 doi:10.20944/preprints202109.0370.v1

Article

The SERL observatory dataset: longitudinal smart meter electricity and gas data, survey, EPC and climate data for over 13,000 GB households

Ellen Webborn[1*], Jessica Few[1], Eoghan McKenna[1], Simon Elam[1], Martin Pullinger[1], Ben Anderson[2], David Shipworth[1] and Tadj Oreszczyn[1]

[1] UCL Energy Institute, 14 Upper Woburn Place, University College London, WC1H 0NN, UK;
[2] Faculty of Engineering and Physical Sciences, University of Southampton, Southampton, UK
* Correspondence: e.webborn@ucl.ac.uk;

Abstract: The Smart Energy Research Lab (SERL) Observatory dataset described here comprises half-hourly and daily electricity and gas data, SERL survey data, Energy Performance Certificate (EPC) input data and 24 local hourly climate reanalysis variables from the European Centre for Medium-Range Weather Forecasts (ECMWF) for over 13,000 households in Great Britain (GB). Participants were recruited in September 2019, September 2020 and January 2021 and their smart meter data are collected from up to one year prior to sign up. Data collection will continue until at least August 2022, and longer if funding allows. Survey data relating to the dwelling, appliances, household demographics and attitudes was collected at sign up. Data are linked at the household level and UK-based academic researchers can apply for access within a secure virtual environment for research projects in the public interest. This is a data descriptor paper describing how the data was collected, the variables available and the representativeness of the sample compared to national estimates. It is intended as a guide for researchers working with or considering using the SERL Observatory dataset, or simply looking to learn more about it.

Keywords: smart meter data, household survey, EPC, energy data, energy demand, energy consumption, longitudinal, energy modelling, electricity data, gas data

1. Introduction

In 2019 the residential sector was responsible for 21% of UK carbon emissions [1] and the UK has committed to reaching net zero carbon emissions by 2050. Understanding domestic energy consumption now and going forward will support the transition to net zero. For example, electrification of heating and transport will become increasingly common and this will require managing increases and time shifts in energy demand [2]. Similarly, understanding the potential flexibility of demand response will support future energy system operation [3]. Identifying households in fuel poverty, or those at risk of missing out on green policy incentives, would help to ensure a just transition that benefits sections of society who might otherwise be left behind [4]. To understand these and other related issues, a nationally-representative empirical database of building, occupant and high-resolution energy data is imperative [5]–[7].

Historically, gathering large samples of energy consumption has been expensive and time-consuming, requiring manual read-outs of traditional energy meters (typically performed once or twice a year), or installation of specific monitoring equipment [7]. However, the smart meter rollout in the UK is ongoing and so far over 22 million smart meters have been installed in homes in Great Britain (GB), amounting to 44% of all domestic meters [8]. In the UK, smart meters record consumption every half hour and can store up to 13 months of historic import data. This data can be accessed via the Data Communications Company (DCC) Gateway; a centralised system, potentially offering a significant advance on the traditional metering or monitoring approach for gathering energy consumption data.

A Department for Business, Energy and Industrial Strategy (BEIS) policy paper demonstrates the UK government's understanding of the need for good-quality energy data throughout the energy system: "digitalisation will enable

 doi:10.20944/preprints202109.0370.v1

millions of low carbon assets, including solar PV, electric vehicles and heat pumps to be optimised across our energy system" [9]. The BEIS strategy focuses on system data and explicitly excludes consumer data which requires informed consent for collection. Despite the potential for using smart meter data in energy consumption research, obtaining householders' informed consent and ensuring compliance with General Data Protection Regulation (GDPR) and data governance requirements laid out in the Smart Energy Code (SEC) [10] is a potentially onerous task, reducing the extent to which the potential benefits of the smart meter rollout to energy researchers can be practically realised.

The aim of the Smart Energy Research Laboratory (SERL) is to provide UK researchers with high-quality, contextualized, half-hourly energy data for a representative sample of GB households in a secure environment that meets all data GDPR and SEC requirements; this is called the SERL Observatory dataset. SERL also has a 'Laboratory' capability where researchers can recruit their own intervention sample of households and use the Observatory as a counterfactual. These Laboratory datasets are not part of this paper; further information about the Laboratory function of SERL can be found on the SERL website [11].

Table A1 in Appendix A summarises seven UK energy datasets that may be used to investigate granular (daily or finer) patterns of UK domestic energy use at the household level. In terms of sample size, after the SERL Observatory dataset, the Customer Led Network Revolution dataset [12], [13] is the largest with up to around 3 years of half-hourly electricity data for approximately 11,000 participants in GB. Some of the other datasets provide higher-resolution energy data; every 15 mins (Wh) and 10 seconds (W) in the case of the SAVE project's representative sample of 4,000+ households from the South of England [14], and per second in the case of IDEAL Household Energy Dataset [15]. Various additional data, such as indoor temperatures, weather, demographic information, and EPC data are also included with some datasets. However, what has previously been lacking for UK researchers is a large, nationally representative, longitudinal sample with contextual data linked at the household level [5], [16]. The SERL Observatory is also unique in collecting data before, during and after COVID-19 lockdowns, which allows for the impact and potential development of a 'new normal' to be investigated.

In this paper we report on the Smart Energy Research Laboratory (SERL) Observatory dataset, which has recruited over 13,000 GB households who have consented to collection of their smart meter data for research purposes. The usefulness of energy consumption data is greatly enhanced by linking with other relevant contextual data and the SERL participants have provided their informed consent for their smart meter data to be linked to other relevant datasets. The SERL Observatory primarily consists of six core datasets: electricity and gas smart meter data, location data (such as region, Lower Super Output Area (LSOA) and Index of Multiple Deprivation quintile), weather data, survey data and energy performance certificate (EPC) data (Figure 1). These are linked at the household level, and accompanied by auxiliary data created by the University College London (UCL) SERL team, and described in the supplementary files with this paper. It is also possible for researchers to bring in data for linking at a local level, subject to approval by the SERL Data Governance Board. The combination of high-resolution energy and linked contextual datasets provides a detailed data resource facilitating exploration of co-variates and drivers of energy consumption as well as the potential implications of energy policies. The SERL Observatory dataset is freely available to accredited researchers[1] in the UK provided they comply with the relevant ethics and data governance processes.

[1] https://ukdataservice.ac.uk/help/secure-lab/training-requirements/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

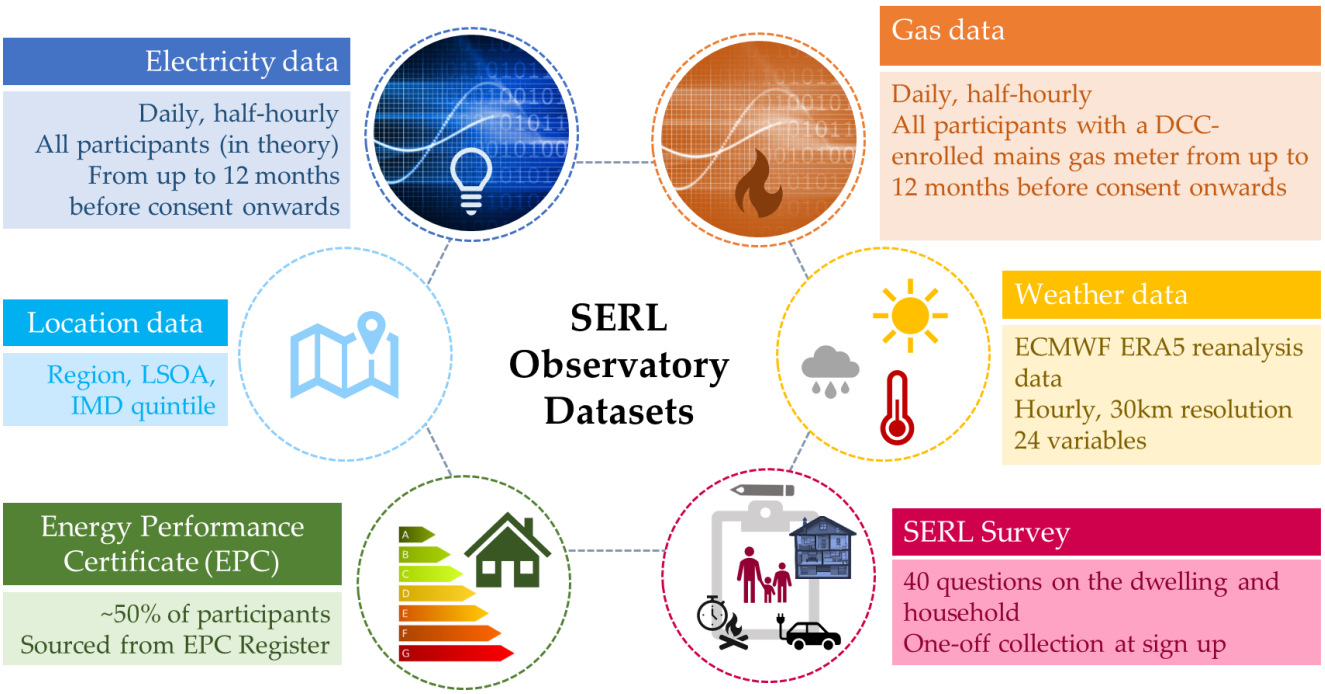

Figure 1: Overview of the SERL Observatory datasets.

Compared to the studies listed in Table A1 in Appendix A, the advantages of the SERL Observatory dataset are the combination of the higher number of participants (~13,000) from across GB, the inclusion of gas data (for around 10,000 households), the longitudinal collection, and significant contextual data linked at the household level. Table 1 shows the number of participants by recruitment wave (details in Section 2.2) and with different types of data available. Nearly 12,980 participants (97%) have provided detailed survey data relating to their dwelling, appliances and household members. 52% of participants have EPC data with 80 variables relating to their dwelling. All participants have region, LSOA and IMD quintile data and 24 climate reanalysis variables for every hour at locations within 30km of each dwelling.

Table 1. Number and percent of participants in the SERL Observatory with different types of data. By 'active' we mean that they have not withdrawn consent or moved house (i.e. data collection remains ongoing). Note that we do not include participants who were recruited but withdrew consent too soon for smart meter data collection to begin, or for whom smart meter data collection is not possible for technical reasons.

Participants…	***Number***	***Percent (%)***
… recruited	13,321	~
… recruited in wave 1 (Aug/Sept 2019)	1,708	12.8
… recruited in wave 2 (Aug/Sept 2020)	3,169	23.8
… recruited in wave 3 (Jan/Feb 2021)	8,444	63.4
… with an electricity smart meter	13,320	100.0
… with an electricity smart meter with any valid data and still 'active' on 31st May 2021	12,823	96.3
… with a gas smart meter	10,202	76.6
… with a gas smart meter with any valid data and still 'active' on 31st May 2021	9,730	73.0
… with SERL survey data	12,977	97.4
… with EPC data	6,921	52.0
… with weather data	13,321	100.0
… with region known	13,321	100.0
… with LSOA known	13,321	100.0
… with IMD quintile known	13,321	100.0

The aims of this paper are to explain how the SERL Observatory dataset was created, describe the data available including the results of data quality analysis, assess the representativeness of the SERL Observatory sample, and explain how researchers can access the data. Section 2 describes the research design and participant recruitment, Section 3

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

describes the datasets included in the SERL Observatory dataset, Section 4 reports the data quality analysis, Section 5 considers sample bias, Section 6 advises researchers how to apply for data access, and Section 7 concludes.

2. Research Design

2.1. Overview

The SERL system diagram is shown in Figure 2 Smart meter data is recorded by the electricity or gas meter in the home and sent to the Communications (Comms) Hub, also within the home, via the home area network (HAN) where it is stored for up to 13 months[2]. The Data Communications Company (DCC) Gateway is a messaging service which allows approved DCC Users to query a Comms Hub for which they have explicit consent from the occupants; further details of participant recruitment are given in Section 2.2. Some types of user are able to modify information on the Comms Hub (such as energy suppliers updating tariff information), but so-called 'Other Users' such as SERL are restricted to a reduced set of information. SERL sends data requests to participants' Comms Hubs via our DCC Adaptor Service, which returns data received along with any error messages/alerts. Consumer consents are stored on one database and smart meter data on an OLTP (Online Transaction Processing) database. Data then undergo processing and data quality analysis/error flagging by the SERL team, along with the contextual datasets which are brought into the research portal network and linked at the participant level by a pseudo-unique property reference number (PUPRN). Periodically the latest processed datasets will be made available to accredited researchers on approved research projects in a secure-lab environment. Researchers in the UK can apply to access the data, and following statistical disclosure control (SDC) checking by the UK Data Service (UKDS), approved outputs can be released and made publicly available; further details of these stages are given in Section 2.3.

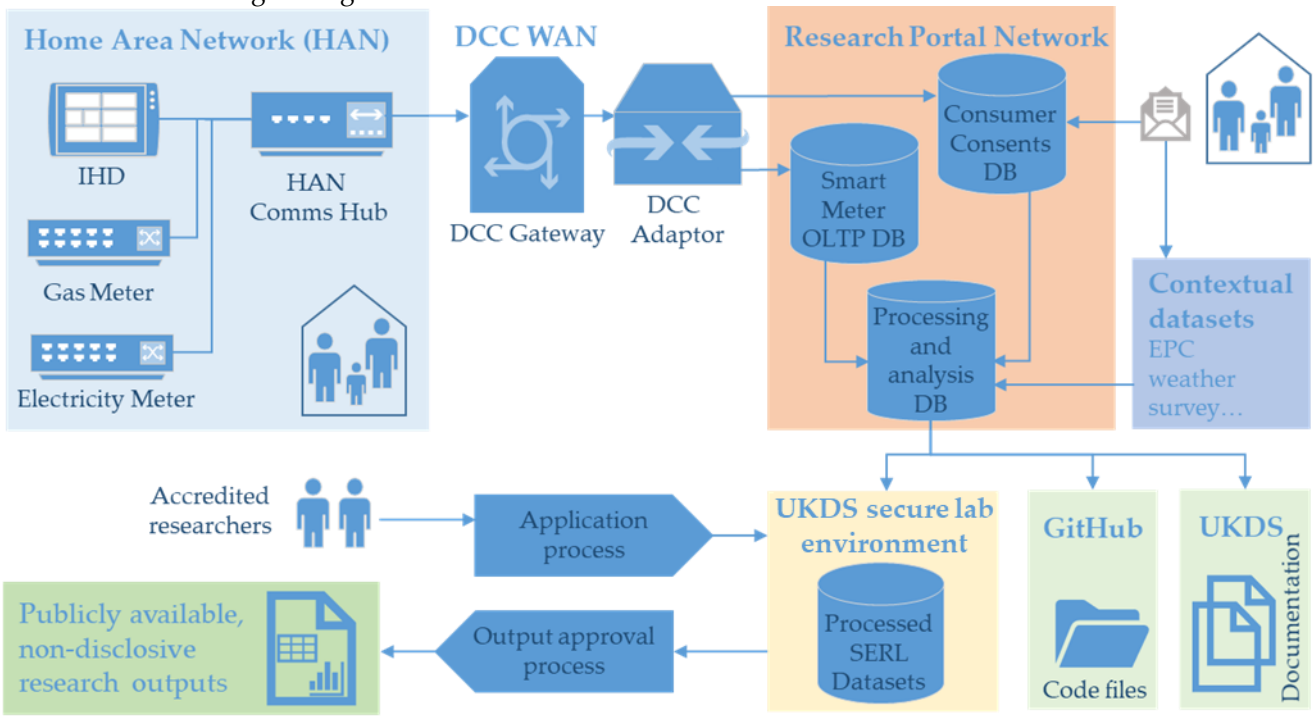

Figure 2: SERL system diagram summarising the data flows and access processes for SERL Observatory data.

2.2. Participant recruitment

[2] Reactive power and electricity export data are only stored for 3 months. The SERL project collects data for up to 12 months prior to consent, as asking participants whether they moved in less than 13 months previously is more complicated than asking if they moved in less than a year ago.

 doi:10.20944/preprints202109.0370.v1

The goal of participant recruitment was to recruit a representative sample of the GB population, acknowledging that not all households have a smart meter, and that bias will occur from both the unevenness of the smart meter rollout[3] and from response bias. We used a stratified random sample approach for SERL participant recruitment. Our variables for stratification were region and index of multiple deprivation (IMD) quintile (an area-based metric of affluence based on factors such as education levels and crime) as these variables could be determined pre-mailing. Participants were recruited over three recruitment waves: in August/September 2019, August/September 2020, and January/February 2021. The process for participant selection for each recruitment wave used a stratified random address sampling approach as follows:

1. Determine the number of households in each region and IMD quintile in the UK Address Base dataset to be as representative as possible of the domestic housing stock, after filtering out those with an organisation name, those listed as not 'in use', without an approved delivery point or no geographical (local authority) address. The percentage in each gave us our target percentage to recruit for SERL.
2. Query a large random sample of these addresses via the DCC gateway to find those which returned a DCC-accessible smart meter, from which we could (in theory) collect data if we had consumer consent. We achieved a positive match with around 2-3% of addresses queried.
3. Estimate what the likely response rate from contacted households would be to provide an upper bound to attempt to contact. For wave 1 we estimated a 5-20% response rate and the actual response rate was 9.5%. Subsequent response rate estimates were based on the response rates from earlier waves and were IMD quintile-specific.
4. Query additional random samples of UK Address Base addresses as required in order to achieve a stratified random sample with the number in each region-quintile determined by the expected response rate and number recruited to date.

The first recruitment wave piloted different recruitment options such as incentives which then informed the subsequent recruitment strategies. The details and results of the first recruitment wave are reported in Webborn *et al.* [17]. For waves 2 and 3 participants were sent the most successful 'Content Version' from wave 1 and up to 2 reminder letters were sent out. This means that in the first mailing participants were invited to participate online with a unique code. They would access a 'participant portal' to consent to smart meter data access and contextual data linking and follow a link to complete the SERL survey. For those who did not respond, a second mailing was sent 13 days later, again encouraging participants to sign up online. Non-responsive participants were sent a final reminder 13 days after this with both online signup details and a paper consent form and survey which they could post back in a pre-paid envelope. This is known as 'push-to-web' recruitment as the postal response is initially withheld to encourage online participation. In wave 1 some participants were offered two postal response mailings which did achieve a higher response rate, but the higher costs and lower data quality (since automatic response checking and question routing could not be implemented) meant that the push-to-web approach was selected instead. Wave 1 also found that offering a £5 conditional 'Love2Shop' voucher increased response by 1 percentage point. However, there were concerns that many of the participating retailers would be closed/unavailable due to Covid-19 restrictions, so no incentive was offered for wave 2. The response rate to wave 2 was lower than wave 1 (possibly impacted by Covid-19) and so the voucher was then offered to all wave 3 participants to boost response for the final recruitment drive in January 2021.

Figure 3 shows the number of 'active' SERL participants from 1st August 2018 (0) to 1st May 2021 (12,992) who responded to this recruitment strategy. Participants are 'active' from the time they sign up until they withdraw consent or move house, at which point data collection stops (but collected data remains).

[3] For example, flats are less likely to have smart meters due to technical issues with early versions of comms hubs, and in GB smart meters are not mandatory.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

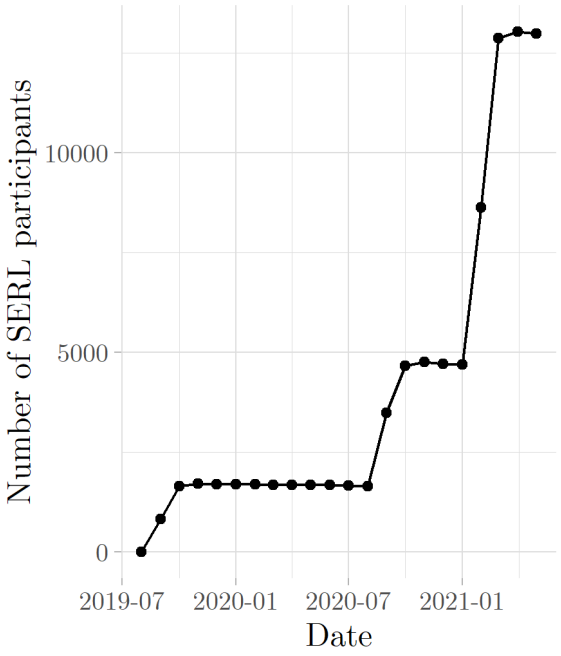

Figure 3: Number of active SERL participants over time.

2.3. Data Governance, Ethics and Consent

To access smart meter data via the DCC Gateway SERL is required to adhere to a strict set of criteria set out in UK legislation known as the Smart Energy Code (SEC) [10]. As part of this, SERL participant and data management processes have been reviewed and approved by the SEC Independent Privacy Auditor, which particularly focused on consent and consumer authentication (Sections I1.2 to I1.5 of the SEC). The UCL Research Ethics Committee approved the operation of the research data portal, including participant recruitment processes, the optional survey, data management and data provisioning (the principle of providing the SERL Observatory dataset to researchers via a secure lab environment). The SERL Observatory dataset is also registered with the Information Commissioner's Office (ICO) via UCL's data protection officer.

Research projects using SERL data must comply with additional data governance and ethics processes. University ethics approval may be required for projects using SERL data, and all projects must be approved by the independent SERL Data Governance Board (DGB) [18] as the final step in the project review process. SERL data can then be made available to accredited UK researchers in a secure lab environment, following the established '5 safes' protocol [19] in provisioning data to researchers:

1. Safe people: all researchers must obtain Office for National Statistics (ONS) Accredited Researcher status;
2. Safe projects: all projects must be approved by the independent SERL DGB;
3. Safe setting: data available via the UKDS secure lab environment;
4. Safe data: data is pseudo-anonymised appropriately for the secure lab environment;
5. Safe outputs: all outputs are SDC- (Statistical Disclosure Control) checked prior to release.

3. Data records

Data records are released together and classed as 'editions' approximately twice per year following data quality checking and basic processing. For example, 'edition02' was released in April 2021 containing data up to the end of October 2020. All filenames are suffixed with the edition number, e.g. 'serl_survey_data_edition02.csv'. Documentation is provided with each data release describing the variables available and any updates since the previous edition. The documentation for the third edition which contains the datasets described here is provided in supplementary files with this paper.

3.1. Smart meter data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

In GB there are SMETS1 (Smart Metering Equipment Technical Specifications 1) (first generation) and SMETS2 (second generation) smart meters. SMETS2 meters all communicate through the DCC gateway, and SMETS1 meters are being migrated onto the DCC network, allowing access for any supplier following supplier switching, and also allowing access for DCC 'other users' such as SERL. Both meter types record half-hourly energy readings; SMETS2 meters also store daily readings. Electricity smart meters record both active and reactive power and export readings for dwellings that export to the grid. Daily readings are stored for active electricity import (Wh) and gas import (m^3) only.

SERL smart meter data records are split by temporal granularity and by time period (electricity and gas appear in the same table). There are currently 12,823 electricity smart meters and 9730 gas smart meters with valid data in the SERL Observatory dataset. Of these, SERL is still able to collect smart meter data for almost all; for example, 12,528 electricity meters and 9344 gas meters have recorded valid data since the start of 2021. Table 2 summarises the availability of the smart meter data for each read type. Note that fewer households have daily reads than half-hourly as some of the SERL meters are SMETS1 (which do not store daily reads). The start dates are different for the different read types because reactive reads and active export reads are only stored on the meter for three months, and initial SERL data collection did not start with all read types at once. Mean and median availability are the mean (or median) of the percent of reads that are available and valid for each read type for each household, given their data collection date range (unique to each household based on when data collection started, when they moved in, when they signed up to SERL, etc.). The median availability is higher than the mean, implying that most households have good availability, with only a small number struggling with very little valid data. We discuss the data quality analysis of the datasets in Section 4.1.

*Table 2: summary of smart meter data availability by read type for the SERL Observatory sample. By 'available' we mean the read exists as expected and is valid according to our criteria (see below). *Households gives the number of households in the sample with at least one valid read. **Mean (or median) of the data availability for each household.*

Type	*Resolution*	*Details*	*Units*	*House-holds**	*Earliest first read date*	*Mean first read date*	*Mean availa-bility***	*Median availabil-ity***
Electricity	daily	active import	Wh	9513	2018-08-18	2020-02-17	88.9%	100.0%
Electricity	half-hourly	active import	Wh	12492	2019-01-13	2020-01-31	95.1%	99.9%
Electricity	half-hourly	reactive import	varh	12819	2019-09-25	2020-08-31	85.7%	92.5%
Electricity	half-hourly	active export	Wh	806	2019-11-07	2020-08-28	84.8%	86.6%
Electricity	half-hourly	reactive export	varh	806	2019-11-07	2020-08-28	84.9%	86.7%
Gas	daily	import	m^3	8594	2018-08-19	2020-02-04	88.9%	99.7%
Gas	half-hourly	import	m^3	9729	2019-01-13	2020-02-07	90.4%	98.8%

Daily electricity data was originally released in one file, and is now split by year (currently 2018, 2019, 2020 and 2021). Half-hourly data was originally released in one file but as the file size increased it became necessary to split the data by month (and year). Original data has not been modified; additional columns have been added to flag potential errors and convert between units. In the daily datasets, if no readings exist for a given participant on a given day then no row for that participant on that day will exist in the dataset, unless a full day's half-hourly readings[4] exist for one of the read types, in which case the row is included in order to show the sum of the half-hourly reads for that day. In the half-hourly datasets, if no data is available for a given participant on a given half hour, then that row will be missing from the dataset. Error flags are discussed in Section 4.1.

Figure 4 shows how the number of meters recording at least 20 days' worth of valid reads each month changes as the number of participants increases. Note that for technical reasons data collection started later for half-hourly reads than daily, and latest for half-hourly electricity active export and for half-hourly reactive reads (not shown). Half-hourly electricity active import reads have greater availability/validity than their daily equivalent, and the sum of half-hourly reads each day is provided in the daily smart meter data table to allow for use where daily reads are unavailable. Note that SMETS1 meters only record half-hourly data.

[4] 'A full day' is 48 half hours except for the start and end of British Summer Time (BST) when there are 46 and 50 half hours, respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

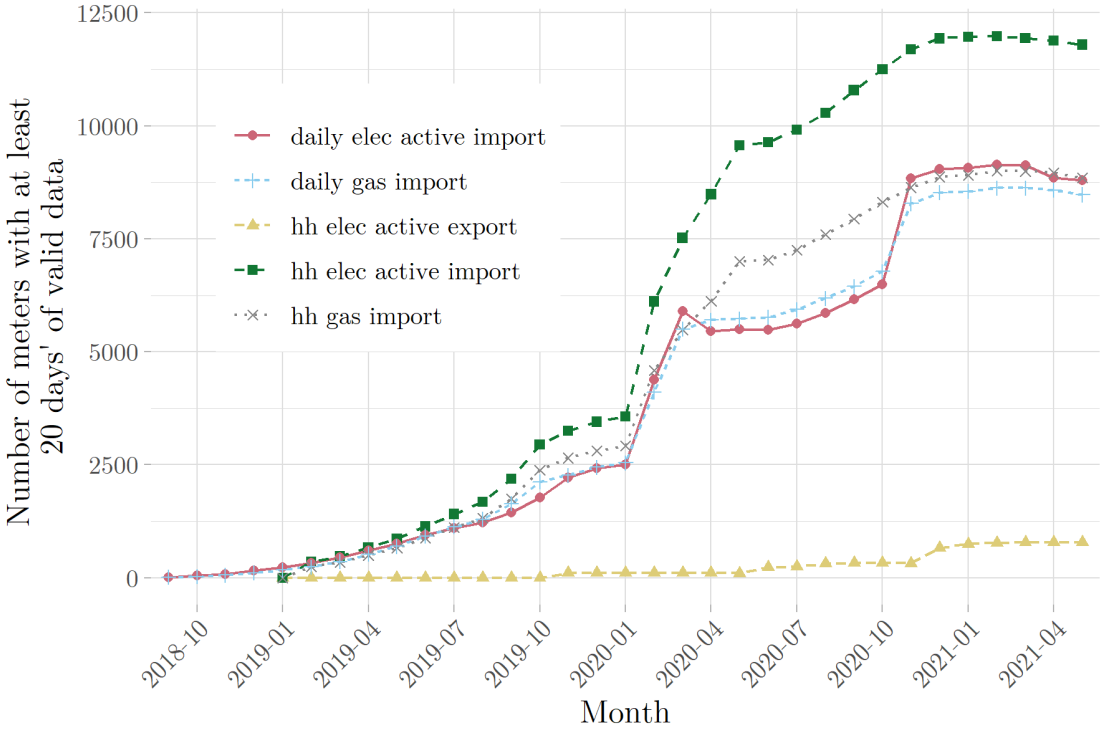

Figure 4: Number of meters with at least 20 daily reads (or 20 days' worth of half-hourly reads) available, recorded at the correct time, and flagged as 'valid' each month.

The smart meter data-related documentation in the supplementary files are:

- *serl_smart_meter_documentation_edition03.pdf* – describes the smart meter datasets in detail including all variables provided.
- *serl_smart_meter_data_quality_report_edition03.pdf* – describes the data quality analysis done and data availability.

3.2. *Survey data*

The SERL survey asks 40 questions of occupants about their dwelling, household members, and attitudes and behaviours related to energy use. Where possible, questions were harmonized with those in existing surveys (such as the English Housing Survey or Understanding Society). They were chosen collaboratively between the eight SERL consortium partners; designed to capture the representativeness of the SERL sample compared to the GB population and for analysis of factors likely to affect energy consumption.

Participants received the same survey except for question A5 which asked about temperature set points. In wave 1 only one box for temperature in degrees Celsius was provided, but since it was clear that many people responded in degrees Fahrenheit, two boxes were provided in the future survey versions. Both versions of the (postal) survey are provided in the supplementary files. Online and postal survey versions were identical apart from automatic question skipping and answer checking for online respondents.

The survey data contains 12,977 rows (one per participant (responding household) who started the survey). The survey data has been cleaned to allow for maximum compatibility between the responses of online and postal respondents. This means that if a question would have been skipped for a participant automatically had they completed online, we have removed the response for the postal participant as if automatic skipping had occurred. Additional cleaning has been done to flag likely erroneous responses, impute variables where error correction is possible (*e.g.* clear that tally notation has been used so a value should be 2 rather than 11), and summarise data (*e.g.* to indicate that no one is working or everyone is aged over 65 years). In Sections 5.2 and 5.3 we present the results from the harmonized survey questions and discuss the representativeness of the SERL sample.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

The supplementary files contain a table showing the frequencies of each response option to the survey questions (merged or rounded where necessary for statistical disclosure control[5]). The survey data-related documentation in the supplementary files are:

- *serl_survey_documentation_edition03.pdf* – describes the variables in the SERL survey dataset
- *serl_pilot_recruitment_survey_copy.pdf* – the postal version of the survey sent to wave 1 participants
- *serl_main_recruitment_survey_copy.pdf* – the postal version of the survey sent to wave 2 and 3 participants
- *serl_survey_response_frequencies_edition03.csv* – table of frequencies of responses to the SERL survey questions, merged or rounded for SDC checking where values are fewer than 10.

3.3. *Energy Performance Certificate (EPC) data*

Of the 13,321 registered SERL participants (including those who have withdrawn consent and for whom we are no longer collecting data), 52% (6921) have an Energy Performance Certificate (EPC). The dataset consists of one row per participant and 80 columns including the PUPRN identifier. The SERL Observatory edition 3 (described here) does not include any Scottish EPC data; this will be provided in future editions. EPC data was collected for the SERL Observatory using house number and postcode with the Domestic Energy Performance Certificates API[6]. Note that where more than one EPC is registered for an address we only provide the latest version. Other than replacing address data with our participant identifier (the PUPRN) we have not modified or performed data quality analysis on this dataset. The EPC documentation in the supplementary files is named *serl_epc_documentation_edition03.pdf.* In Section 5.4 we present a short analysis of the representativeness of our sample in terms of current EPC ratings, and consider the potential bias among households with an EPC.

3.4. *Climate data*

The SERL Observatory climate dataset comes from the Copernicus/ECMWF ERA5 hourly reanalysis dataset[7]. This is reanalysis (modelled) data based on readings from weather stations across GB at a horizontal resolution of 0.25 x 0.25 degrees latitude and longitude (approximately 28 square km). The *grid_cell* variable is provided in the climate data and participant summary table for linking at the household level. It takes the format *xx_yy* where *xx* indicates latitude and *yy* indicates longitude, where *00_00* refers to 61 degrees latitude and -8 degrees longitude (the most Northwesterly point in our original data grid) and an increase of 1 in *xx* represents a 0.25 degree *decrease* in latitude; an increase of 1 in *yy* represents a 0.25 *increase* in longitude. Data are provided for all grid cells containing at least one household. For more information about this dataset see the ERA5 documentation[8].

The climate data is split into monthly files (currently 34 from August 2018 to May 2021), with approximately 340,000 rows per file. Data are provided for the 466 grid cells containing participants. There are 24 climate variables in addition to the grid cell and time/date variables, as described in *serl_climate_documentation_edition03.pdf* in the supplementary files. The variables provided relate to temperature, precipitation, pressure, solar radiation, cloud cover, wind and snow.

3.5. *BST dates*

British Summer Time (BST) dates is a small table stating the start and end date of BST (the dates the clocks change) from 2018 to 2024 (more years may be added going forward). There are 14 rows and 3 columns: *Read_date_effective_local*, *type* (‘start’ or ‘end ‘of BST) and *n_hh* (either 46 or 50: the number of half-hours on the days the clocks change). Daily data is recorded in local time and therefore corresponds to a different number of hours per day on the dates that BST starts and ends. Half-hourly data was originally provided in Universal Time Coordinated / Universal Coordinated Time (UTC) and local date times are included in the half-hourly datasets for reference.

4. Data Quality Analysis

[5] For statistical disclosure control, as recommended by the UK Data Archive, we do not report figures relating to fewer than 10 households. To avoid this, categories may be merged or counts rounded to the nearest 10.

[6] https://epc.opendatacommunities.org/docs/api/domestic

[7] https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5

[8] https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

We have performed data quality checking and analysis for the smart meter and survey data as a guide for data users. The climate variables are derived data from re-analysis of multiple original datasets and therefore do not suffer from missing data, although other issues may exist that we have not investigated. EPC data quality has been investigated previously and found to suffer from inaccuracies [20], [21]. Researchers are encouraged to treat EPC data with caution and consider potential data quality issues, in addition to the potential bias that requiring an analytic sample to have 100% EPC data will entail (Section 5.4). All data processing for the edition 03 data and the analysis presented here was performed using R version 4.0.1 (2020-06-06) [22]. Code used to perform the pre-processing of SERL datasets is available on GitHub[9].

4.1. Smart meter data

As part of the smart meter data preprocessing, each read is given an error flag to indicate our belief about its validity. Each row also contains a 'valid_read_time' flag (TRUE if on the hour/half-hour for half-hourly data or at midnight for daily data). The error flags are described in Table 3. Note that if all reads for a particular time are missing for a meter then the data will be a missing row rather than an empty row in the dataset.

Table 3: Error flags and their meanings.

Code	Name	Meaning/Details
3	Ignore	Invalid read time, no read. The row exists for a different read type that was taken at the wrong time so we do not require a read for this time stamp.
2	No meter	The meter does not exist in the DCC system; the read is not actually 'missing' because we do not expect it. For example, when there is no gas meter all gas read flags will be '2' for rows that exist because there are electricity reads.
1	Valid	The read exists and does not meet any of the other error flag criteria, and valid_read_time = TRUE thus presumed valid.
0	Missing	The read should exist as far as we are aware but it is missing.
-1	Max read	The largest number storable on the meter (equal to 16777215, all 1s in 32-bit binary (the 64-bit equivalent is converted to 32-bit to save memory). Likely due to a technical fault.
-2	Very high but not max	Table 5 in the *serl_smart_meter_documentation_edition03.pdf* (supplementary file) shows the threshold for flagging a read as larger than we (cautiously) deem plausible, excluding those flagged as 'max reads'.
-3	Negative	Value less than 0 (no occurrences).
-4	Elec in kWh	Electricity reads are required to be recorded in Wh but for some meters we believe that the readings have incorrectly been stored in kWh. Some properties do not suffer from this problem for the entire period of collection; the issue may start or stop when a meter is replaced. Due to the difficulty in automatically assessing this issue, any meter with at least 5 rows of electricity data where the daily reads are approximately 1/1000th of the sum of the half-hourly reads for that day are flagged with this error for their entire recording period. A new column with unit correction has been included, but since the reads are rounded to the nearest kWh instead of the nearest Wh researchers may wish to exclude such data.
-5	Valid read, invalid read time	Originally flagged as valid (1) but valid_read_time = FALSE therefore we cannot say over what time period the energy has been recorded. E.g. at 15:01 we may wish to keep the read, and assume it is for 14:30 – 15:01. But at 15:15 perhaps it is 14:30 – 15:15 but it may be 15:00-15:30, depending on the previous read. We suggest researchers filter out reads at invalid times.

Zero reads may indicate erroneous data, as some meters have been found to record zeros during British Summer Time (BST), but these are left to researchers to consider. It is also possible that some half-hourly electricity data is recorded in tens of Wh, as a handful appear to have half-hourly sums approximately 1/100th of the daily read. Table 4 shows the number and percentage of each error flag by read type for the SERL Observatory edition 3 dataset. Missing data is the most substantial issue, affecting 11.5% - 24.2% of reads. The worst affected are the daily reads and we provide the sums of half-hourly reads alongside the daily data for imputation and comparison. There are very few reads flagged with error codes other than valid or missing; 'Max read' is the next most prevalent, affecting around 0.7% of gas reads

[9] https://github.com/smartEnergyResearchLab/observatoryData

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

(half-hourly and daily). In terms of individual meters, 82% of meters have at least 90% of their half-hourly electricity active import reads flagged valid; 75% of accessible gas meters have at least 90% of their half-hourly gas reads valid. These numbers are lower for daily reads: 56% of electricity meters have at least 90% of reads valid; 71% for daily gas. Fewer electricity meters have high-quality export data; only 36% with at least 90% valid, but this rises to 84% if we only require at least 75% valid.

Table 4: Smart meter error flag prevalence by read type.

Flag	Half-hourly reads					Daily reads	
	Electricity active import	Electricity active export	Electricity reactive import	Electricity reactive export	Gas import	Electricity active import	Gas import
Valid	280,433,417 88.5%	10,236,704 87.0%	158,556,764 84.1%	10,246,482 87.0%	208,866,909 86.6%	4,009,715 66.0%	3,803,714 82.7%
Missing	36,290,271 11.5%	1,526,268 13.0%	29,901,735 15.9%	1,526,268 13.0%	30,480,701 12.6%	1,467,930 24.2%	749,785 16.3%
Max read	542 0.0%	0 0.0%	2 0.0%	0 0.0%	1,736,169 0.7%	0 0.0%	37,880 0.8%
Very high	182 0.0%	9,778 0.1%	1 0.0%	0 0.0%	7,622 0.0%	1,196 0.0%	4,394 0.1%
Negative	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%
Elec in kWh	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	594,690 9.8%	0 0.0%
Valid but invalid time	228 0.0%	18 0.0%	186 0.0%	18 0.0%	69,431 0.0%	6 0.0%	1,260 0.0%

Figure 5 and Figure 6 summarise the availability of valid data by read type; each point represents the percent of meters with at least x% of their reads available and valid (flag 1). We see that nearly 25% of electricity meters have less than 10% valid daily reads over their possible data collection date range; significantly lower than for their half-hourly equivalent. However, the meters with at least 10% valid typically have a very high percentage valid; over 50% of meters have at least 99% of their daily and half-hourly electricity reads valid. For electricity export data, almost all export meters have at least 50% validity, but this starts to decrease, with only around 20% having at least 95% data valid. For gas meters (Figure 6), daily and half-hourly data quality issues seem to affect meters similarly, and around 75% of meters record at least 90% valid reads.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

Figure 5: Percent of meters with x% of data valid over their potential valid read date range, by type of electricity read. 'hh' stands for half-hourly; 'elec' for electricity.

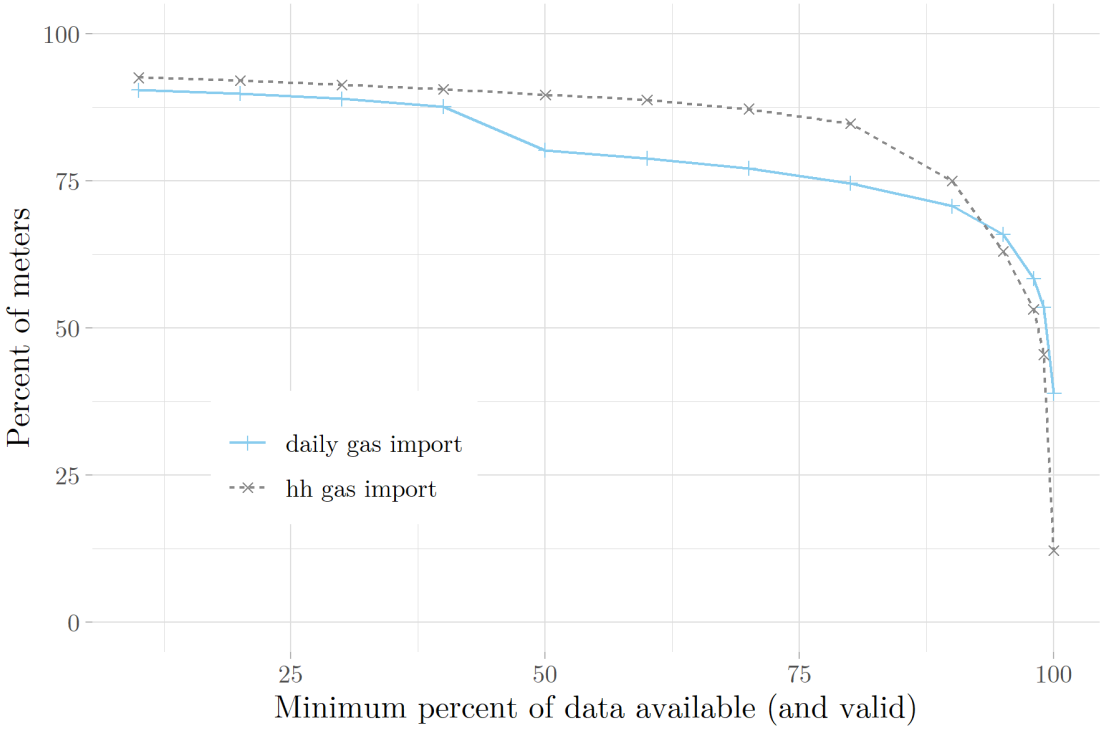

Figure 6: Percent of meters with x% of data valid over their potential valid read date range, by type of gas read (daily and half-hourly ('hh')).

SMETS1 meters do not store daily readings and SMETS2 daily reads can suffer from missing data or other data quality issues highlighted above. For these reasons we provide researchers with additional columns in the daily datasets that show the sum of the half-hourly reads each day. In order to be included, there must be the correct number of valid half-hourly reads taken at the correct time (error flag 1). By 'correct number' we mean 48, unless the clocks change, in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

which case 46 or 50. We also provide an 'Elec_sum_match' and 'Gas_sum_match' to code a comparison between the half-hourly sums and the daily read. A description of these codes can be found in Table 6 in the *serl_smart_meter_documentation_edition03.pdf* supplementary file.

Information about the data quality for each read type for each participant is provided in the read-type summary dataset. It contains information about the 'theoretical' date range when data could have been available (if not missing or erroneous), the actual date range when valid data was recorded, the number and percent of reads with each error code, and the minimum and maximum recorded valid reads. Note that in this table missing data is counted as the number of theoretical reads minus the number of reads with error flags 1, -1, -2, -3, -4 or -5 (as some missing data will be in a missing row rather than flagged as missing). For more details about this dataset see the *serl_smart_meter_documentation_edition03.pdf* supplementary file.

4.2. Survey data

The SERL survey questions underwent in-person cognitive testing with ten smart meter owners to check for clarity and, in some cases, adapt their wording [17]. Copies of the wave 1 and wave 2/3 versions of the survey are available as supplementary files. 12,977 participants started the SERL survey, and each question had a response rate of at least 95.0%. The highest response rate, 98.9%, was for the first question on the survey: "Before we contacted you, did you know you had a smart meter?". The question with the lowest response rate (95.5%) asked about the main heat source for their hot water taps and shower, with tick-boxes for the shower and taps. The following questions had detectable potential data quality issues:

- A5 - What temperature do you set your controller to in the winter months for the late afternoons or evenings? A box with °C was provided in the wave 1 survey, but it was clear that some respondents were reporting temperatures in °F. Reported temperatures above 35 were assumed to be temperatures in °F and were converted to °C. In subsequent surveys two boxes (°C OR °F) were provided. Some issues remained despite this, including: °F continuing to be reported in the °C box, a range of temperatures being provided, temperatures reported in both °C and °F but not matching within 1°C, and excessively high or low temperatures (defined as less than 10°C, or more than 35°C after assuming answers more than 35 were given in °F); all these issues have been flagged as errors in the dataset. After converting responses above 35 in the °C box (assuming they were in °F) (1.6%) 0.4% of responses remain flagged as errors.
- A16 - Which of the following, if any, is your household considering replacing or adding to your heating or energy supply in the next 12 months? This question had several possible options to tick and a free text "other" option. The free text response contained comments about various energy related issues, including describing works previously completed, comments on why various options are not possible and comments regarding installed solar panels. This may be useful data and has been retained in the dataset, but it is not necessarily the case that a participant answering 'other' is considering a change to their heating or energy supply in the next 12 months.
- B3 - How many other households do you share with at the moment? This question is applicable only to those who responded in the previous question that their accommodation is not self-contained. 0.6% of respondents who answered that their accommodation is not self-contained (*i.e. shared with other households)* went on to contradictorily indicate that they lived with zero other households.
- B5 - How many rooms are available for use only by this household? And B6 - How many of these rooms are bedrooms? Some respondents reported more bedrooms than rooms; these have been flagged as errors (0.4%). Some reported zero bedrooms, these have been edited to one bedroom; the same imputation as in the UK 2011 Census [23].
- C1 - How many people currently live in your household, including you? And C2 - Including you, how many males and females are there in each of the following age groups in your household? In some cases, the total number of householders reported in these two questions did not match; it appears that some respondents reported the ages of each householder, rather than the number of householders in each age and gender bracket. 1.7% of all respondents have C2 flagged as an error.
- C3 - Thinking about the working situation of each member of your household aged 16 and over, including you, how many would you say fall into each category below? In some cases, the numbers reported were larger than the number of occupants reported in the previous two questions. Sometimes it appears that people were reporting the number of hours worked in each category. 5.3% of all respondents have C3 flagged as an error.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

- C4 - Including you, how many people in your household hold a degree (e.g. BA, BSc) or higher qualification (e.g. MA, PhD, PGCE)? In some cases more people with degrees were reported than people living in the house (C1), and in some cases non-numeric responses were given. 0.1% of all respondents have C4 flagged as an error.

5. Sample bias and representativeness

SERL's original target was to recruit 10,000 households in GB to allow for national-level estimates for population sub-groups such as building types with a 2-3% margin of error [16] (page 25). The aim was for the sample to be maximally representative of GB, considering the limitations of a) bias in the smart meter rollout and b) response bias. Recruiting in three waves allowed us to attempt to redress the balance of participants by region and IMD quintile, by over-recruiting in areas/quintiles that were currently under-represented. In this section we assess the representativeness of the full SERL sample, those who responded to the survey (95%), and those with an EPC.

5.1. All SERL Observatory participants

We can use estimates for the 'true' proportions of households from Ordnance Survey's Address Base dataset [24] to assess the representativeness of the SERL Observatory sample by region and IMD quintile. Figure 7 compares the locations of SERL Observatory participants with the estimated national breakdown. Most of the regions are very close to the national proportion; Wales and Yorkshire being the exceptions with a -2.9 and +3.5 percentage point difference, respectively. Table B1 in Appendix B gives the number of participants in each region.

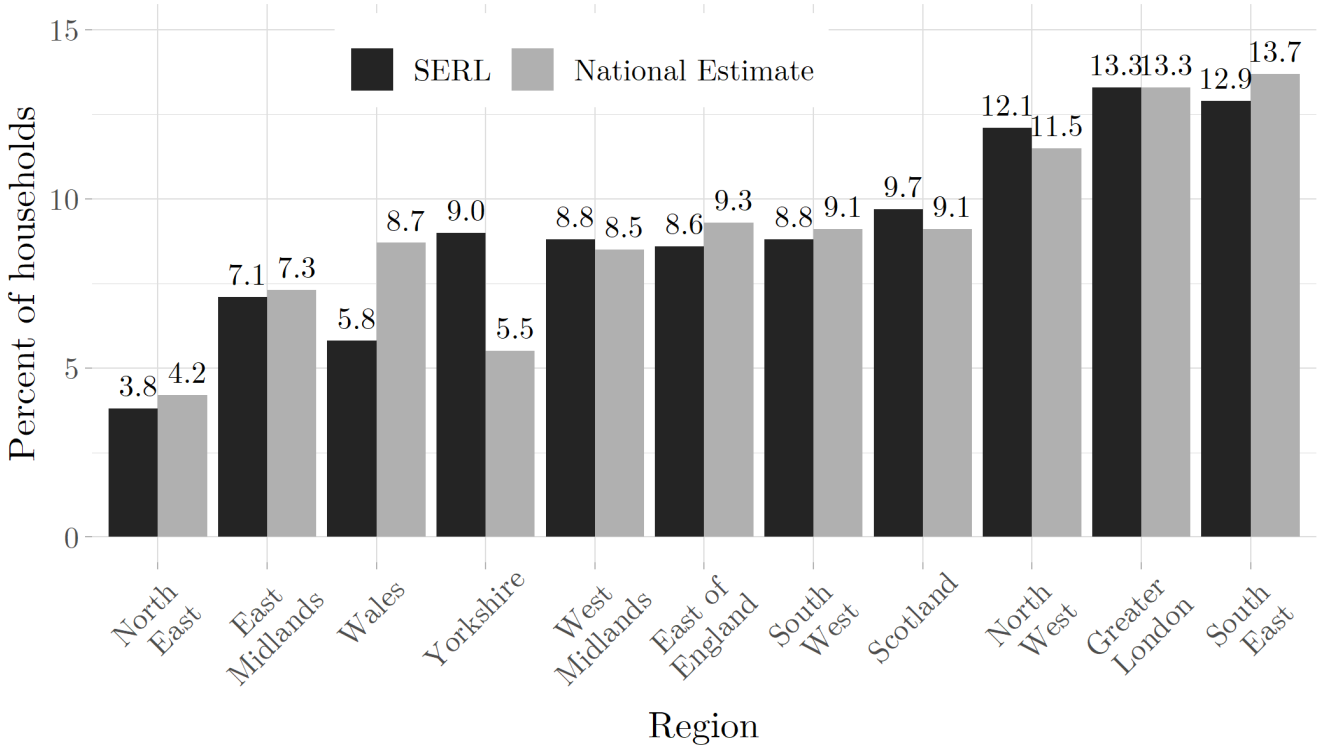

Figure 7: Percentage of the SERL Observatory in each GB region compared with our national estimate from Address Base.

Figure 8 shows the equivalent results to Figure 7 but split by IMD quintile (quintile 5 has the greatest affluence). Overall, the SERL Observatory slightly over-represents the two quintiles with the greatest deprivation (1 and 2) but only by a small amount. Initial recruitment strongly under-represented the lower quintiles and so efforts were made to redress this balance in later recruitment waves. Note that Index of Multiple Deprivation is an area-based metric rather than household-specific, and while it means that a household is more or less likely to experience deprivation, the socio-economic circumstances of individual households within the area could still play a role in response bias of the SERL participants. Region and IMD quintile were used for sample selection as these variables were available for all addresses pre-contact.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

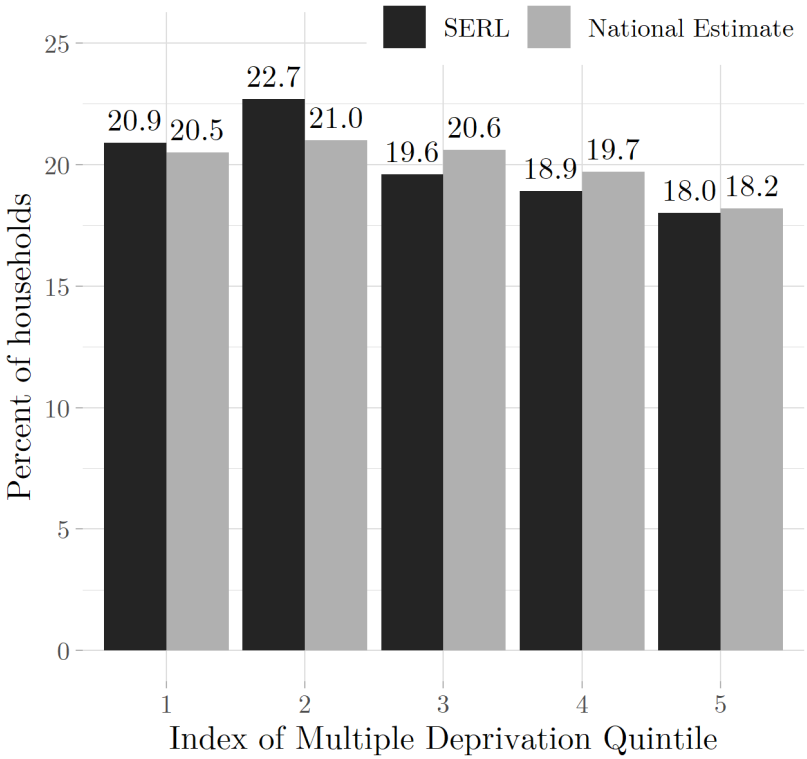

Figure 8: Percentage of the SERL Observatory in each Index of Multiple Deprivation quintile compared with our national estimate from Address Base.

5.2. SERL Observatory survey respondents: households

Many of the SERL survey questions are designed to assess household demographics and attitudes towards energy-saving practices. In this section we compare responses to the SERL survey with those from harmonized questions in the English Housing Survey (EHS) 2019/20 [25], Understanding Society Wave 10 [26], and the UK 2011 Census [23]. Note that the figures presented from these external datasets have not been weighted for response bias. The EHS only surveys households in England, and so comparisons can only be made with SERL survey respondents in England. Understanding Society covers the UK including Northern Ireland, and it was not possible to restrict this data to GB-only, so we have to assume that removing households in Northern Ireland would not have a significant effect on the Understanding Society statistics. The UK Census data can be filtered by region, so unless stated otherwise, '2011 Census' refers to households in GB. The Census has far more respondents than either of the other two surveys, and can be filtered for the regions required, but was carried out approximately 9 years before the SERL survey. For these reasons, none of the datasets should be considered to be 'ground truth', rather, indicative of potential bias in the SERL sample.

Figure 9 compares the number of occupants in SERL Observatory households in England with those in EHS 2019/20 households. The main difference is an overrepresentation of two-person households. Responses to question C2 about the number of males and females in each age category indicates that the SERL Observatory may also overrepresent households with all occupants aged 65 and over (30.9% in SERL compared to 23.7% in the 2011 Census).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

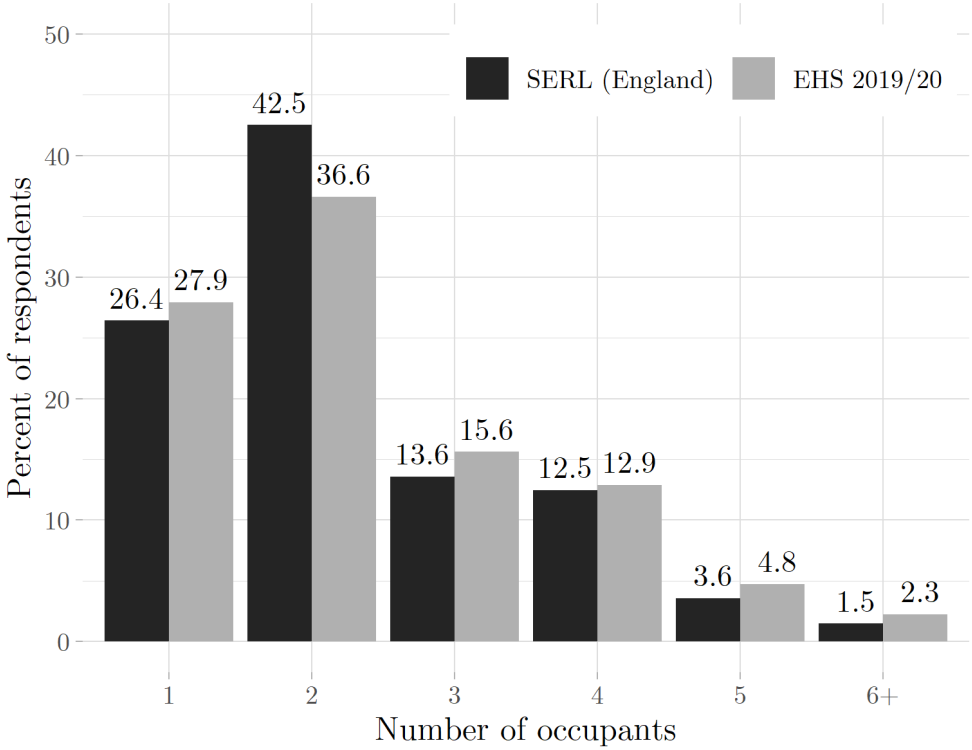

Figure 9: The number of occupants in SERL Observatory households in England (question C1) compared to EHS 2019/20 households.

Question B4 asks about tenure with five response options: 1) own it outright/buying it with a mortgage/loan, 2) part own and part rent (shared ownership), 3) rent privately, 4) rent from council (local authority) or housing association, and 5) live here rent free. The rental options are all stated as 'with or without housing benefit'. In order to compare with other surveys some of the categories need to be combined. The EHS has high-level response categories 'all owner occupiers', 'all social renters' and 'all private renters'. We compare the SERL responses (in England) by excluding "no answer" (1.9%) and "live here rent free" (0.6%) and merging "own it outright/buying with a mortgage/loan" (77.6%) with "part own and part rent (shared ownership)" (0.8%). We see that the SERL Observatory significantly overrepresents owner-occupiers (15.7 percentage points higher than EHS) and underrepresents both private renters (8.8 percentage points lower) and social renters (6.9 percentage points lower) (Figure 10). Understanding Society Wave 10 also asks this question but only has one 'rented' option (for full details see Figure B1 in Appendix B). Comparing with this dataset we find a substantially smaller disparity between the SERL percentage of owner-occupiers (77.7%) and the percentage in Understanding Society (70.6%), while the SERL renters make up 19.2% compared with their 27.3%. This bias is likely to be due to response bias from people renting, fewer smart meter installations in rental properties where renters may not be invested in changing something in a property they plan to live in temporarily, and flats (which suffer from lower smart meter install rates) are more likely to be rented than dwellings such as detached houses.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

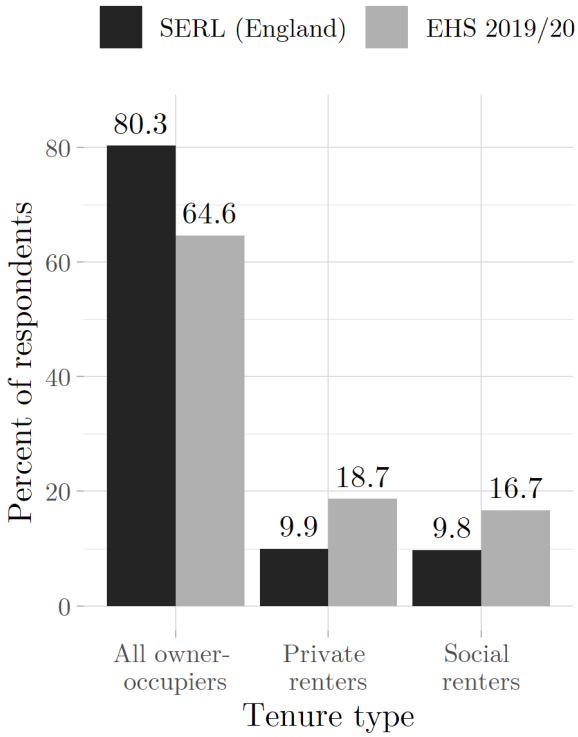

Figure 10: Comparing SERL Observatory tenure types in England (counts rounded down to the nearest 20 for statistical disclosure control) with the EHS 2019/20 survey (question B4).

The SERL survey also asks respondents how well they would say they personally are managing financially. Figure 11 compares the results to the responses from the same question in Understanding Society Wave 10. Note that the SERL 'null response' includes those who didn't respond, those who responded "prefer not to say" and those who responded "don't know", while the Understanding Society 'null response' includes those with response options "missing", "proxy", "refusal", or "don't know". SERL respondents were less likely to answer the question, either because they had stopped answering the survey already, or perhaps because they found the question too intrusive/personal. SERL respondents were much more likely to report 'living comfortably' compared to 'just about getting by', 'finding it quite difficult', or 'finding it very difficult'.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

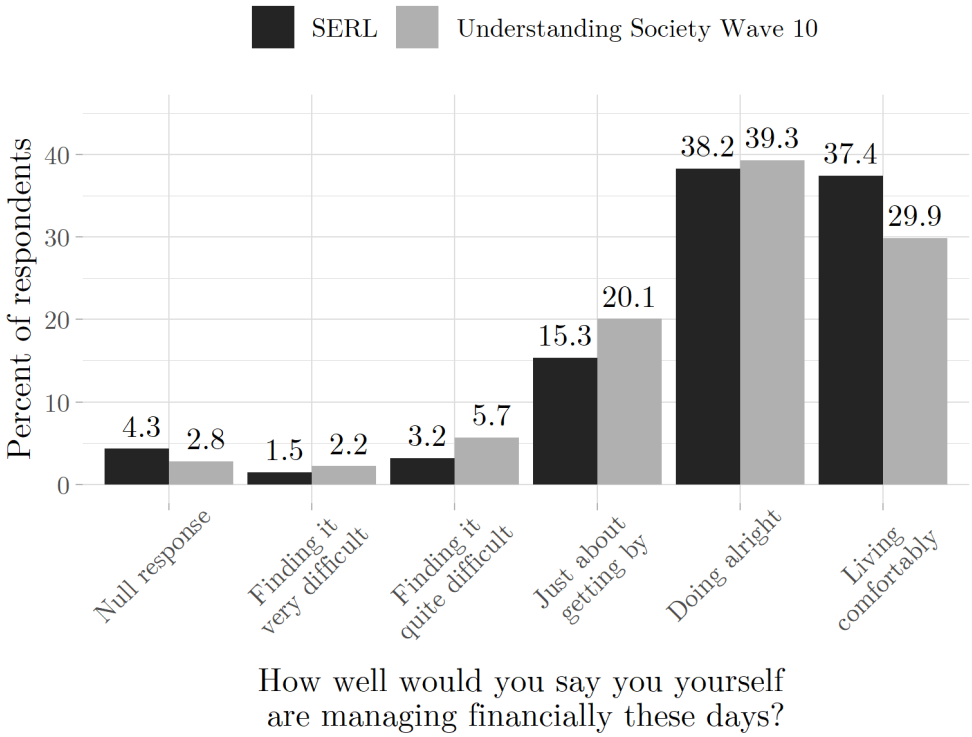

Figure 11: SERL Observatory responses to question D4 about financially wellbeing, compared to Understanding Society Wave 10.

Question A13 in the SERL survey concerns energy-saving behaviours, asking "how often do you switch off lights in rooms that aren't being used?" and "how often do you put more clothes on when you feel cold rather than putting the heating on or turning it up?". The responses can be compared with those to the same question in the Understanding Society Wave 10 survey. Note that "Null response" in the SERL survey includes those who answered "not applicable, cannot do this" and those who did not answer. Understanding Society had more response options, and "null response" includes those who did not respond or whose response was "proxy", "refusal", "don't know" or "not applicable, can't do this". Figure 12 shows that SERL respondents were much more likely to switch off lights in unused rooms 'very often' although slightly less likely to switch off lights 'always'. SERL respondents were a little less likely not to respond or respond with 'never', 'not very often', or 'quite often'. Interestingly we see a similar pattern of response differences when comparing the question about putting more clothes on instead of using the heating more (Figure 13); fewer respondents with the most energy-conscious response but more with the second-most, and significantly fewer choosing 'never'. When combining the two most energy-saving behaviour options, the results from both questions show that the SERL participants are slightly more energy conscious than those in Understanding Society, which is possibly to be expected given the framing of the SERL recruitment information [17].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

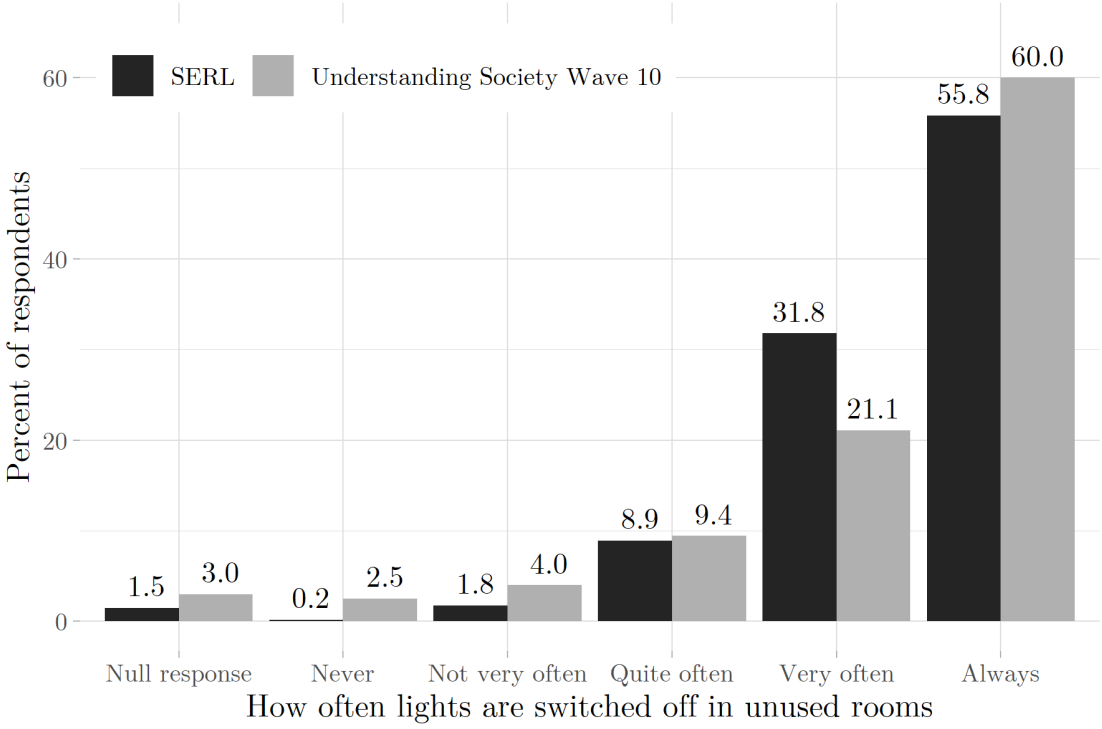

Figure 12: Responses given to question A13a about saving energy by switching off lights (SERL survey respondents compared to Understanding Society Wave 10).

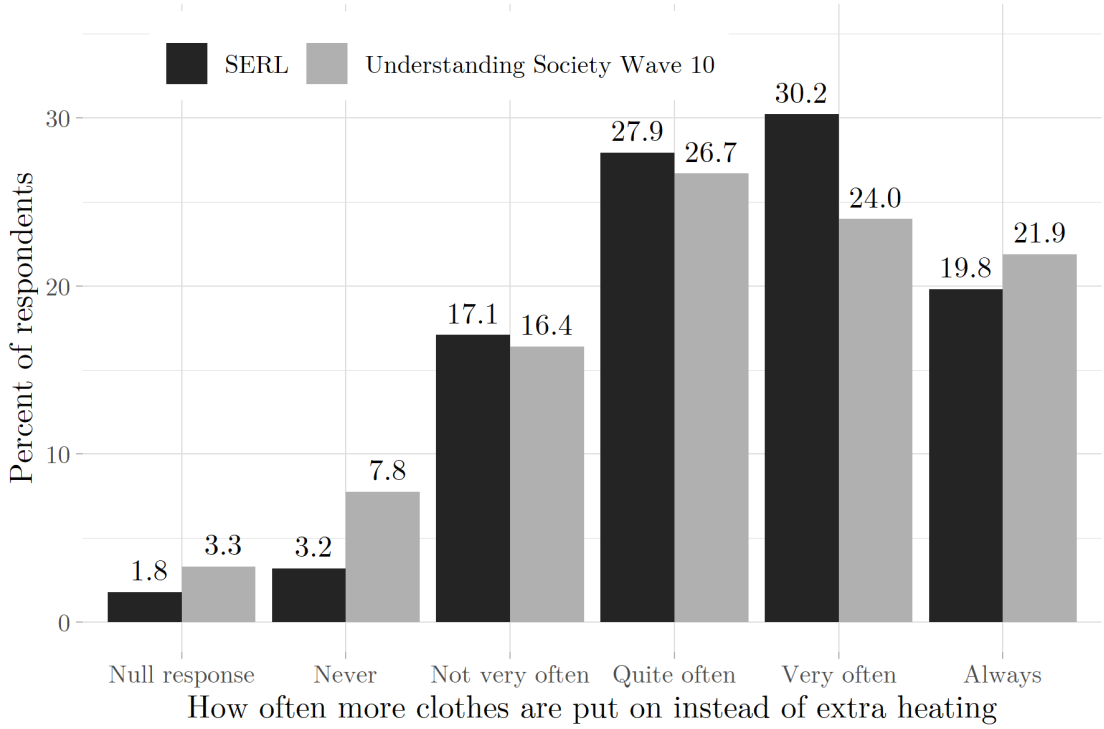

Figure 13: Responses given to question A13b about saving energy by choosing to wear more clothes over using more heating (SERL survey respondents compared to Understanding Society Wave 10).

5.3. SERL Observatory survey respondents: dwellings

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

Dwelling characteristics play a key role in how households consume energy, and many of the SERL survey questions were designed to capture relevant dwelling information, in some cases harmonized with other surveys for representative comparisons.

We can compare the responses to question B1 about dwelling type with the results from the 2011 Census (Figure 14). The SERL Observatory slightly overrepresents detached dwellings (4.5 percentage points higher) and slightly underrepresents flats (7.4 percentage points lower for all flat categories combined). As mentioned above, smart meter installs are known to have been less prevalent in flats. Note that 2.1% of survey respondents did not answer this question, and the percentages shown exclude non-response.

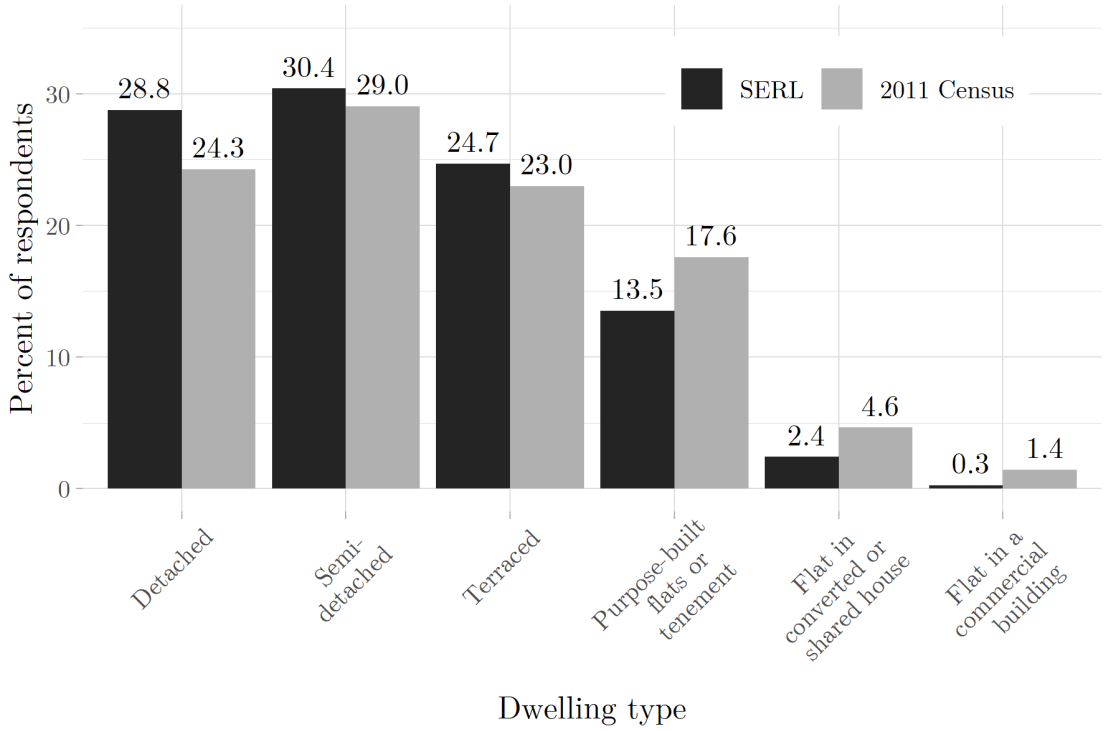

Figure 14: Dwelling types for SERL survey respondents (question B1) compared to the 2011 Census.

Dwelling size can also play a role in energy consumption, as larger dwellings require more energy to heat. Question B5 asked SERL respondents to count the rooms in their dwelling available for use only by the household, excluding bathrooms, toilets, halls, landings and storage-only rooms such as cupboards. Compared to EHS 2019/20 respondents, SERL survey respondents in England are substantially more likely to live in very large properties (more than 8 rooms) or very small (1 or 2 rooms), and substantially less likely to live in 4 or 5 room properties (

Figure 15). The reason for this difference is unclear but could be associated with the different data collection methods between SERL and EHS; SERL is a self-reported survey whereas EHS collects this data either via a surveyor or a telephone interview. When comparing the number of bedrooms reported in the SERL survey with the 2011 Census data[10], SERL also overrepresents dwellings with 4 or more bedrooms, but underrepresents those with 1 or 2 bedrooms (Figure 16).

[10] Data on the number of bedrooms in Scotland is not reported in the 2011 Census so we compare with SERL survey data for England and Wales.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

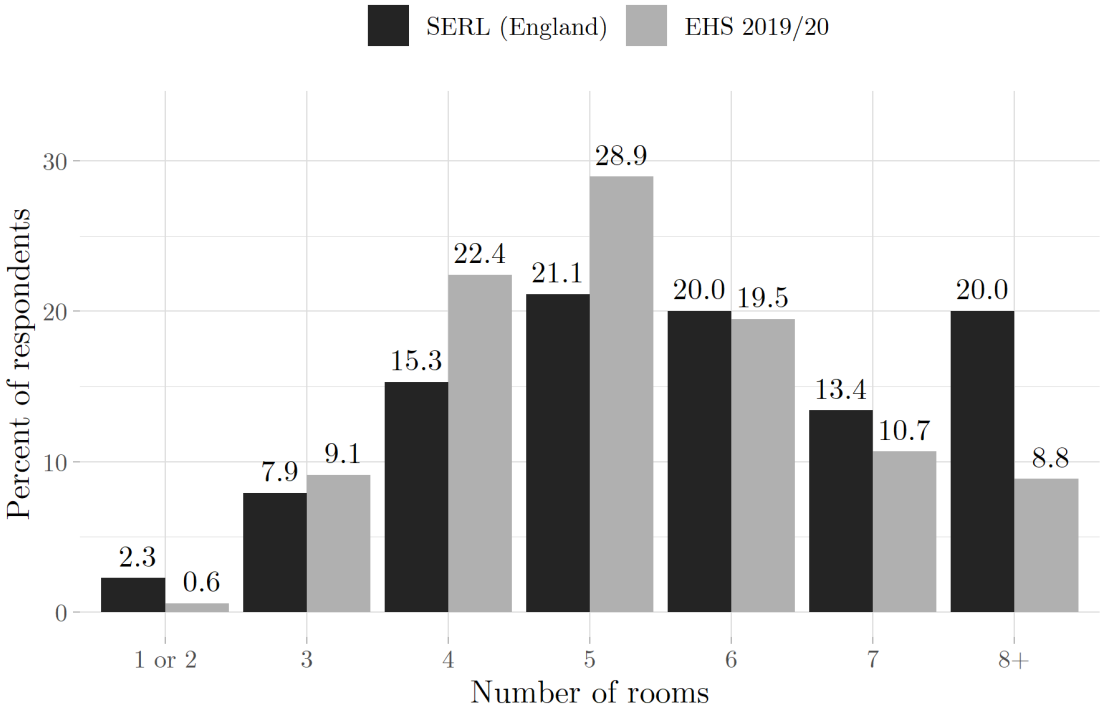

Figure 15: The number of rooms in SERL Observatory dwellings in England (counts rounded down to the nearest 20 for statistical disclosure control) compared to those reported in the EHS 2019/20 survey.

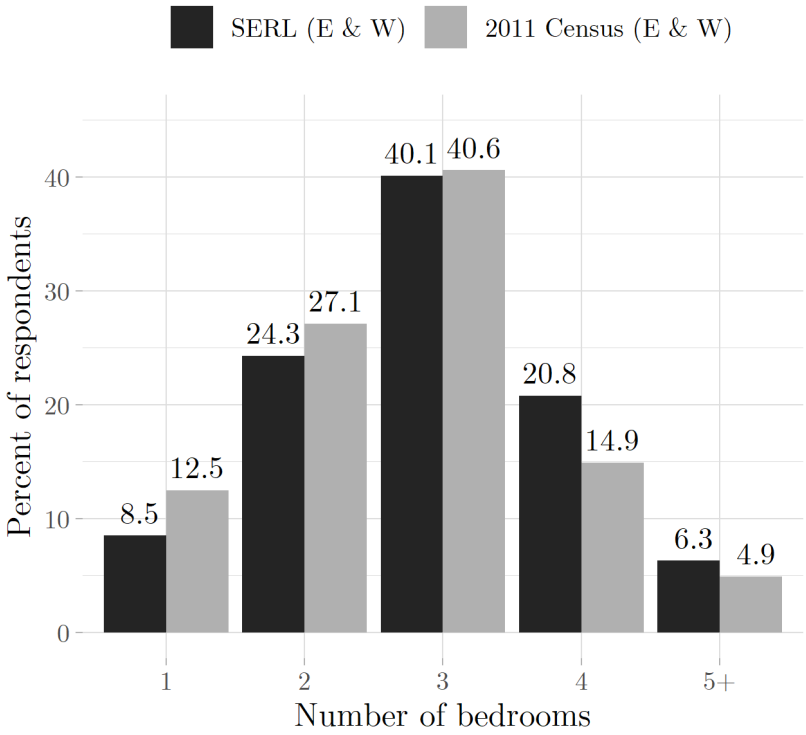

Figure 16: The number of bedrooms in SERL Observatory dwellings in England and Wales (E & W) compared to those reported in the 2011 Census (England and Wales only as bedroom data not available for Scotland).

Understanding whether a household can achieve their desired thermal comfort is important for understanding their heating choices. We can compare the results of asking "during the cold winter weather, can you normally keep

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

comfortably warm in your living room?" (question B7) with a similar question in Understanding Society Wave 10 which asks "in winter, are you able to keep this accommodation warm enough?" and merging their categories "refusal" and "doesn't apply" into one which we call "no answer" (Figure 17). We see similar results between the two, with the SERL respondents being a little more likely to report not being able to keep comfortably warm (6.5% compared to 4.8%, and more of the SERL participants did not answer).

A related question (B8) asks "do you have any problems with condensation, damp or mould in your home?". Problems with damp and mould can be indicators that dwellings are insufficiently ventilated, poorly constructed or under-heated, either because the heating system is incapable of providing sufficient warmth or because householders are unable or unwilling to heat their dwelling sufficiently to prevent these issues. The EHS reports only 3.4% of households with 'any damp' (rising or penetrating damp or condensation/mould) whereas in the SERL survey 26.1% responded yes to the question "do you have any problems with damp or mould in your home?" (compared to 70.2% who said no). This very large difference may be because of how the two surveys collect their data (self-reported in SERL; by a surveyor or telephone interview in EHS). Moreover, the wording between the two surveys is different in that EHS surveyors report instances of rising or penetrating damp, or serious condensation/mould growth, whereas SERL asks about any problems with damp or mould, which may result in a lower threshold for reporting of an issue.

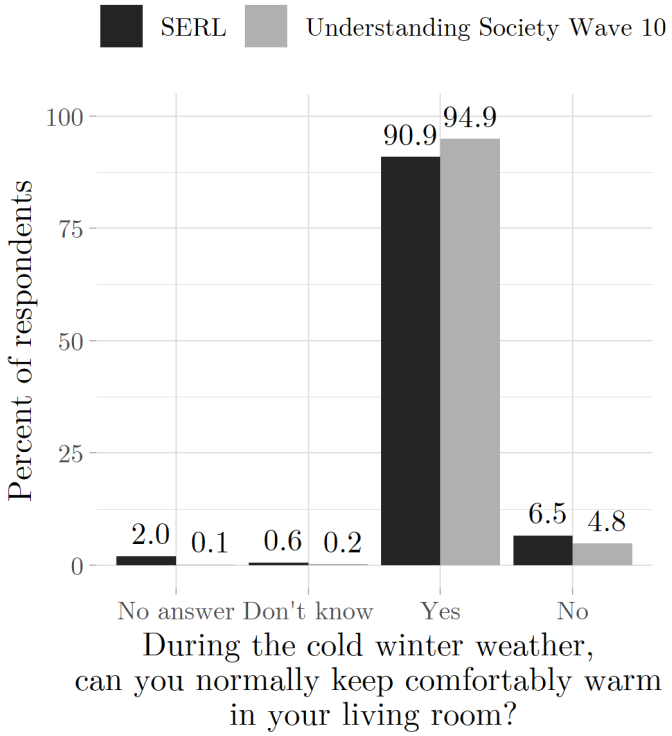

Figure 17: Comparing responses to SERL survey question B7 about being able to keep comfortably warm with a similar question in Understanding Society Wave 10.

For participants whose heating can be controlled by a temperature setting or smart device, question A5 asked "what temperature do you set your controller to in the winter months for the late afternoons or evenings? If you have more than one controller, choose what you would consider the main one". Figure 18 is a histogram of the responses between 10°C and 35°C, where the least popular choices have been merged into wider bars at the extremities for statistical disclosure control. The modal reported set point in the SERL dataset is 20°C, the same modal value as found by Shipworth *et al*. [27] in their study of 164 reported temperature set-points. Currently air conditioning is fairly rare (indeed less than 4% of respondents reported an air conditioning unit) and so temperature set points in summer were not part of the SERL survey.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

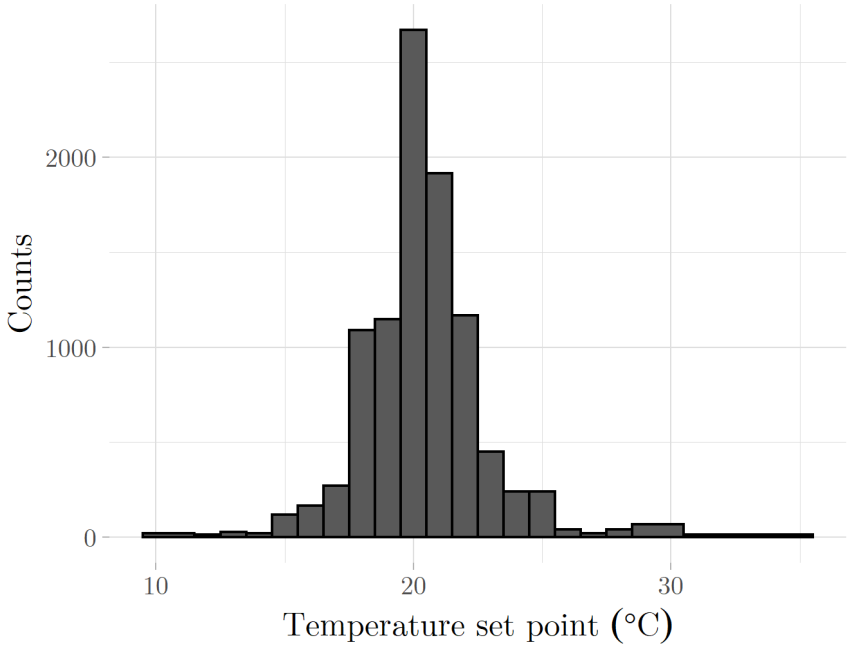

Figure 18: Self-reported winter afternoon temperature set points in °C for those who could control the temperature of their heating (question A5). Outliers are excluded and bars with fewer than 10 participants have been merged.

5.4. SERL Observatory participants with EPC data

Just over half of SERL participants have an EPC rating logged in our database, around 58% of those in England and Wales (we do not yet have Scottish EPC data). Figure 19 compares the EPC ratings of dwellings in the SERL sample with a national estimate from the English Housing Survey 2019 to 2020: headline report data [28]. It should be noted that the EHS data does not include Welsh homes unlike the SERL EPC dataset and the processes for generating EPCs for EHS dwellings are not identical to those used to generate standard domestic EPCs. In particular, EHS EPCs are carried out by a different group of assessors, their results are more likely to undergo quality control, and only RdSAP version 9.93 was used to calculate EPC ratings (depending on when a standard domestic EPC was generated different versions of SAP would have been used). Nonetheless, we include the following comparison of EHS and SERL EPCs as an indication of possible differences between the SERL Observatory sample and the wider population of English dwellings. When compared to the EHS EPC ratings, the SERL Observatory has roughly the same proportion of dwellings in EPC band D, F and G, while those with rating A/B and E are overrepresented and those with rating C are underrepresented in SERL. This will need to be borne in mind when generalising from studies that require an EPC rating.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

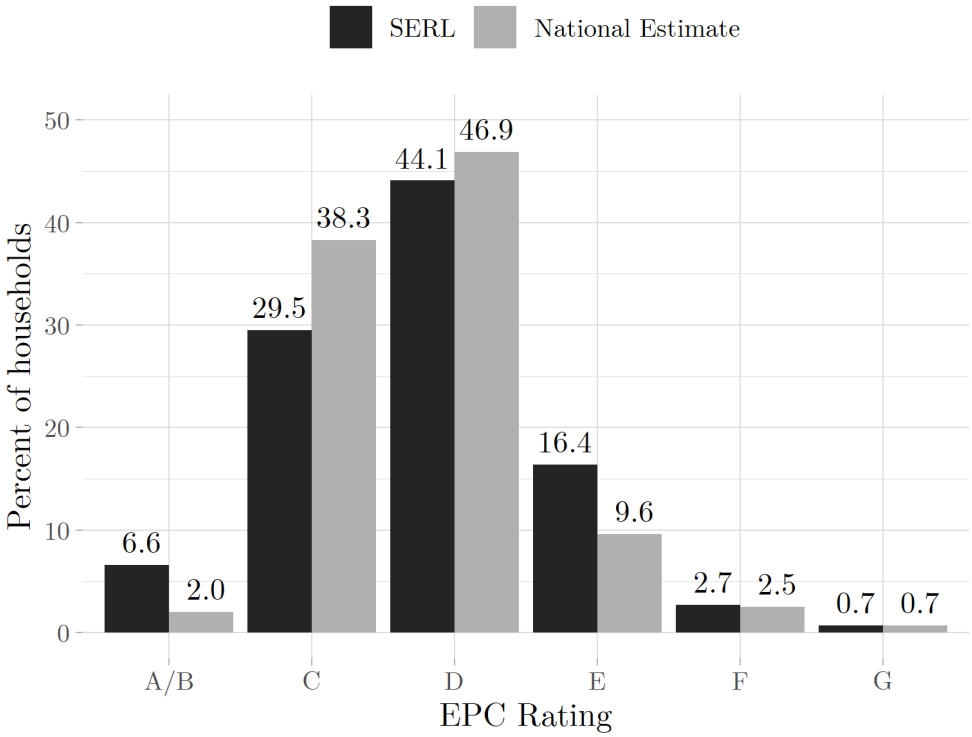

Figure 19: Percentage of dwellings in the SERL Observatory sample with each EPC rating (in England and Wales) compared to an estimate for England (EHS). Note that EPC ratings A and B have been merged as fewer than 10 households have EPC rating A.

Regionally, EPC prevalence in the SERL sample is fairly consistent within England; from 55.6% in the Midlands to 59.4% in the South West. The prevalence in Wales is higher at 62.3% (Figure B2 in Appendix B). Figure 20 shows the regional representation of the SERL sample with an EPC compared to the full SERL sample and the national estimate from the Address Base dataset[11]. Overall, the sample with an EPC is more representative regionally (in England and Wales) than the full SERL sample; for the North West, South West, West Midlands and Wales, filtering out homes without an EPC makes the sample at least 0.3 percentage points more regionally representative, while the opposite is true for the East Midlands (all other regions show very similar percentages). Future SERL Observatory editions will include EPC data in Scotland, but this has not been included yet for edition 3 as Scottish EPC data was only recently released in a machine-readable format.

[11] Since we do not yet have Scottish EPC data, the percentages are calculated excluding households in Scotland.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

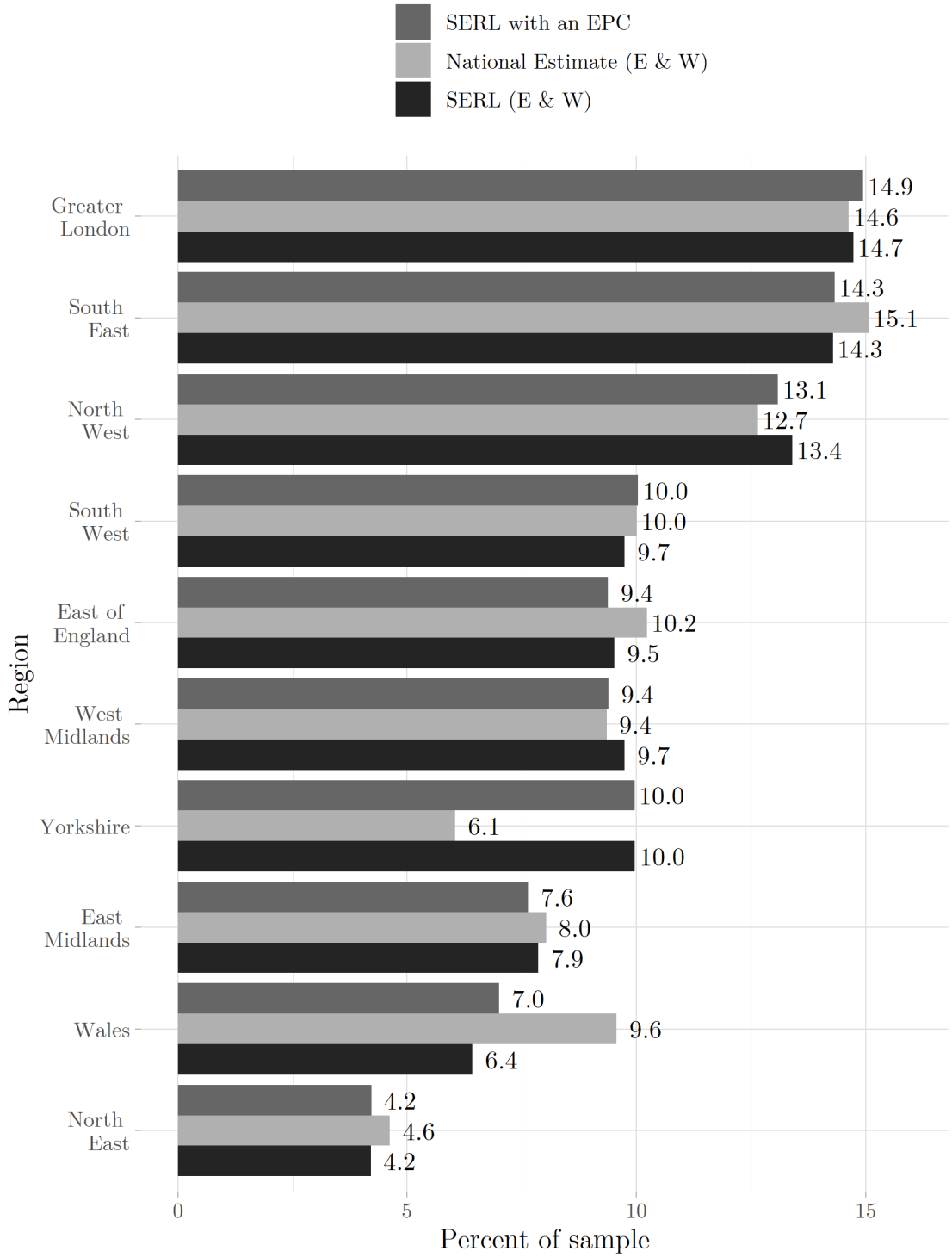

Figure 20: Regional split of dwellings in SERL (England and Wales), national estimate from Address Base (England and Wales), and the SERL sample of participants with an EPC (Scottish data to be included in future).

The difference between EPC prevalence among IMD quintiles within the SERL Observatory is significantly greater than the regional differences, with those in the areas of least deprivation far less likely to have an EPC (only 43.5%) than those with the most (56.3%) (Figure 21). This may be because rental properties (required to have an EPC) are more likely to be in areas with greater deprivation. Almost all IMD quintiles are less representative of the national distribution when filtering out homes without an EPC (Figure 22), with the filtered sample biased towards lower IMD quintiles.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

Note that we are not have national estimates for EPC ratings by IMD quintile, so it may be that our sample reflects the skew in EPCs by IMD quintile.

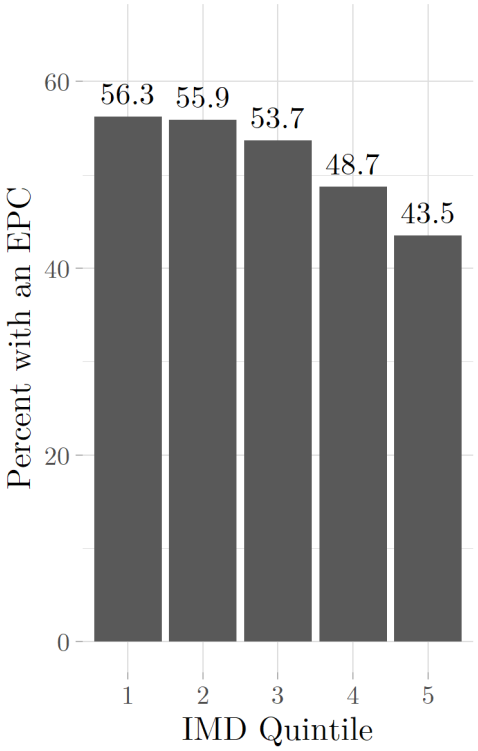

Figure 21: Prevalence of EPCs among SERL Observatory households by IMD quintile.

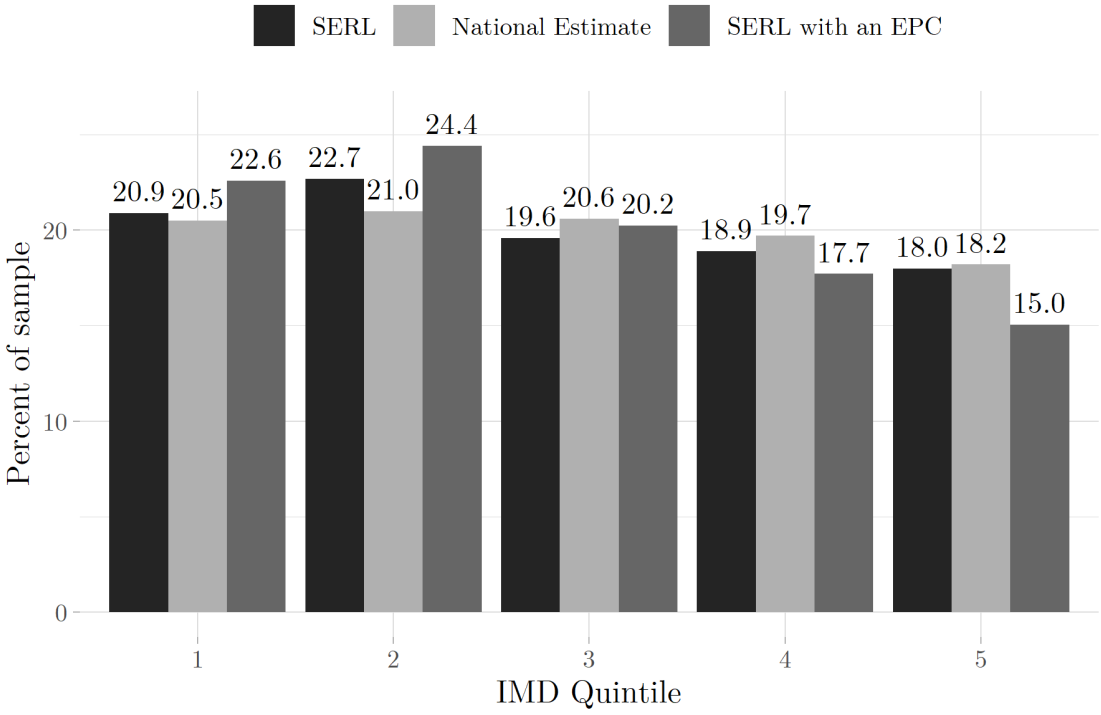

Figure 22: Representation of IMD quintiles in each sample: the full SERL sample (GB), a national estimate from Address Base (GB), and the SERL sample with an EPC (England and Wales).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

There are advantages for representativeness of EPC filtering; owner occupiers are substantially less overrepresented and private (and to a lesser extent social) renters are less underrepresented (Figure 23). In terms of dwelling type, the subsample of SERL dwellings with an EPC accurately represents semi-detached houses and purpose-built flats (previously over- and under-represented, respectively) and the overrepresentation of detached houses less than for the SERL survey respondent sample (Figure 24). In general, researchers should take care when reporting results to note the bias in the selected (sub)sample, in order to allow the generalisability of their results to be better understood.

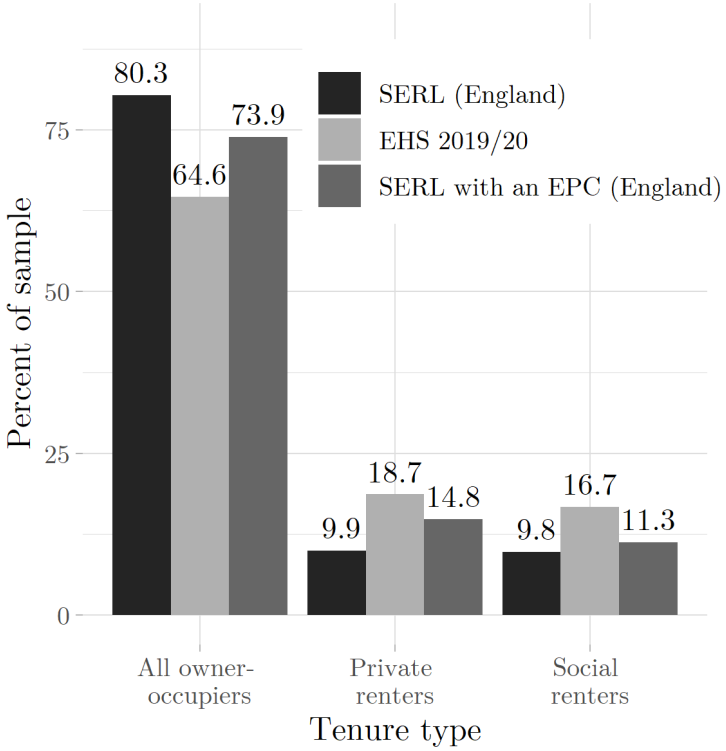

Figure 23: Tenure types among SERL survey respondents in England (counts rounded down to the nearest 20 for statistical disclosure control), EHS respondents and SERL survey respondents in England with an EPC.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

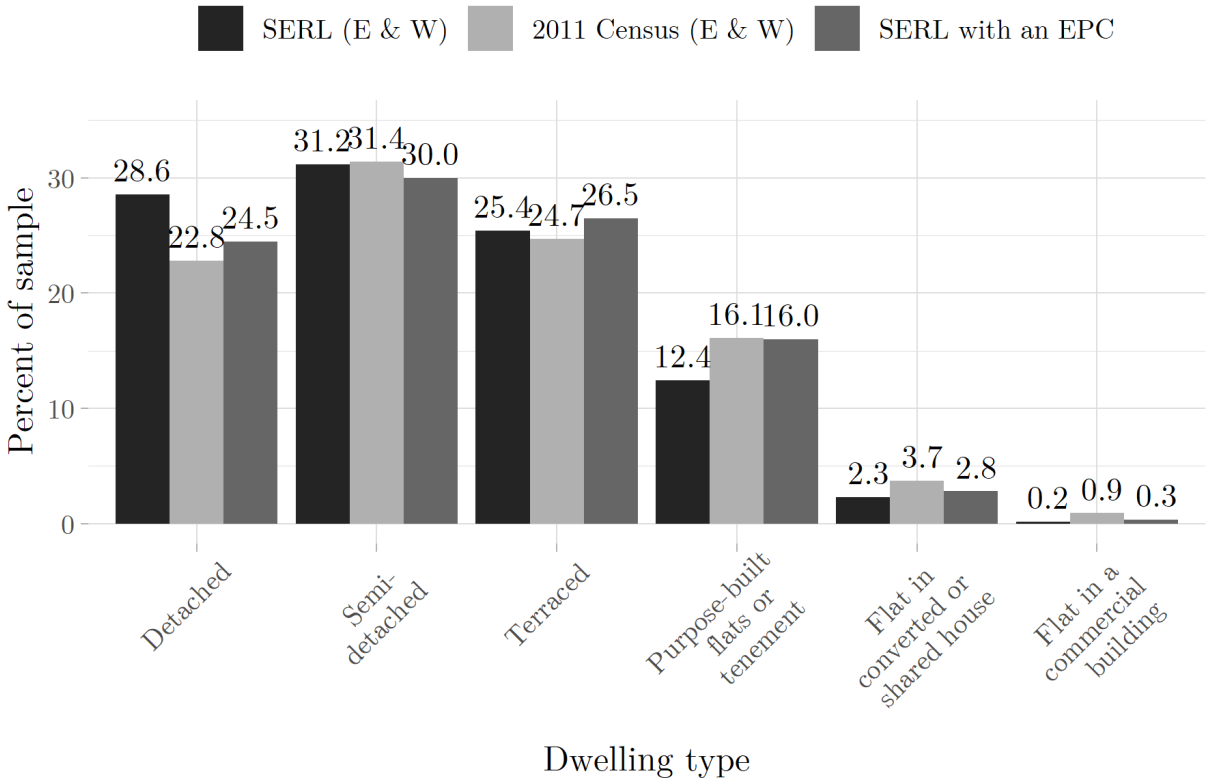

Figure 24: Dwelling type among SERL survey respondents (counts rounded down to the nearest 20 for statistical disclosure control), the 2011 Census respondents and SERL survey respondents with an EPC (all England and Wales only).

6. Usage notes and code availability

Access to the SERL Observatory dataset is provided via a secure virtual lab environment and is restricted to accredited researchers working on approved projects. This aligns with the 5 Safes protocols used by the UK Data Service (UKDS). Currently, SERL only provides controlled data via a secure lab environment.

The process to obtain access to controlled data is appropriately rigorous to comply with the data governance, privacy and ethics requirements described in Section 2.3. It is strongly recommended that researchers read the detailed documentation associated with the SERL catalogue record (SN 8666) provided as supplementary files and with the latest data edition in the UKDS data catalogue to ensure that SERL data is appropriate for their research project before starting the application process. In terms of timelines, the following is a rough guide to the application process:

1. Accredited researcher status (safe researcher training and exam): 1 month
2. University ethics approval: this varies by institution but allow at least 3-4 weeks
3. Project application (UKDS triage and SERL Data Governance Board review): 4-6 weeks.

Therefore, for new applicants it is recommended that at least 3 months is allowed to access the data. Full details and guidance on submitting an application are available on the UKDS website [29].

7. Conclusions

The SERL Observatory dataset comprises 13,321 households recruited to date with linked daily and half-hourly gas and electricity use, household survey, weather and dwelling energy performance data. The data will remain available on UKDS secure lab indefinitely, with ongoing longitudinal data collection until at least August 2022, and longer if funding allows. The role of SERL has been explicitly acknowledged by BEIS as a platform "allowing energy researchers to carry out valuable public interest work" [9] (page 11). The Public Interest Advisory Group on access to smart meter energy data final report further highlights the value of SERL as a "unique resource spanning gas and electricity and linking energy consumption to socio-demographic and other data" and recommending "that funding is provided for it to continue" ([30], page 6).

In this paper we have described the SERL Observatory dataset and detailed the key processes involved from participant sample selection, recruitment, data collection, analysis, provisioning and researcher access. Our three recruitment waves during 2019-2021 and stratified random sampling approach ultimately resulted in a very close representation of IMD quintiles (within 1.7 percentage points of the target in all quintiles), and reasonable regional representation, with the exceptions of Yorkshire (overrepresented) and Wales (underrepresented). Our sample with an EPC overrepresents the most energy-efficient dwellings (EPC rating A-C) compared to the EHS EPC distribution, although it should be noted that EHS EPC ratings are carried out slightly differently to standard domestic EPCs, which may influence this result. We have used the SERL survey data to identify sources of bias, which are likely to have been introduced by a combination of response bias and an uneven distribution of smart meter installations. The main sources of bias that we can detect with the data available are the overrepresentation of owner-occupiers (underrepresenting private and social renters), detached houses (underrepresenting flats), dwellings with a large number of rooms/bedrooms, those who are more energy conscious and those with the greatest self-reported financial wellbeing. Researchers should consider the change in sample bias introduced by filtering. For example, filtering out households without an EPC results in an overrepresentation of lower IMD quintiles and certain regions, but for some characteristics (such as owner-occupiers, renters, detached houses and flats), the filtered sample is more representative than the full SERL Observatory sample.

Ongoing research (such as [31], [32]) is demonstrating the value of the SERL linked data in a number of policy contexts and we encourage researchers to apply to use the SERL Observatory dataset (see Section 6 for details) to make the most of this unique UK research community resource.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Report S1: readme_serl_supplementary_files.pdf, Report S2: serl_climate_documentation_edition03.pdf, Report S3: serl_epc_documentation_edition03.pdf, Report S4: serl_main_recruitment_survey_copy.pdf, Report S5: serl_participant_summary_documentation_edition03.pdf, Report S6: serl_pilot_recruitment_survey_copy.pdf, Report S7: serl_smart_meter_data_quality_report_edition03.pdf, , Report S8: serl_smart_meter_documentation_edition03.pdf, , Report S9: serl_survey_documentation_edition03.pdf, Table S1: serl_survey_response_frequencies_edition03.csv.

Author Contributions: Conceptualization, S.E., T.O., D.S., E.M., E.W. and B.A.; methodology, S.E., T.O., D.S., E.M., E.W. and B.A.; investigation, E.W. and J.F.; software, E.W. and J.F.; validation, E.W. and J.F.; formal analysis, E.W.; data curation, E.W. and J.F.; writing—original draft preparation, E.W., J.F., M.P.; writing—review and editing, E.W., J.F., M.P., T.O., S.E., B.A.; visualization, E.W.; supervision, S.E., T.O.; funding acquisition, S.E., T.O. and D.S; Project administration, S.E.. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by EPSRC, grant number EP/P032761/1.

Data Availability Statement: Data is available in a secure lab environment for accredited researchers working on approved projects. Guidance on access to data is provided in Section 6.

Acknowledgments: There are over 30 individuals across 8 organisations in the SERL Consortium (University College London, the University of Essex (UK Data Archive), University of Edinburgh, Cardiff University, Loughborough University, Leeds Beckett University, the University of Southampton and the Energy Saving Trust) who have contributed to the development of SERL and thus the content of this paper. Particular thanks go to the SERL technical team at the UK Data Archive: Darren Bell, Deirdre Lungley, Martin Randall and Jacob Joy for software, data curation and dataset development; James O'Toole at UCL for project administration; Andrew ZP Smith (MaREI), Adam Cooper (UCL) and Abubakr Bahaj (University of Southampton) for conceptualization; and Ipsos MORI for participant recruitment design and implementation. Support from the SERL Advisory Board, Data Governance Board and Research Programme Board played a critical role in the establishment and ethical operation of SERL.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results, other than being members of the advisor board, whose role was to oversee the work.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

Appendix A: UK domestic energy datasets

Seven UK energy datasets that may be used to investigate granular (daily or finer) patterns of UK domestic energy use at the household level. Datasets are included according to the following criteria: in terms of variables, they include whole-home energy use for at least one fuel type (electricity or gas), at daily resolution or higher, and measurements of at least some factors which are likely to influence that energy use, e.g. contextual, building, occupant and/or appliance characteristics. In terms of coverage and size, they include households sampled from within Great Britain, and sample sizes of at least 100 homes to allow for investigation of the influence of home-level variation, relatively contemporary (covering a period that falls at least partly within last 10 years), for any duration. Datasets based on data from simulated homes rather than occupied by actual residents are not included. The table also includes only datasets that have been made available free at point of use for research purposes. Depending on the intended use case for the data, other suitable datasets may exist that are not listed here.

Table A1: Previously released energy datasets with comparable features to the SERL dataset, most recent first (by end of data collection). See Section 1 for criteria for inclusion of datasets described in this table. 'T' stands for 'temperature'. See below table for additional relevant notes. Datasets were identified from: Elam 2016 [33]*, CEEDS: The Centre for Energy Epidemiology Data Service 2017* [34]*, Pullinger et al 2021* [35]*, Georgia Tech n.d.* [36]*, authors' existing knowledge.*

	Smart Energy Research Lab (SERL) Observatory [37]	**Solent Achieving Value from Efficiency (SAVE) Data** [14], [38]	**IDEAL Household Energy Dataset** [15], [35]	**DEFACTO: Digital Energy Feedback and Control Technology Optimisation** [39]	**Smart Meter Energy Consumption Data in London Households** [40]	**Customer Led Network Revolution** [41]	**North East Scotland Energy Monitoring Project, 2010-2012** [12], [13]
Number of homes*	>13,000	>4,000	255	393++	5,567	~11,000	215
Sampling method**	Stratified random sample	Stratified random sample	Mixed methods, with quota sampling.	Stratified random sample	Participants of UK Power Networks Low Carbon London project: "balanced sample representative of the Greater London population"	CLNR trial participants	Purposive selection/case studies

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

Geographic coverage	Mainland England, Scotland and Wales	Hampshire, Southampton, Portsmouth, Isle of Wight	Edinburgh, Lothians and S. Fife, Scotland	English Midlands	London	GB	North East Scotland
Temporal coverage**	Aug. 2018 – Aug. 2022 and beyond	Jan. 2017 – Dec. 2018	Aug. 2016 – Jun. 2018	2015 – 2018	Nov. 2011 – Feb. 2014	2011 – 2014	2010 – 2012
Whole-home energy data (fuels and resolution) * **	Electricity (import and export) and gas, 30min	Electricity, 15min^{+}	Electricity, 1s. Gas, 1 reading per 1dm^3 or 1ft^3	Electricity, 2min. Gas, 30min	Electricity, 30min	Electricity, 30min	Electricity, 5min
Building and room data* **	Dwelling attributes	Dwelling attributes	Property type, age, entry floor, outdoor space. Room type, external doors and windows, floor area, height, radiators, thermostat presence – per room.	Property floor plan. Home energy survey, including a domestic energy assessment. EPC+ input data (project-collected). T per room.	None	Indoor temperature (homes with air source heat pumps only)	Temperature, 1 room
Occupant data	Occupant characteristics, sociodemographics, self-reported energy awareness and behaviours	Occupant characteristics and household behaviours	Sociodemographics, values, attitudes, self-reported energy awareness and behaviours, including occupancy, household income band and stability	Demographic and occupancy. Heating system and appliance usage. Self-reported zonal control usage	None	Mosaic consumer classification	Demographic, "psycho-social measures including individual environmental attitudes, household characteristics, and everyday behaviours"

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

Contextual data	IMD quintile, region, LSOA, EPC, 24 weather variables	Urban-rural classification; IMD decile; modelled % of LSOA in fuel poverty	Weather, Urban-rural classification	Weather	None	External temperature (homes with air source heat pumps only)	Urban-rural classification
Appliance data* **	Presence of 14 appliances	None	Inventory, presence of smart systems; T for boiler pipes, radiator pipes, hot water outlets, fires, cookers; Electricity for selected appliances, main sub-circuits.	Heating system details.	None	Homes with and without: solar PV (with automatic or with manual in-premises balancing); air source heat pumps; EVs. Electricity for EV charge points	None
Other data* **	Potential for linking	Time use diary	Tariffs, meter readings	Changes to property near end of trial vs start.	Tariff – 1100 approx. had a dynamic Time of Use tariff; remainder flat rate.	Tariff type, flat or Time of Use	Carbon footprint questionnaire
Availability	Accredited researchers on approved projects (see Section 6)	Registered users via UK Data Archive	Open access, CC BY 4.0	Contact data controllers at Loughborough University regarding access	Open access, unspecified license	Open Access, CC BY-SA 4.0	UK Data Archive, safe-guarded

* Maximum number of homes that include some whole-home electricity and/or gas sensor/smart meter data. Not all variables and/or the full time period may be available for all of those homes.

** Refers to sample and period with electricity and/or gas sensor/smart meter data collection.

*** Key to abbreviations: E: Electricity; G: Gas; T: Temperature; H: Humidity; L: Light level; EV: Electric Vehicle

+ SAVE dataset: 10s electricity data available for staff and students at the University of Southampton on request [14].

++ DEFACTO dataset: Information in table refers to main sample only (not pilot homes).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

Appendix B: Supporting Tables and Figures

Table B1: Number of participants in each region and percentage of the SERL Observatory sample.

Region	Total	Percentage
East Midlands	952	7.1%
East of England	1151	8.6%
Greater London	1773	13.3%
North East	508	3.8%
North West	1610	12.1%
Scotland	1297	9.7%
South East	1716	12.9%
South West	1171	8.8%
Wales	779	5.8%
West Midlands	1170	8.8%
Yorkshire	1194	9.0%

Understanding Society Wave 10 asks about tenure but only has one 'rented' option (rather than splitting into privately and from the council/local authority/housing association). This survey also gives multiple options which we combine into 'null response': 'inapplicable', 'refusal', 'don't know', and 'other'. The SERL 'null response' is the percent who did not answer. The full results are shown in Figure B1.

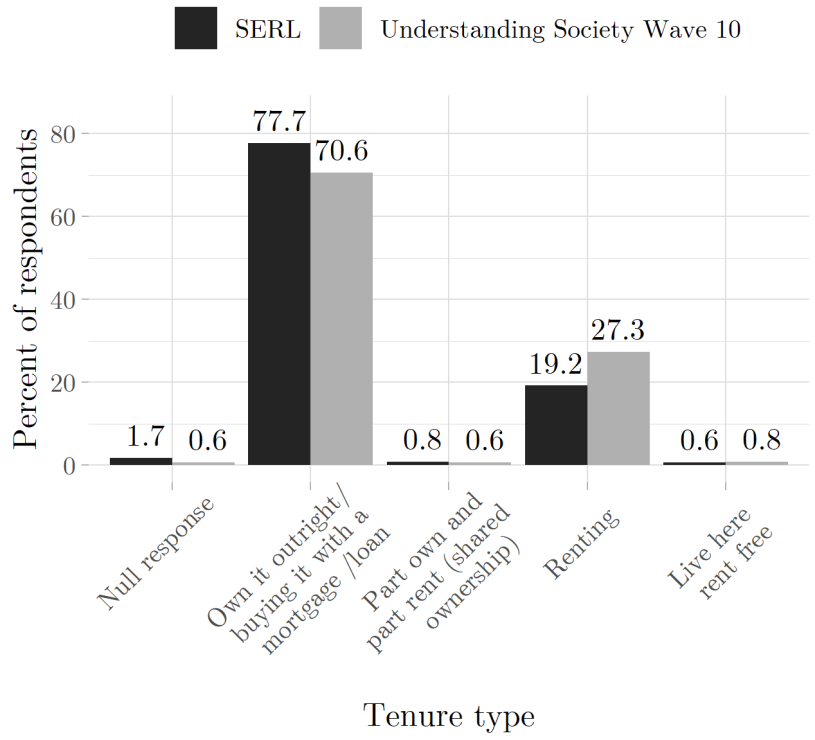

Figure B1: Comparing SERL Observatory tenure types (question B4) with the Understanding Society Wave 10 survey.

Figure B2 shows the percentage of SERL households in each region (excluding Scotland) with an EPC. Scottish EPC data will be included in future SERL Observatory editions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2021 doi:10.20944/preprints202109.0370.v1

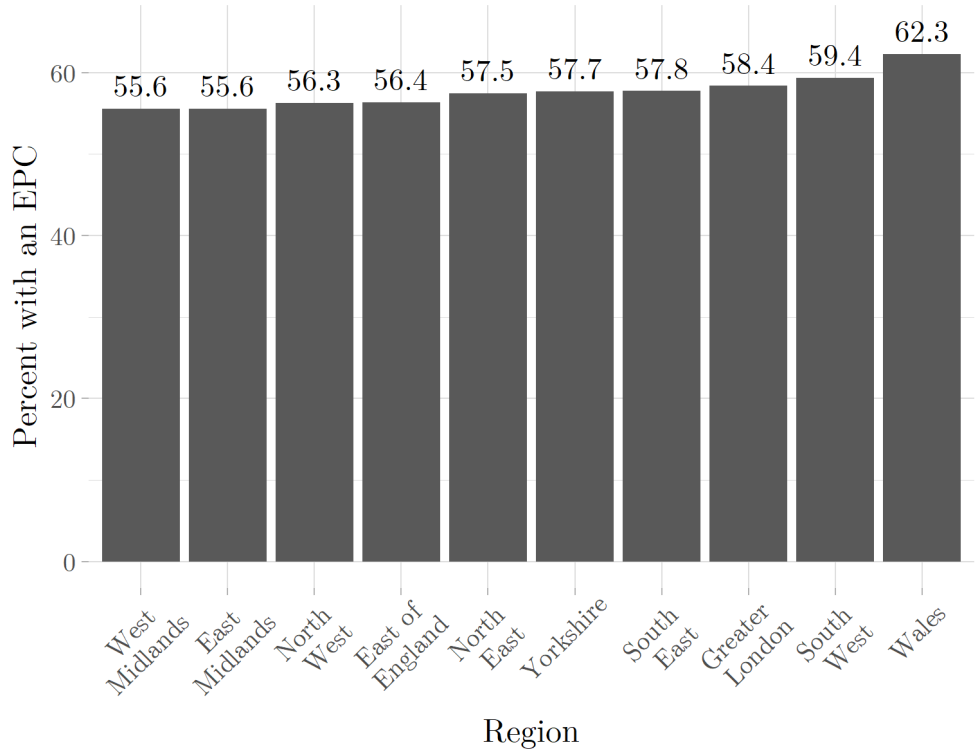

Figure B2: Prevalence of EPCs among SERL Observatory households by region.

References

[1] Department for Business Energy & Industrial Strategy, *Annex: 2019 UK Greenhouse Gas Emissions, final figures by end user and fuel type*. 2019.

[2] National Grid, "System Operability Framework 2015," 2015.

[3] G. Strbac, "Demand side management: Benefits and challenges," *Energy Policy*, vol. 36, pp. 4419–4426, 2008.

[4] O. Golubchikov and K. O'Sullivan, "Energy periphery: Uneven development and the precarious geographies of low-carbon transition," *Energy Build.*, vol. 211, p. 109818, 2020.

[5] E. Webborn and T. Oreszczyn, "Champion the energy data revolution," *Nat. Energy*, vol. 4, no. 8, pp. 624–626, 2019.

[6] I. Hamilton, T. Oreszczyn, A. Summerfield, P. Steadman, S. Elam, and A. Smith, "Co-benefits of Energy and Buildings Data: The Case for supporting Data Access to Achieve a Sustainable Built Environment," *Procedia Eng.*, vol. 118, pp. 958–968, 2015.

[7] A. J. Summerfield and R. Lowe, "Challenges and future directions for energy and buildings research," *Build. Res. Inf.*, vol. 40, no. 4, pp. 391–400, 2012.

[8] E. & I. S. Department for Business, *Smart Meter Statistics in Great Britain: Quarterly Report to end March 2021*. 2021.

[9] Department for Business Energy & Industrial Strategy, "Digitalising our energy system for net zero: Strategy and Action Plan 2021," London, 2021.

[10] "Smart Energy Code," vol. 2015, no. September 2013, 2015.

[11] Smart Energy Research Lab, "Welcome to the Smart Energy Research Lab." [Online]. Available: www.serl.ac.uk. [Accessed: 18-Jun-2020].

[12] T. Craig and I. Dent, "North East Scotland Energy Monitoring Project, 2010-2012, Study Level Documentation," UK Data Service, 2016.

[13] T. Craig and I. Dent, "North East Scotland Energy Monitoring Project, 2010-2012 [Data Collection]. SN: 8122," 2017. .

[14] T. Rushby, B. Anderson, P. James, and A. Bahaj, "Solent Achieving Value from Efficiency (SAVE) Data, 2017-2018 [data collection]. SN: 8676." UK Data Service, 2020.

[15] N. Goddard *et al.*, "IDEAL Household Energy Dataset." Edinburgh DataShare, 2021.

[16] A. Cooper, D. Shipworth, and A. Humphrey, "UK Energy Lab: A feasibility study for a longitudinal , nationally representative sociotechnical survey of energy use Synthesis Report," 2014.

[17] E. Webborn, E. McKenna, S. Elam, B. Anderson, A. Cooper, and T. Oreszczyn, "Increasing response rates and improving research design: Learnings from the Smart Energy Research Lab in the United Kingdom," *Energy Res. Soc. Sci.*, vol. (in press), 2021.

[18] E. McKenna, M. Frerk, and S. Elam, "Smart Energy Research Lab (SERL) Data Governance Board (DGB) Terms of reference," 2020.

[19] L. Corti and R. Welpton, "Access to sensitive data for research: 'The 5 Safes,'" *Data Impact blog*, 2015. [Online]. Available: https://blog.ukdataservice.ac.uk/access-to-sensitive-data-for-research-the-5-safes/. [Accessed: 13-Sep-2021].

[20] J. Crawley, P. Biddulph, P. J. Northrop, J. Wingfield, T. Oreszczyn, and C. Elwell, "Quantifying the measurement error on England and Wales EPC ratings," *Energies*, vol. 12, no. 18, 2019.

[21] D. Jenkins, S. Simpson, and A. Peacock, "Investigating the consistency and quality of EPC ratings and assessments," *Energy*, vol. 138, pp. 480–489, 2017.

[22] R Core Team, "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria, 2019.

[23] Office for National Statistics, "2011 Census," 2011. [Online]. Available: https://www.ons.gov.uk/census/2011census. [Accessed: 02-Sep-2021].

[24] Ordnance Survey, "AddressBase." [Online]. Available: https://www.ordnancesurvey.co.uk/business-government/products/addressbase. [Accessed: 02-Sep-2021].

[25] Ministry of Housing Communities & Local Government, "English Housing Survey," 2021. [Online]. Available: https://www.gov.uk/government/collections/english-housing-survey. [Accessed: 02-Sep-2021].

[26] Institute for Social and Economic Research, "Wave 10 data released," *Understanding Society*, 2020. [Online]. Available: https://www.understandingsociety.ac.uk/2020/11/26/wave-10-data-released. [Accessed: 02-Sep-2021].

[27] M. Shipworth, S. K. Firth, M. I. Gentry, A. J. Wright, D. T. Shipworth, and K. J. Lomas, "Central heating thermostat settings and timing: building demographics," *Build. Res. Inf.*, vol. 38, no. 1, pp. 50–69, 2010.

[28] Ministry of Housing Communities & Local Government, "English Housing Survey 2019 to 2020: headline report - Section 2 housing stock tables," 2020.

[29] UK Data Service, "Apply to Access Smart Energy Research Lab Data," *UK Data Service*, 2021. [Online]. Available: https://ukdataservice.ac.uk/find-data/access-conditions/secure-application-requirements/apply-to-access-serl/. [Accessed: 08-Sep-2021].

[30] M. Frerk, "Smart Meter Energy Data: Public Interest Advisory Group Final Report - Phase 2," 2021.

[31] E. McKenna *et al.*, "Explaining daily total energy demand in British housing using linked smart meter and socio-technical data in a bottom-up statistical model," *OSF Prepr.*, no. 15 Sept, 2021.

[32] G. Huebner *et al.*, "Self-reported energy use in UK homes during the first COVID-19 lockdown: A survey study," *SocArXiv*, no. 14 July, 2021.

[33] S. Elam, "Smart Meter Data and Public Interest Issues – The National Perspective: Discussion Paper 1. Annex A – Existing data," 2016.

[34] CEEDS: The Centre for Energy Epidemiology Data Service, "CEE Data Asset Register (2017-08)." RCUK Centre for Energy Epidemiology, London, 2017.

[35] M. Pullinger *et al.*, "The IDEAL household energy dataset, electricity, gas, contextual sensor data and survey data for 255 UK homes," *Sci. Data*, vol. 8, no. 1, Dec. 2021.

 doi:10.20944/preprints202109.0370.v1

[36] Georgia Tech, "Smart Meter Data Portal." [Online]. Available: https://smda.github.io/smart-meter-data-portal/. [Accessed: 26-Aug-2021].