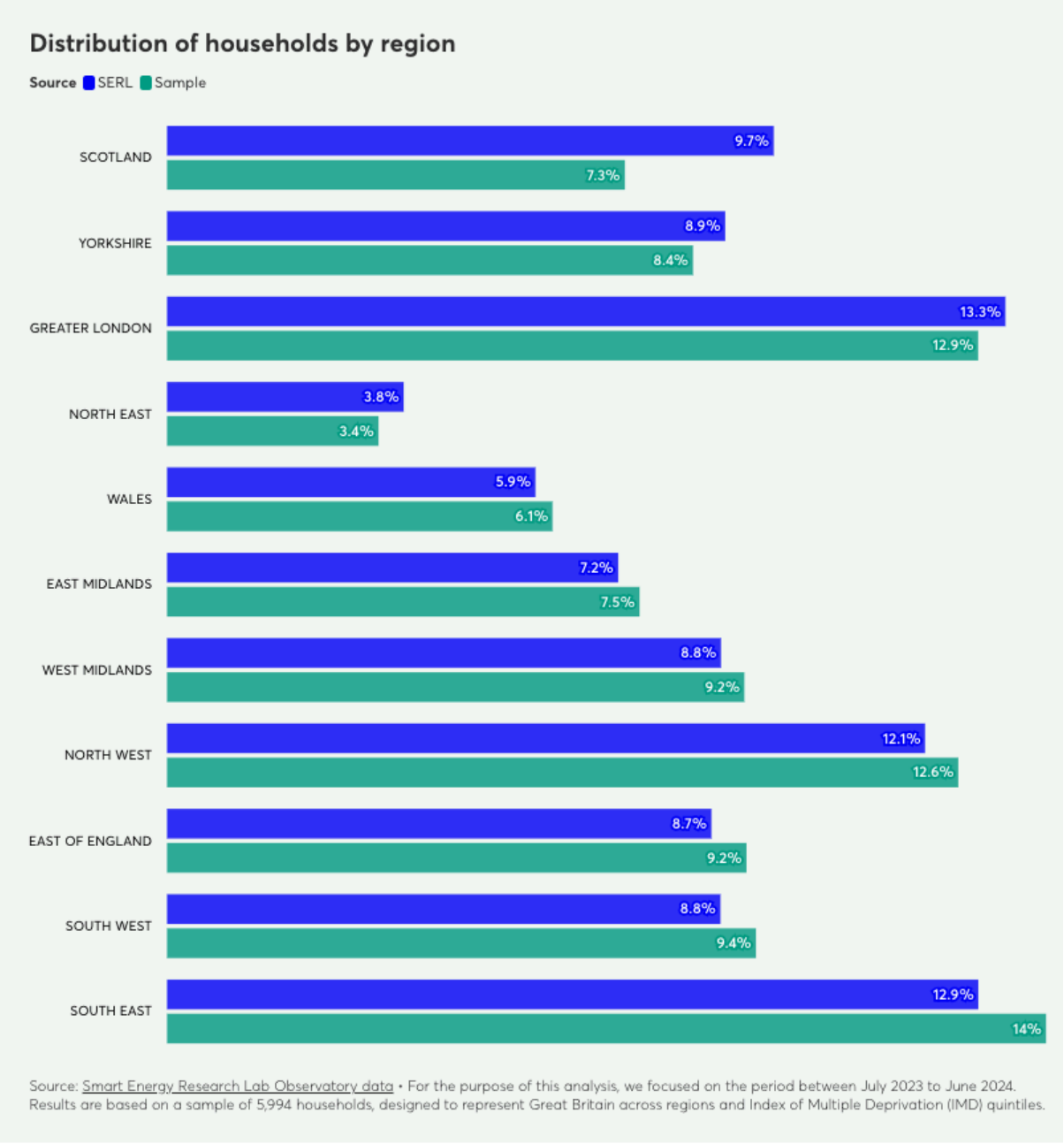

Figure 3. Distribution of households by region in SERL households and sample used in analysis

Data citation

1. Elam, S., Few, J., McKenna, E., Hanmer, C., Pullinger, M., Zapata-Webborn, E., Oreszczyn, T., Anderson, B., Department for Levelling Up, Housing and Communities, European Centre for Medium-Range Weather Forecasts, Royal Mail Group Limited. (2024). *Smart Energy Research Lab Observatory Data, 2019-2024: Secure Access*. [data collection]. *8th Edition.* UK Data Service. SN: 8666, DOI: http://doi.org/10.5255/UKDA-SN-8666-8