Explaining daily total energy demand in British housing using linked smart meter and socio-technical data in a bottom-up statistical model

Eoghan McKenna a, Jessica Few a, Ellen Webborn a, Ben Anderson b, Simon Elam a, David Shipworth a,Adam Cooper c, Martin Pullinger a, Tadj Oreszczyn a

a UCL Energy Institute, 14 Upper Woburn Place, University College London, WC1H 0NN, UK

b Faculty of Engineering and Physical Sciences, University of Southampton, Southampton, UK

c Department of Science, Technology, Engineering and Public Policy, Shropshire House, Capper Street, London WC1E 6JA, UK

# Abstract

This paper investigates factors associated with variation in daily total energy consumption in domestic buildings using linked pre-COVID-19 smart meter, weather, building thermal characteristics, and socio-technical survey data covering appliance ownership, demographics, behaviours, and attitudes for a sub-sample of 617 British households selected from the Smart Energy Research Laboratory (SERL) Observatory panel.

Linear mixed effects modelling resulted in marginal/conditional R2 of 0.68/0.83 and root mean squared error of 17.7 kWh/day, for daily gas and electricity use combined. Increased daily energy consumption was significantly associated (*p*-value<0.05) with: households living in buildings with larger floor area, more rooms, that are older, have lower energy efficiency, and experience colder or less sunny weather; households with more adult occupants, more children, older adult occupants, fewer adults with qualifications, higher heating temperature setpoints, that do not try to save energy, and that do not put on more clothes rather than turning the heating on.

The results demonstrate the value of smart meter data linked with contextual data for improving understanding of energy demand in British housing. Accredited UK researchers are invited to apply to access the data which has recently been updated to include over 13,000 households from across Great Britain.

**Key words:** building; energy; heating; gas; electricity; demand; consumption; household; residential; domestic; smart meter; daily; longitudinal; regression; mixed effects; random effects; survey; energy performance certificate; weather; temperature; solar radiation; building physics; sociodemographic; occupant; behaviour; attitudes.

# Introduction

Increasingly governments are pledging to reduce greenhouse gas emissions to net zero by 2050 [1]. The building sector requires rapid decarbonisation of energy supply and wide-spread reduction in energy demand through improvements in energy efficiency, changes in behaviour and avoided energy use [1,2]. A critical starting point to achieve this is an effective characterisation of energy demand in buildings i.e. ‘what norms, values, preferences and structural factors shape energy demand?’ [3]. Cooper has emphasised the need for better integration of research approaches across social and physical science research for energy policy impact [4] and provides a conceptual framework for reasoning how to integrate data validly from different physical and social sources to enable socio-technical research [5]. This integration is required to better explain patterns in demand, identify the factors which are associated with greatest impact on demand, and to better inform effective policy instruments targeted at improving energy efficiency or changing occupant behaviour [4,6]. Moreover, effective characterisation can enable improved predictions of demand. This could reduce demand in buildings if used to identify (and potentially reward) changes in demand (e.g. due to an intervention, or in response to a tariff or energy efficiency installation). Prediction can also be used to diagnose problems such as malfunctioning heating systems, poor quality build, or energy waste in the form of heating or lighting in unoccupied buildings.

A greater understanding of energy demand in buildings has been impeded by limited data about energy demand and its influencing factors [7–9]. The Smart Energy Research Lab (SERL) is a five-year UK research council funded project which aims to address this by bringing together, for the first time, half-hourly resolution household-level electricity and gas demand data with detailed socio-technical and weather data for a representative sample of over 13,000 households in Great Britain (GB) (the ‘SERL Observatory’). In this respect the data captures a much wider array of energy demand co-variates *and* more detailed energy use data than has previously been reported in the literature.

The first aim of this paper is therefore to evaluate the SERL Observatory as a data resource to improve current characterisations of household-level energy demand. Linear regression is commonly used in the literature to characterise household demand given multi-variate demand-side datasets. This is usually done in two ways; first by assessing the overall *explanatory power* of an appropriately validated statistical model applied to the data, usually in the form of the R2/adjusted R2 or coefficient of determination. High explanatory power implies that the data includes appropriate variables, and the model captures appropriate relationships between them, such that variation in the variable of interest can be explained given the model and data. The model can be applied to other data and tested for prediction or forecasting purposes. Second, studies scrutinise the results of the model to identify specific variables which are statistically significantly associated with variation in the variable of interest *and* which have a substantively significant effect. Such variables can be interpreted as important factors related to household-level demand, leading to a more detailed understanding of residential demand, and can inform policies aimed at targeting such key factors and reducing demand in future. This leads to our first two research questions:

*1: What is the overall explanatory power of SERL Observatory data with respect to variation in household-level daily residential energy consumption and does this improve on studies reported in the literature?*

*2: Which variables observed in SERL Observatory data are most strongly, and statistically significantly, associated with household-level daily residential energy consumption and do these confirm and extend results reported in the literature?*

The SERL Observatory links energy consumption data from smart meters (at daily and half-hourly resolution) with three contextual datasets (described in more detail later):

* Basic data: dwelling region, local area Index of Multiple Deprivation[[1]](#footnote-2) (IMD) for 2019, and local area hourly weather variables;
* SERL survey: occupant-reported household-specific sociodemographic characteristics and energy saving behaviour, and some building-specific physical characteristics;
* EPC (Energy Performance Certificate) data: building-specific physical and thermal characteristics.

These three datasets have different levels of availability: all households in the SERL Observatory have basic data, around 80% have complete SERL survey data, and approximately half have EPC data as only about half of British properties have an EPC. Researchers using SERL data are therefore presented with a choice: to increase sample size but reduce contextual data, or decrease sample size and increase contextual data. Determining the usefulness of the datasets separately and together is therefore important for the overall objective of characterising demand. This leads to our final research question:

*3: What is the additional value of the EPC and SERL survey contextual data beyond that of the basic data?*

We answer this question by investigating how the explanatory power of the model changes with different levels of contextual data*.*

Our analysis uses the SERL Observatory Edition 2 dataset [10]. This contains data from almost 5000 households and energy demand data from August 2018 to October 2020. Data collection is ongoing and subsequent editions will be updated with this newly collected data. The first coronavirus lockdown in GB started on 23rd March 2020, meaning Edition 2 includes data from before the onset of the coronavirus pandemic. It is important to understand the impact of the coronavirus pandemic on residential energy consumption in buildings and what constitutes the post-pandemic ‘new normal’, and the SERL Observatory is a well-suited data resource to do this and currently supports several research projects investigating the effects of the pandemic. This paper aims to provide a foundation to this forthcoming research by seeking to understand and characterise residential energy consumption *pre-coronavirus*.

Given the requirements of full contextual data availability from waves 1 and 2 only, and the focus on the pre-lockdown period, the resulting sample of households analysed here is a relatively small subsample (N=617) compared to the number of households that will be available in later editions of the SERL Observatory (Edition 3 increases the sample size to >13,000). These results should therefore be seen as an initial analysis and should be interpreted with caution. In particular, results should not be viewed as representative of the future full-size SERL Observatory sample, nor indeed the GB population.

# Literature review

A substantial body of existing literature investigates the factors which shape household energy use. The first section below describes the literature on relevant factors which shape energy use. The remaining sections describe relevant literature on quantitative approaches to modelling building energy demand. The literature has informed the development of the SERL Observatory dataset and the analytical methods used in this study.

## Factors influencing household energy demand

Household energy demand can be viewed as the outcome of occupants making use of the energy-using appliances and equipment in their home, largely through everyday activities such as cleaning, food preparation, leisure and keeping warm (or cool). At the population level and over periods of years, structural drivers can be seen as having a major effect on average energy use for these activities, with policy-driven incremental energy efficiency improvements in technology often being offset by an increasing intensity of appliance use [11,12]. The development and diffusion of new technologies and more radical changes in social norms and expectations can also lead to more significant changes in the average energy intensity of such practices [13]. Inter-household variation in energy use at shorter timescales is then a matter of variation between households, the currently available technologies they have, and how they make use of them. Multiple studies have attempted to identify which aspects of occupants, of their activities, and of the technologies they use and buildings they occupy, are most important for explaining inter-household variation in final energy demand. Huebner et al [8] found building characteristics, particularly size, type and energy performance rating (as provided by EPCs), dominate in explaining between-household variation in energy use in a sample of contemporary English households, with household size (number of occupants) also important, as well as the length of the heating season and reported beliefs about climate change.

Other quantitative studies unpack overall energy use. Gram-Hanssen [12] separately investigates energy used for heating and energy used for appliances and lighting, in Danish households. Drawing on multiple data sources and analytical approaches, she concludes that user behaviour (including appliance ownership) is a more substantial factor shaping energy used for appliances and lighting than is appliance energy efficiency, noting, for example, that energy use in physically similar houses can vary by a factor of 5. Energy use for heating meanwhile is found to be roughly equally explained by building characteristics, including size and age, and by user behavioural factors, whilst sociodemographic characteristics (age, income and education) explain very little, indicating that they only weakly correlate with a person's heating behaviours.

Many other studies focus on a single fuel type rather than end use. Jones et al. [14] provided a literature review of nearly 40 empirical studies of household electricity use, identifying 62 factors that potentially affect it, with 20 "found to unambiguously have a significant positive effect on electricity use" (defined by the authors as the number of papers confirming a positive effect being more than three higher than the number finding a negative or non-significant effect). These 'unambigious' variables were classed by the authors into socio-economic factors (more occupants, presence of teenagers, higher income and higher disposable income), characteristics of the dwelling (older dwellings, and higher number of rooms or number of bedrooms, or larger total floor area; presence of an electric space heating system, air-conditioning and/or an electric water heating system) and appliance-related factors (higher number of appliances, ownership of: desktop computers, televisions, electric ovens, refrigerators, dishwashers, tumble dryers; greater use of: washing machines, tumble dryers). The categorical variables 'age of household reference person' and 'level of detachment of the building' also significantly affected electricity use. Further quantitative studies aim to specifically consider the influence of occupant behaviour, by combining time use data and electricity use data. Satre-Meloy et al [15] find that variation between occupants in when and how electricity-using activities are performed does have a statistically significant effect on energy use, at least over the course of the day, finding from their own data and a review of previous studies that quotidian activities related to chores, food consumption and preparation, and leisure are particularly high energy intensity, and sleep and rest low intensity.

Regarding heating use, a review by Wei et al [16] of 41 papers found 27 factors identified in them as affecting space heating behaviour in residential buildings, concluding the following eight factors 'unambiguously' influenced it (using the same definition of unambiguous as above): "outdoor climate, dwelling type, room type, house insulation, type of temperature control, occupant age, time of day and occupancy".

Overall, the literature provides evidence that inter-household variation in energy use is related to building and appliance characteristics, occupant sociodemographics, behaviours, and contextual factors around climate, indoor conditions and time. Although existing studies provide some insight into which factors within these broad classes are 'unambigiously' important, Wei et al [16] note that the literature does not definitively rule out the influence of any factor that has been studied. Huebner et al [8] highlight that limitations in measurement methods, particularly for measuring behaviours, and collinearity and interaction effects between variables, can lead to factors appearing to have non-significant effects or being excluded from models, while Jones et al. [14] note that the often incomplete contextual information about sample characteristics (such as the fuel type used in the dwellings for space heating and cooling and water heating, or if there was mechanical ventilation) could explain some of the conflicting results found between studies regarding the influence of certain factors. In sum, there is value in continued research to investigate the effects of a wide range of variables within these broad classes of building, appliance, occupant and contextual factors.

## Characterising energy demand in buildings

Characterising building energy demand is an active field of research employing a wide range of methods, depending on the data available and research objectives. Swan and Ugursal [17] provide a taxonomy of residential energy demand modelling approaches, grouping them into two broad categories of ‘top-down’ (a ‘macro’ approach where the housing stock as a whole is usually the unit of analysis) and ‘bottom-up’ (a ‘micro’ approach where the basic unit of analysis is usually individual dwellings), with the latter further sub-categorised into ‘statistical’ and ‘engineering’ methods. As this paper aims to characterise individual households, we adopt a bottom-up approach. Statistical regression is a common bottom-up approach that, while requiring large, detailed datasets, offer simple implementation and relatively easy interpretability. Jones et al. [14] provide a recent systematic review of studies using regression methods to explain energy demand in residential buildings. Satre-Meloy et al. [18] provide a complementary and updated summary of the literature. Rather than duplicating these works, we draw broad observations relevant to the present work from the literature. The focus is on studies that used statistical approaches to characterise energy demand in residential buildings. Detailed reviews of alternative ‘bottom-up’ approaches (e.g. engineering, artificial neural networks) can be found in [19–22].

## Explanation versus prediction

Multiple linear regression using ordinary least squares (OLS) is a technique commonly used in studies seeking to characterise energy demand in buildings using linked contextual data [18,23–26]. However, OLS relies on an assumption of independent observations which reduces its appropriateness for longitudinal data, in which there are repeated observations of individual cases [27]. Anderson et al. [26] used a linear mixed effects model to address this in their study of daily electricity consumption using daily aggregates of sub-half-hourly household level energy use, similar to the current study.

Recent advances in statistical learning [28] have resulted in the emergence of new techniques in this field of research. There has been increased interest in techniques such as tree-based methods, support vector machine, and artificial neural networks [21,22]. These can be considered more ‘flexible’ than OLS because they allow non-linear relationships between variables. However, increased flexibility can come at a cost, with greater risk of over-fitting, increased model variance error, and potentially less interpretability [21,29]. These techniques tend to be more suitable for prediction, rather than inference which is the primary interest of this work.

Jones et al’s review [14] found that at least 62 factors have been studied that potentially affect residential electricity demand. Energy demand in buildings can therefore be characterised by its large number of potentially influential factors. Studies seeking to characterise demand are therefore often faced with ‘dense’ models i.e. with many explanatory variables. While adding more variables to a model can increase its overall explanatory power, this can be accompanied by a reduction in model interpretability and the reliability of estimates for individual variable coefficients (e.g. due to multi-collinearity). Increased model complexity can also result in over-fitting and a decrease in the model’s predictive power [29].

Numerous techniques have been developed to deal with this issue and reduce the complexity of the model by selecting a subset of the total number of variables to include in the final model [30] and some of these have been applied in the field. For example, Kavousian et al. [31] use forward stepwise variable selection, while Huebner et al. [8] and Satre-Meloy et al. [18] use regularisation methods (or ‘shrinkage’ or ‘penalised’ regression). These techniques can be useful for improving interpretation and, depending on the nature of the underlying data, can also improve model prediction. As the present paper is intended to determine the explanatory power of the SERL data in total, these techniques are not applied here, although future research is likely to do so.

## Heating demand and gas meter data

In GB natural gas is widely used for space and water heating and cooking e.g. 86% of dwellings in England supplied by the gas grid [32]. Moreover, heating demand is strongly weather dependent, and so it is crucial to understanding how total domestic energy demand changes over time. Therefore, observing both gas and electricity demand is necessary to achieve a data-driven characterisation of *total* residential energy consumption in GB dwellings *including heating* (note that cooling is currently very uncommon in UK homes) where gas and electricity are the only fuel sources. In the case of the 14% of English dwellings that are not connected to the gas network, data on oil, LPG and solid fuel use would also be required but is not collected by SERL. We note, however, the relative difficulty of accessing gas demand data compared to electricity data (smart electricity meters are more widespread than smart gas meters [33] and it is easier to retro-fit sensor equipment to measure electricity demand data than it is for gas demand). This is reflected in the literature, which predominantly focuses on analysis of electricity demand compared to gas demand, even in countries where gas demand is present [15,23,25,34–36]. One of this paper’s key contributions is that we analyse electricity and gas data (where used by the household as discussed below) thereby focussing on *total* domestic energy consumption including space and water heating for these households.

## Temporal resolution of demand data

As noted above, the majority of studies in the literature focus on data of relatively low temporal resolution i.e. monthly, seasonal or annual summaries [14]. Studies focussing on daily or higher-resolution data are comparatively rare, presumably because of the relative difficulty of accessing large, high-resolution energy meter data sets which also have the necessary linked contextual data to perform robust regression analyses [9]. Table 1 summarises key characteristics of studies from the literature chosen for their relevance (focus on household-level residential energy consumption, use of regression, household-level contextual data, annual or higher time resolution) and shows how much of the variation in demand (the ‘coefficient of determination’ or R2) was explained by their models.

Table 1 – summary of characteristics and key results of previous relevant studies.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Study** | **Data source** | **Country or area** | **Sample size (N)** | **Resolution of demand data** | **Observes heating/ cooling?** | **Contextual data** | **Coefficient of determination (R2/Adjusted R2)** |
| [24] | Korea Energy Economics Institute survey | Korea | 2436 | Annual | Yes | Building physical characteristics, socio-demographics, appliance usage | R2: 0.009 - 0.017 |
| [8] | Energy Follow-Up Survey | England | 924 | Annual | Yes | Building physical characteristics, socio-demographics, heating behaviour, attitudes and other behaviours | Adjusted R2: 0.44 |
| [18] | City of Palo Alto Utilities survey | Palo Alto, California | 1008 | Annual | Yes | Building physical characteristics, socio-demographics, energy literacy, attitudes | R2: 0.373-0.398 |
| [35] | Smart Grid Smart City | New South Wales | 3446 | Annual | No (no gas meter data) | Building physical characteristics, socio-demographics, appliances | Adjusted R2: 0.55 |
| [31] | Convenience sample | Silicon Valley, California | 952 | Averaged over a period of 238 days | Yes | Weather, building physical characteristics,  appliances, and behaviour | Adjusted R2: 0.43-0.68 |
| [25] | Irish Commission for Energy Regulation's (CER) Smart Metering Electricity Customer Behaviour Trials | Ireland | ~4200 | Averaged over a period of 6 months | No (no gas meter data) | Building physical characteristics, socio-demographics, appliances | R2: 0.32 |
| [23] | Convenience sample | Japan | 740 | Monthly averaged demand | No (no gas meter data) | Weather, building and heating system information, household and appliance ownership and usage | Adjusted R2: 0.18-0.60 |
| [37] | Smart Grid Smart City | New South Wales | 3446 | Daily peak demand | No (no gas meter data) | Building physical characteristics, socio-demographics, appliances | Adjusted R2: 0.29 |
| [26] | Irish Commission for Energy Regulation's (CER) Smart Metering Electricity Customer Behaviour Trials | Ireland | 3488 | Daily | No (no gas meter data) | Income, employment status, presence of children, number of residents | Marginal R2: 0.20. Conditional R2: 0.81. |
| [15] | Convenience sample | UK | 173 | Daily | No (no gas meter data) | Building physical characteristics, socio-demographics, appliances, activity | Adjusted R2: 0.44 |
| [38] | National Energy Efficiency Data (NEED) | England and Wales | 11.3M | Annual | Yes | Property characteristics  Energy efficiency measures installed  Household characteristics  Local area characteristics | R2: 0.38 |

McLoughlin et al. [25] analysed half-hourly electricity smart meter data and linked survey data for a representative sample of approximately 4200 Irish households involved in a time-of-use tariff trial. Multiple linear regression was used to estimate the influence of survey data (which covered dwelling and occupant characteristics) on the dwelling-level variability of total electricity demand, maximum demand, load factor, and the time of maximum demand all averaged over six months.

Anderson et al. [26] analysed the same dataset and investigated the extent to which the data from the survey can explain variability in load profile ‘indicators’ such as 97.5% percentile load, lunchtime load, morning maximum, etc. Dependent variables were sampled for midweek days (Tuesday-Thursday) over a four-week period resulting in 12 observations for each variable per participant. A mixed effects framework was used including a random effects coefficient to quantify how much each household deviated from the average.

Kavousian et al. [31] examined structural and behavioural determinants of residential energy consumption for a convenience sample of 1628 Californian households (952 used in final analysis). Participants were all employees of a single Silicon Valley technology company. As such the sample was biased towards higher income, higher education, and higher interest in energy efficiency households. 10-minute resolution electricity data were collected over 238 days and survey data were collected covering weather, location, building physical characteristics, appliances, and occupant information. While high-resolution energy data were collected, and daily electricity consumption was used as the dependent variable in the regression analysis, this was averaged over the collection period (238 days).

Iwafune and Yagita [23] analysed high-resolution (30-60 min) energy data for 740 Japanese households collected over a period of one year (Dec 2013-Nov 2014), and performed a regression on monthly-averaged daily electricity consumption using data on weather, building and heating system information, household and appliance ownership and usage. The study used a convenience sample determined by the presence of specific home energy management systems. Separate regressions were conducted for the different seasons of the year. Unlike Anderson et al. [26] the authors of the study did not include a household-specific effect but instead used a time-specific effect for each month.

Satre-Meloy et al. [15] analysed high-resolution (1 second) electricity data measured over a period of 28 hours on a convenience sample of 173 GB households. Electricity data were averaged over the collection period and within day sub-periods. Satre-Meloy et al. used ‘de-minned’ electricity demand in their regression in addition to average demand. De-minning subtracts each household’s minimum demand from its average demand to remove its baseload and is particularly appropriate for studies aiming to characterise intra-day variations in demand that are affected by occupant activities.

Fan et al. [37] conducted a statistical analysis of drivers of peak demand by analysing half-hourly electricity consumption data collected over one year (2013) for 9900 households from the greater Sydney region linked with survey data for 3500 of these households covering housing type, demographics, appliance ownership, occupant living habits, and socioeconomic status. The study estimated individual peak demand over 12 selected peak demand periods in a year. A General Linear Model was used with 5-fold cross-validation. A mixed effects framework for analysing panel data was not used, unlike in Anderson et al. [26] or Iwafune and Yagita [23].

The review indicates that a linear mixed effect framework with random effects is appropriate when analysing panel data (cross-sectional plus time series data) and so will be used here for the analysis of daily household-level total energy consumption as we have repeated (daily) observations at the household level alongside cross-sectional socio-technical and contextual data. Analysis of panel data without using mixed effects would not be correct as the structure of the model would not account for the grouped nature of the data [39], effectively assuming that every observation is independent, even where they are from the same household.

Finally, we note the high variability of R2 for the studies above and, without performing a systematic analysis, make the general observations that higher R2 appears to be associated with studies with smaller sample sizes, lower data resolution, more contextual data, and that do not include heating or cooling.

# Method

This section describes the datasets, data preparation and analysis methods used to address the research questions.

## Datasets

This paper uses Edition 2 of the SERL observatory which contains data from almost 5,000 households who have consented for SERL to collect their smart meter data and to link to other datasets, including Energy Performance Certificate (EPC), Index of Multiple Deprivation 2019 (IMD) quintile and weather data, as well as an (optional) survey completed at sign up. The first participants were recruited during wave 1 in Autumn 2019 and the second tranche were recruited in wave 2 in August 2020 which broadened the sample to include the North of England and Scotland as described in [40,41]. The data used in this paper is drawn from the ~5,000 participants recruited in these two waves.

### SERL smart meter data

Half-hourly and daily[[2]](#footnote-3) electricity and gas readings are collected via the DCC gateway[[3]](#footnote-4) [42,43] for all participants with an accessible gas (76%) and/or electricity (100%) smart meter. The observations run from 19th August 2018 – 29th February 2020. The latest date of meter data was chosen to be sufficiently in advance of the start of the first COVID-19 lockdown period in GB (23rd March 2020) for the observations to be reasonably unaffected by the pandemic. The data documentation describes the quality of the SERL smart meter data in detail [10] and data quality processes were applied before conducting the analysis for this research, as described below.

### SERL survey

The SERL survey consists of 40 questions covering physical dwelling characteristics, household and respondent sociodemographic characteristics, energy use and heating behaviour, environmental attitudes, and appliance ownership. A copy of the survey is available in the documentation [10]. The aims of the survey were to collect contextual data to enable the production of nationally representative estimates, allow the creation and comparison of matched samples, and to help explain the variability of energy demand in the sample based on variables representing factors which existing research indicates are likely to influence household energy consumption (see literature review above), while also being reliably self-completed by a member of the public in about 10 minutes. Questions were designed in consultation with SERL consortium partners and Ipsos MORI and, where possible, were harmonised with national surveys such as the English Housing Survey, the 2011 Census and Understanding Society. Survey data are available for 4,753 (Edition 2) participants.

### Energy performance certificate (EPC)

Energy performance certificates (EPCs) are EU-mandated ratings of domestic building energy performance which aim to rate a building’s energy performance to enable comparisons of buildings energy use independent of occupant behaviour [44]. An EPC assessment is required by law when properties are sold or let in England and Wales. Address-level EPC data is publicly available [45], along with a description of variables (which include descriptions and energy efficiency estimates for building components, heating and lighting), and approximately half of the participants have an EPC[[4]](#footnote-5). At present, EPCs are not available for SERL participants in Scotland. While many dwelling-characteristic variables are available, it should be noted that there are measurement uncertainties associated with EPCs [46] e.g. due to surveyor error, or inaccuracies due to age of EPC and not reflecting subsequent retro-fitted measures. As we limit analysis to those households with an EPC in the final sample, this could be a source of bias as buildings which have not been sold or let since EPCs were introduced (in 2008) will not appear in the sample used here.

### Weather data (ECMWF ERA5 reanalysis)

Weather data linked to the SERL observatory households is sourced from the Copernicus ERA5 reanalysis of the ECMWF (European Centre for Medium-Range Weather Forecasts) global weather data [47]. This combines observations and modelled data to produce a global, complete, and consistent dataset. The data are available hourly on a grid with spatial resolution of approximately 28 sq. km. The SERL observatory provides over 20 variables relating to temperature, wind, irradiance, precipitation, and humidity conditions for participant grid cells. The ERA5 website gives full details of the data, and details of the data available through the SERL observatory are provided in [10]. The present analysis made use of two weather variables: air temperature 2m above the surface (°C) and global horizontal irradiance reaching the surface (MJ/m2 per day).

## Data preparation

To avoid the influence of coronavirus lockdowns, this analysis used smart meter data from 19th August 2018 to 29th February 2020. The number of households with smart meter data increases over this period due to the second recruitment wave in 2020 and the lack of historical data for some households (up to 12 months before sign-up depending on move in date and meter installation date[[5]](#footnote-6)). The following criteria were applied to the initial sample of approximately 5000 households, such that households were excluded if any of the following applied:

* Any missing SERL survey data (any question not completed).
* No EPC data available
* Gas and electricity data did not record most of the energy used in the home (any of the following):
  + Solar thermal hot water heating or solar PV reported in the survey or EPC data, or electricity export readings in the smart meter data (indicates presence of solar PV). This will bias the sample away from buildings that tend to have solar PV e.g. more recent, larger, more likely to be detached, as well as households that are more likely to have retro-fitted energy efficiency technologies [48].
  + Any form of central heating other than gas or electric (for example an oil boiler) reported in the survey. A consequence of this will be to bias the sample away from rural households where oil is more commonly used.
  + Gas heating reported in the survey or EPC but no gas smart meter data available.
  + Electric vehicle reported in the survey, since we are only concerned in this paper with energy use *within* the home. This will bias the sample away from the wealthy, middle-aged, male, well-educated, and affluent [49].
  + Buildings of multiple occupancy (not ‘self-contained’ in the survey) because the smart meter data relates to occupants not considered in the survey.
* Insufficient valid smart meter data available:
  + The smart meter documentation provided by Elam et al. [10] describes the conditions used to flag a read as valid (below a high threshold and in the correct units). In addition, we required valid read times (midnight for daily, on the hour/half-hour otherwise). Due to higher quality of half-hourly data overall, the sum of half-hourly readings was used where valid; otherwise daily reads were used.
  + The proportion of missing data for the household exceeded a maximum threshold. The missing data threshold was calculated separately for monthly average daily gas and electricity demand using data from 19 August 2018 to 29 February 2020. The missing data thresholds were 94% for electricity data and 72% for gas data.
* Ages of adult occupants not self-reported in the survey as this precludes the calculation of the average age of adult occupants.

This ensured that households in the analysis had equivalent levels of data and that the same sample could be analysed with different levels of contextual data, allowing the explanatory power of different levels of contextual data to be compared. Table 2 shows the number of households excluded at each stage. The large drop due to insufficient data can be attributed to participants having smart meters installed close to SERL recruitment and therefore not having smart meter data for the analysis period. This will not be an issue in future editions of the SERL data but we expect the level of survey non-response and EPC absence to remain roughly constant. The exclusion rates shown in Table 2 are therefore the worst case we anticipate.

Table 2. Sample size after the application of successive exclusion criteria

|  |  |  |
| --- | --- | --- |
| **Exclusion criteria** | **Households excluded in each step** | **Sample size remaining** |
| Initial sample of households with smart meter data |  | 4716 |
| Excluding dwellings with insufficient data and where not all energy use in the home was recorded by smart meters | 3063 (65%) | 1653 |
| Excluding dwellings without complete SERL survey | 372 (22%) | 1281 |
| Excluding dwellings without EPC data | 664 (52%) | 617 |

Both the SERL survey and EPC data contained categorical variables for which small categories of less than 10 were merged to avoid statistical disclosure. Where possible categories were merged with the most similar category, otherwise with the next smallest category.

Daily summaries were derived from hourly weather data for use in the regression models. To account for increased space heating in cold weather, hourly temperature data for each grid cell was transformed into heating degree days (*hdd*) using the method described by [50]. We used a UK standard base temperature of 15.5⁰C. The daily sum of the hourly solar radiation reaching a horizonal plane at the surface of Earth was also included in the models. This acts as a proxy for solar gains and day length. Future work will explore the use of different base temperatures and whether more sophisticated methods, perhaps making use of the hourly resolution of the weather data, could improve the models. The models contained continuous and categorical variables. The continuous variables were centred on the population mean to remove structural multicollinearity [51]. Similarly, the categorical variables were ‘one-hot’ encoded i.e. dummy encoded with the largest category used as the reference to reduce multicollinearity [52].

## Sample representativeness

The SERL Observatory sample was designed to be representative of households in GB with a DCC-enrolled smart meter (see [43] for further details), but response bias, the exclusion of the final recruitment wave and the application of the above exclusion criteria will result in a biased final analytic sample. Our results should not be taken as generalisable to the SERL Observatory as a whole, nor to the broader GB population. Future work will explore the use of larger and weighted samples to enable results that are more generalisable. Table 3 compares the regional distribution of the sample compared to the population of England and Wales, showing that in particular it under-represents the North of England (due to delayed smart meter rollout) and areas with greater deprivation (low IMD quintiles).

Table 3. Regional representation of the dwellings in the (N=617) sample used for analysis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Characteristic** | **Category** | **Sample number** | **Sample Percentage** | **Population Percentage (England and Wales)[[6]](#footnote-7)** |
| Region | SOUTH EAST | 120 | 19.4% | 15.0% |
| GREATER LONDON | 110 | 17.8% | 14.6% |
| SOUTH WEST | 78 | 12.6% | 10.0% |
| EAST OF ENGLAND | 73 | 11.8% | 10.2% |
| EAST MIDLANDS | 62 | 10.0% | 8.0% |
| WEST MIDLANDS | 60 | 9.7% | 9.6% |
| NORTH WEST | 46 | 7.5% | 12.7% |
| WALES | 34 | 5.5% | 6.0% |
| YORKSHIRE AND NORTH EAST | 34 | 5.5% | 13.9% |
| IMD quintile | 1 | 88 | 14.3% | 20.5% |
| 2 | 118 | 19.1% | 21.0% |
| 3 | 143 | 23.2% | 20.5% |
| 4 | 135 | 21.9% | 19.7% |
| 5 | 133 | 21.6% | 18.2% |

Table 4 compares some key characteristics of the sample with the population in England using data from the English Housing Survey 2018-2019 [53]. The sample under-represents flats and rental tenures, though is comparable to the national average in terms of size of dwelling and household and building energy efficiency rating (SAP).

Table 4. Key characteristics of the dwellings in the sample used for analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Characteristic** | **Category** | **Sample number** | **Sample Proportion** | **Sample mean** | **Population Proportion or mean (England)[[7]](#footnote-8)** |
| Built form | Detached | 155 | 25.1% | - | 26.1% |
| Semi-detached | 188 | 30.5% | - | 25.4% |
| Terraced | 182 | 29.5% | - | 28.4% |
| Purpose built flat | 75 | 12.2% | - | 16.5% |
| Converted house or commercial building | 17 | 2.8% | - | 3.6% |
| Tenure | Own / part-own | 515 | 83.5% | - | 63.3% |
| Private rental | 62 | 10.0% | - | 19.9% |
| Social rental / rent free | 40 | 6.5% | - | 16.8% |
| Household size | Number of persons per household | - | | 2.39 | 2.39 |
| Building energy efficiency | SAP rating | - | | 62.2 | 63.2 |
| Size of dwelling | Floor area (m2) | - | | 99.1 | 94 |

Table 5 shows key statistics regarding the energy consumption of the dwellings in the sample, and the degree days during the period of analysis. For comparison, in 2019 the mean UK daily domestic consumption was 31.56 kWh/day for gas and 10.22 kWh/day for electricity [54], mean gas use is 27% higher in the SERL sample and 7% lower in electricity. The higher gas use is consistent with larger properties (less flats) and wealthier occupants. Note our sample is drawn from England and Wales not UK. Note also that (as to be expected) the distributions of energy variables are highly skewed, hence the large relative standard deviations.

Table 5. Energy consumption and temperature statistics for the sample used for analysis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mean (SD) | 1st quartile | Median | 3rd quartile |
| Total daily household energy consumption (kWh) | 47.73  (42.74) | 14.80 | 36.48 | 69.14 |
| Daily gas consumption (kWh) | 39.91  (39.67) | 7.90 | 30.23 | 60.30 |
| Daily electricity consumption (kWh) | 9.50  (8.05) | 4.67 | 7.6 | 11.78 |
| Daily mean external temperature (⁰C) | 10.45 | 6.71 | 9.56 | 14.21 |
| Heating degree days (⁰C per day) | 5.63 | 1.86 | 5.95 | 8.79 |

## Statistical analysis

As noted in the method section, linear mixed effects models are an appropriate method for longitudinal panel data, as this allows the structure of the data (repeated observations for different individuals) to be explicitly accounted for in the model [55]. Coupled with this, mixed models are relatively straightforward to interpret and for these reasons this method was selected for this work. As one of the objectives of this paper is to assess the explanatory power of the SERL contextual datasets separately and in combination, variable subset selection is not implemented.

To investigate how the explanatory power of the model and results for individual coefficients change given different levels of contextual data, linear mixed models of daily total (gas + electricity) energy consumption per participant were fitted using four levels of contextual data. The first level of contextual data is ‘basic data’ consisting of: region, IMD quintile, day of the week, bank holiday indicator, heating degree days and solar radiation. These are widely available, area-based variables that are easily linked to smart meter data and available for all participants. Additional models were developed with further levels of added contextual data: from the SERL survey and EPC data separately and then in combination.

We note that log-transforming of the dependent variable is sometimes performed in previous studies to address heteroscedasticity [18] or symmetry in residuals [8]. We do not log-transform as the residuals of the model are normally distributed and not strongly affected by heteroscedasticity, and log-transforming has the adverse effect of producing residuals which are not normally distributed.

To take advantage of the longitudinal (repeated measures) nature of the dataset we applied a random effects (RE) approach, similar to that used by Anderson et al. [26]. We use both random intercepts and random slopes applied to the heating degree day (*hdd*) variable; this allows each individual dwelling to deviate from the group mean intercept and gradient. The thermal performance of each dwelling will strongly affect the gradient of the *hdd* variable and the random slope component allows this to deviate from the mean for each participant. Every variable is included by itself as well as interacting with the *hdd* slope variable. Following Snijders and Bosker [39], the random slope model with all contextual data therefore takes the form:

Equation 1

*Yt,i* is the energy consumption of dwelling *i* at time period *t*. The first part of the equation with coefficients is the *fixed part* (because the coefficients are fixed i.e. non-stochastic), while the remainder is the *random part* of the model, comprising ‘level two’ residuals (i.e. participant-level) random intercept and random *hdd* slope for each participant *i*, and ‘level one’ residual (i.e. measurement-level) error . It is assumed that level one and two residuals have mean 0, and that the pair of level two residuals, and the level one residual, are independent and identically distributed.

The fixed part includes the intercept for the average participant ; regression coefficients associated with *P* measurement-level variables i.e. those that change for each participant *i* and each time step *t* (heating degree days and solar radiation); and regression coefficients associated with *Q* participant-level variables , consisting of all other variables, all of which are also interacted with the heating degree day variable *hdd*. is the *hdd* slope for the average participant. A full list of the variables from each dataset used in the regressions is given in Supplementary Data.

EPC variables related to cost, carbon dioxide emissions and environmental efficiency were excluded. Text descriptions of building elements such as type of external wall were also excluded as a categorisation of the element’s thermal performance was included instead[[8]](#footnote-9). All SERL survey variables were included except for those relating to the respondent (‘About you’ section) as the unit of analysis of interest here is the household, not the respondent. An interaction term between solar radiation and heating degree days was included in all models since solar gains can provide space heating during the heating season.

As in [26], the explanatory power of models is evaluated using marginal and conditional R2. The marginal R2 describes the variance attributable to fixed effects (approximately equivalent to the conventional OLS coefficient of determination R2), while the conditional R2 describes the variance attributable to both the fixed and random effects (i.e it includes the marginal R2). The residual R2 describes the remaining unexplained variance. See Nakagawa and Sheilzeth [56] for more information regarding these indicators for linear mixed effect models using random factors (intercepts) only, and Johnson [57] for a discussion of extending this to models with a random slope component. Adjusted R2 are not calculated here as while there are many explanatory variables (*k*=359 in model with all data), this is very small compared to the number of observations (*n*=163,107), meaning the adjustment factor used in its calculation [58] and the differences between models with different numbers of explanatory variables are negligible.

We use *p*-values to evaluate statistical significance of independent regression variables and take *p*<0.05 as our statistical significance threshold noting that statistical significance does not necessarily imply substantive significance. We note that as we do not use variable selection and we include many explanatory variables in our model it is probable that this means that some of the variables found to be significant will be spurious results. We also note the population for which the sample is intended to be representative of is not the SERL Observatory nor GB population. It is a biased sample affected by sample design, recruitment strategy, non-response bias, and exclusion criteria (see above for descriptive statistics of the sample that indicate biases).

All analysis was performed within the UCL Data Safe Haven using Python (version 3.8), spyder (version 4.1.4), pandas (version 1.0.5) [59,60], and statsmodels (version 0.11.1) [61] for the linear mixed effects regressions.

## Multicollinearity

Many explanatory variables were included in this analysis and there is multi-collinearity in the original regressors (prior to transformation). An obvious example is that both the SERL survey and the EPC data include categorical variables relating to the age of the building. However, many other variables are also affected by collinearity, a common issue in energy demand research [8].

The effect of multicollinearity is to reduce the accuracy of the estimates of the regression coefficients [28] and thereby reducing the probability of correctly detecting a significant coefficient, and to make the coefficient of collinear variables sensitive to changes in the input data. This makes interpretation of the model’s results challenging, though multicollinearity does not affect the model’s goodness of fit (R2) or (within-sample) predictive accuracy.

Multicollinearity is commonly assessed using the variance inflation factor (VIF). This showed high levels of multi-collinearity for our data (98 variables with VIF > 10).

Correcting multicollinearity for categorical variables can involve removing, combining, or transforming variables. Removing variables is unwelcome as we want to assess the explanatory power of the dataset in total. To reduce multicollinearity, we therefore use combination and transformation. Continuous variables were centered (the population mean was subtracted); categorical variables dummy encoded so that the reference category was the largest, and some categorical variables were combined (total number of rooms; central heating fuel type). Centered variables are denoted with the suffix *\_c­*. The exception to this is that heating degree day *hdd* was not centered for the basic data model as VIF were reasonable i.e. multicollinearity did not need to be addressed and the model is more intuitive where the intercept is when *hdd* equals zero.

This considerably reduced variance inflation factors although some remain high, with an average of 6.8 and 38 with VIF>10 for the final model. Average VIF for the other models were: 5.6 for basic data only, 8.0 for basic plus EPC data, and 2.8 for basic plus SERL survey. VIF for all the models are reported in Supplementary Data.

The coefficients of variables affected by collinearity (commonly taken to be where VIF is greater than 5 or 10) should be interpreted with extreme caution as they may be unstable (if a different data set is used) and have inflated p-values (i.e. reduced significance) due to systematic bias in the underlying standard errors.

# Results

## Comparison of models with different levels of data

Table 6 compares the overall performance of the models with different levels of contextual data. In terms of overall explanatory power, the model with all data explained the most, and produced a marginal R2 of 0.68, conditional R2 of 0.83 and residual R2 0.17, with a root mean square error of 17.7kWh/day (Table 6). The marginal R2 for the model with basic data only is 0.36. Adding EPC data to the basic data increases the marginal R2 to 0.63, while adding SERL survey data to the basic data increases the marginal R2 to 0.65, the effect of including both SERL survey and EPC data only increases the R2 to 0.68.

Table . Statistical performance for linear mixed effect models using different levels of contextual data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Basic data only | Basic + EPC data | Basic + SERL data | All data |
| Marginal R2 | 0.36 | 0.63 | 0.65 | 0.68 |
| Conditional R2 | 0.83 | 0.83 | 0.83 | 0.83 |
| Residual R2 | 0.17 | 0.17 | 0.17 | 0.17 |
| Root mean square error (RMSE) (kWh/day) | 17.7 | 17.7 | 17.7 | 17.7 |

The conditional R2, residual R2, and RMSE is the same for all models (0.83, 0.17 and 17.7 kWh/day). This is because a best-fit intercept and *hdd* slope is fitted for all individual dwellings, regardless of the amount of contextual data provided. Although the conditional R2 does not significantly change, the variance of the random slope parameter decreases from 12.3 kWh/day/°C for the basic daily model to 4.3 kWh/day/°C for the model with all contextual data. This suggests that the model combining all contextual data explains much of the variability in individual temperature response of the dwellings in the sample (evidenced by the larger marginal R2).

## Model with all data

Table 7 shows the estimates of the coefficients for the individual variables included in the model with all data. Variables are ordered by origin dataset and only shown if they were statistically significant at 90% level or above. We take a significance level of 95% to indicate a significant variable as it is conventional for academic publishing (albeit less conservative than the choice of 99%) and only discuss variables in the text where this significance level is reached. A full set of results are provided in the Supplementary Data.

Table 7. Regression results using linear mixed effects model with random intercept and hdd slope using all data (basic, EPC, and SERL survey). Dependent variable total energy consumption per day (kWh/day). Variables that are significant at 90% level are shown, those at 95% level shown in bold, ordered by origin dataset. Number of observations: 163107. Number of groups (households): 617 . Min. group size: 53. Max. group size: 560. Mean group size: 264.4.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | Variable | Coefficient | Standard error | z-score | p-value | 95% CI (lower) | 95% CI (upper) |
|  | **Intercept** | **48.750** | 6.338 | 7.692 | 0.000 | 36.328 | 61.173 |
| Basic | C(bank\_holiday, Treatment(reference=0))[T.1] | -0.921 | 0.500 | -1.842 | 0.065 | -1.900 | 0.059 |
| Basic | C(dayofweek, Treatment(reference=0))[T.1] | 0.323 | 0.166 | 1.950 | 0.051 | -0.002 | 0.647 |
| Basic | **C(dayofweek, Treatment(reference=0))[T.2]** | **-0.487** | 0.167 | -2.924 | 0.003 | -0.814 | -0.161 |
| Basic | **C(dayofweek, Treatment(reference=0))[T.3]** | **0.778** | 0.165 | 4.705 | 0.000 | 0.454 | 1.103 |
| Basic | **C(dayofweek, Treatment(reference=0))[T.4]** | **1.153** | 0.165 | 6.986 | 0.000 | 0.830 | 1.477 |
| Basic | **C(dayofweek, Treatment(reference=0))[T.5]** | **2.964** | 0.165 | 17.948 | 0.000 | 2.640 | 3.288 |
| Basic | **C(dayofweek, Treatment(reference=0))[T.6]** | **2.430** | 0.166 | 14.635 | 0.000 | 2.104 | 2.755 |
| Basic | **C(IMD\_quintile, Treatment(reference=3))[T.4]** | **5.306** | 2.148 | 2.471 | 0.013 | 1.097 | 9.515 |
| Basic | C(IMD\_quintile, Treatment(reference=3))[T.5] | 3.738 | 2.250 | 1.661 | 0.097 | -0.672 | 8.148 |
| Basic | **hdd\_c** | **5.201** | 0.876 | 5.937 | 0.000 | 3.484 | 6.919 |
| Basic | **hdd\_c:C(bank\_holiday, Treatment(reference=0))[T.1]** | **-1.488** | 0.124 | -11.986 | 0.000 | -1.732 | -1.245 |
| Basic | **hdd\_c:C(dayofweek, Treatment(reference=0))[T.1]** | **-0.085** | 0.042 | -2.030 | 0.042 | -0.167 | -0.003 |
| Basic | **hdd\_c:C(dayofweek, Treatment(reference=0))[T.3]** | **-0.231** | 0.043 | -5.375 | 0.000 | -0.315 | -0.147 |
| Basic | **hdd\_c:C(dayofweek, Treatment(reference=0))[T.6]** | **0.107** | 0.044 | 2.443 | 0.015 | 0.021 | 0.193 |
| Basic | **hdd\_c:C(IMD\_quintile, Treatment(reference=3))[T.1]** | **0.734** | 0.345 | 2.127 | 0.033 | 0.058 | 1.411 |
| Basic | **hdd\_c:C(IMD\_quintile, Treatment(reference=3))[T.4]** | **0.821** | 0.297 | 2.765 | 0.006 | 0.239 | 1.403 |
| Basic | **hdd\_c:C(IMD\_quintile, Treatment(reference=3))[T.5]** | **0.759** | 0.311 | 2.442 | 0.015 | 0.150 | 1.368 |
| Basic | **hdd\_c:C(Region, Treatment(reference="SOUTH EAST"))[T.NORTH WEST]** | **-0.969** | 0.469 | -2.063 | 0.039 | -1.889 | -0.049 |
| Basic | hdd\_c:C(Region, Treatment(reference="SOUTH EAST"))[T.WALES] | -0.881 | 0.509 | -1.731 | 0.083 | -1.878 | 0.116 |
| Basic | **hdd\_c:solar\_radiation\_c** | **-1.012** | 0.025 | -40.8 | 0.000 | -1.061 | -0.963 |
| Basic | **solar\_radiation\_c** | **-12.564** | 0.101 | -124.248 | 0.000 | -12.762 | -12.366 |
| EPC | **C(hotWaterEnergyEff, Treatment(reference="Good"))[T.Very Good]** | **11.317** | 5.750 | 1.968 | 0.049 | 0.047 | 22.587 |
| EPC | **C(hotWaterEnergyEff, Treatment(reference="Good"))[T.Very Poor]** | **-12.730** | 5.979 | -2.129 | 0.033 | -24.449 | -1.012 |
| EPC | **C(wallsEnergyEff, Treatment(reference="Good or very good"))[T.Very Poor]** | **-6.274** | 3.006 | -2.087 | 0.037 | -12.165 | -0.383 |
| EPC | **C(windowsEnergyEff, Treatment(reference="Average"))[T.Poor]** | **6.036** | 2.959 | 2.040 | 0.041 | 0.237 | 11.835 |
| EPC | **hdd\_c:C(hotWaterEnergyEff, Treatment(reference="Good"))[T.Very Good]** | **1.57** | 0.793 | 1.98 | 0.048 | 0.016 | 3.123 |
| EPC | **hdd\_c:C(hotWaterEnergyEff, Treatment(reference="Good"))[T.Very Poor]** | **-2.088** | 0.829 | -2.519 | 0.012 | -3.713 | -0.463 |
| EPC | hdd\_c:C(tenure, Treatment(reference="owner-occupied"))[T.rental (social) / Other] | -1.006 | 0.612 | -1.644 | 0.1 | -2.205 | 0.193 |
| EPC | hdd\_c:C(wallsEnergyEff, Treatment(reference="Good or very good"))[T.Very Poor] | -0.774 | 0.416 | -1.863 | 0.063 | -1.589 | 0.041 |
| EPC | hdd\_c:C(windowsEnergyEff, Treatment(reference="Average"))[T.Poor] | 0.702 | 0.409 | 1.716 | 0.086 | -0.1 | 1.504 |
| EPC | **hdd\_c:negativeCurrentEnergyEfficiency\_c** | **0.07** | 0.03 | 2.303 | 0.021 | 0.01 | 0.129 |
| EPC | **hdd\_c:numberOpenFireplaces\_c** | **-0.675** | 0.203 | -3.326 | 0.001 | -1.072 | -0.277 |
| EPC | **hdd\_c:totalFloorArea\_c** | **0.031** | 0.005 | 6.568 | 0.000 | 0.022 | 0.041 |
| EPC | negativeCurrentEnergyEfficiency\_c | 0.375 | 0.218 | 1.717 | 0.086 | -0.053 | 0.803 |
| EPC | **numberOpenFireplaces\_c** | **-4.76** | 1.466 | -3.246 | 0.001 | -7.634 | -1.886 |
| EPC | **totalFloorArea\_c** | **0.208** | 0.035 | 6.01 | 0.000 | 0.14 | 0.276 |
| EPC/SERL survey | **total\_rooms\_avg\_c** | **2.946** | 1.305 | 2.257 | 0.024 | 0.388 | 5.505 |
| SERL survey | **average\_adult\_age\_c** | **0.16** | 0.07 | 2.275 | 0.023 | 0.022 | 0.297 |
| SERL survey | **C(A14, Treatment(reference="Some effort"))[T.A great deal of effort]** | **-6.453** | 2.032 | -3.176 | 0.001 | -10.435 | -2.471 |
| SERL survey | C(A15\_cold, Treatment(reference="Not very often"))[T.Always] | 6.032 | 3.569 | 1.690 | 0.091 | -0.964 | 13.027 |
| SERL survey | C(A15\_warm, Treatment(reference="Always"))[T.Quite often] | -5.968 | 3.119 | -1.914 | 0.056 | -12.081 | 0.145 |
| SERL survey | **C(A2, Treatment(reference="Not at all"))[T.A little]** | **-4.564** | 1.736 | -2.629 | 0.009 | -7.967 | -1.161 |
| SERL survey | C(A4\_timer, Treatment(reference=1))[T.0.0] | 4.838 | 2.672 | 1.810 | 0.070 | -0.400 | 10.075 |
| SERL survey | **C(A5\_setpoint, Treatment(reference=">20.0 & <=22.0"))[T.<=19.0]** | **-6.330** | 2.078 | -3.046 | 0.002 | -10.403 | -2.257 |
| SERL survey | C(A5\_setpoint, Treatment(reference=">20.0 & <=22.0"))[T.>19.0 & <=20.0] | -3.745 | 2.016 | -1.858 | 0.063 | -7.695 | 0.206 |
| SERL survey | C(A5\_setpoint, Treatment(reference=">20.0 & <=22.0"))[T.>22.0] | 4.750 | 2.834 | 1.676 | 0.094 | -0.804 | 10.304 |
| SERL survey | **C(A8\_use\_own, Treatment(reference="None or don't know"))[T.Varies]** | **-10.675** | 4.786 | -2.231 | 0.026 | -20.055 | -1.296 |
| SERL survey | C(B10\_acu, Treatment(reference=0))[T.1] | 7.660 | 4.540 | 1.687 | 0.092 | -1.238 | 16.559 |
| SERL survey | C(B10\_laptop, Treatment(reference=1))[T.0] | -4.864 | 2.856 | -1.703 | 0.089 | -10.462 | 0.734 |
| SERL survey | **C(B9, Treatment(reference="1950 to 1975"))[T.1900 to 1929]** | **7.854** | 3.586 | 2.190 | 0.029 | 0.826 | 14.883 |
| SERL survey | **C(B9, Treatment(reference="1950 to 1975"))[T.1976 to 1990]** | **-6.094** | 2.876 | -2.119 | 0.034 | -11.731 | -0.458 |
| SERL survey | **education\_c** | **-3.859** | 1.876 | -2.057 | 0.04 | -7.537 | -0.181 |
| SERL survey | **hdd\_c:C(A13\_often\_clothes, Treatment(reference="Very often"))[T.Never, or not applicable]** | **1.930** | 0.839 | 2.300 | 0.021 | 0.285 | 3.575 |
| SERL survey | **hdd\_c:C(A14, Treatment(reference="Some effort"))[T.A great deal of effort]** | **-0.896** | 0.281 | -3.191 | 0.001 | -1.447 | -0.346 |
| SERL survey | **hdd\_c:C(A2, Treatment(reference="Not at all"))[T.A little]** | **-0.611** | 0.240 | -2.543 | 0.011 | -1.081 | -0.140 |
| SERL survey | hdd\_c:C(A4\_smart, Treatment(reference=0))[T.1.0] | -0.534 | 0.286 | -1.866 | 0.062 | -1.095 | 0.027 |
| SERL survey | hdd\_c:C(A5\_setpoint, Treatment(reference=">20.0 & <=22.0"))[T.<=19.0] | -0.475 | 0.287 | -1.655 | 0.098 | -1.038 | 0.088 |
| SERL survey | hdd\_c:C(B10\_combined\_wash, Treatment(reference=0))[T.1] | -1.238 | 0.751 | -1.648 | 0.099 | -2.71 | 0.234 |
| SERL survey | hdd\_c:C(B10\_dryer, Treatment(reference=0))[T.1] | -0.426 | 0.236 | -1.807 | 0.071 | -0.889 | 0.036 |
| SERL survey | hdd\_c:C(B10\_freezer, Treatment(reference=0))[T.1] | 0.425 | 0.225 | 1.89 | 0.059 | -0.016 | 0.865 |
| SERL survey | **hdd\_c:C(B9, Treatment(reference="1950 to 1975"))[T.1900 to 1929]** | **1.26** | 0.496 | 2.54 | 0.011 | 0.288 | 2.233 |
| SERL survey | hdd\_c:C(B9, Treatment(reference="1950 to 1975"))[T.1976 to 1990] | -0.698 | 0.398 | -1.755 | 0.079 | -1.478 | 0.082 |
| SERL survey | hdd\_c:C(B9, Treatment(reference="1950 to 1975"))[T.1991 to 2002] | -0.999 | 0.59 | -1.693 | 0.091 | -2.156 | 0.158 |
| SERL survey | hdd\_c:occupants\_0\_15\_c | 0.295 | 0.163 | 1.812 | 0.07 | -0.024 | 0.613 |
| SERL survey | **hdd\_c:occupants\_adults\_c** | **0.336** | 0.141 | 2.375 | 0.018 | 0.059 | 0.613 |
| SERL survey | **occupants\_0\_15\_c** | **4.016** | 1.175 | 3.417 | 0.001 | 1.712 | 6.32 |
| SERL survey | **occupants\_adults\_c** | **5.519** | 1.023 | 5.393 | 0.000 | 3.513 | 7.525 |

### Basic data

The intercept for the average reference-category participant was 48.75 kWh/day (*p*=0.000), equivalent to an average power consumption of 2.031 kW. This corresponds to the average daily consumption associated with the reference categories for all categorical variables (i.e. semi-detached house in South East England, built 1950-1975, etc.) and mean value for all continuous variables (i.e. 5.6 heating degree days, average number of occupants etc.). The slope of the heating degree day variable for the average reference category participant was 5.201 kWh/day/°C per day (*p*=0.000). This corresponds to an increase in average power consumption of 216.7 W for each additional heating degree day. The intercept at zero *hdd\_c* can be estimated from these values as 19.47 kWh/day or an average power consumption of 811.2 W.

Region (*Region*) had a significant effect on *hdd\_c* slope. Compared to the reference category ‘South East’ of England, the ‘North West’ was associated with a shallower *hdd\_c* slope -0.969 kWh/day/°C per day (*p*=0.039).

IMD quintile (*IMD\_quintile*) had a significant effect on intercept and *hdd\_c* slope. Compared to the reference category 3, quintile 4 (less deprivation) was associated with larger intercept +5.306 kWh/day (*p*=0.013), and quintiles 1, 4 and 5 were associated with steeper *hdd\_c* slopes +0.734 kWh/day/°C per day (*p*=0.033), +0.821 kWh/day/°C per day (*p*=0.006), +0.759 kWh/day/°C per day (*p*=0.015) respectively.

Day of the week (*dayofweek*) had a significant effect on intercept and *hdd\_c* slope. Compared to the reference category Mondays (coded 0), other weekdays had larger intercepts (*p*<=0.003) except Tuesdays, while Tuesdays and Thursdays had shallower *hdd\_c* slopes (*p*<=0.042). Weekends had larger intercepts (*p*=0.000) and Sunday had steeper *hdd\_c* slope (*p*=0.015) than Mondays. Bank holidays had shallower *hdd\_c* slopes (*p*=0.000) than non-bank holidays.

Increase in solar radiation from the mean (*solar\_radiation\_c*) was associated with a negative effect on both intercept (-12.6 kWh/day per MJ/m2/day, *p*=0.000) and *hdd\_c* slope (-1.01 kWh/day/°C per day per MJ/m2/day, *p*=0.000).

### Building physical characteristics

The following building physical characteristics had *p*-values <0.05:

* **Floor area:** Increase in total floor area from the mean (*total\_floor\_area\_c*) was associated with larger intercept (+0.21 kWh/day/m2, *p=*0.000) and steeper *hdd\_c* slope (+0.031 kWh/day/m2/°C per day, *p*=0.000). Note: this variable has high VIFs (6.7, 6.8).
* **Rooms:** Increases in total number of rooms from the mean (*total\_rooms\_avg\_c*) was associated with larger intercept (+2.95 kWh/day per room, *p*=0.024). Note: this variable has high VIFs (14.8, 15.1).
* **Building age:** While none of the EPC construction age band (*C(constructionAgeBand)*) categories were significant, the building age question in the SERL survey (*C(B9)*) was significant. Compared to the reference category ‘1950 to 1975’ buildings constructed ‘1900 to 1929’ had larger intercept (+7.85 kWh/day, *p*=0.029) and steeper *hdd\_c* slope (+1.26 kWh/day/°C per day, *p*=0.011), buildings constructed ‘1976 to 1990’ had lower intercepts (-6.09 kWh/day, *p*=0.034). Note: these variables had high VIFs (e.g. 5.4, 5.6).
* **Energy efficiency**:
  + **Hot water:** Buildings with ‘very good’ hot water energy efficiency (*C(hotWaterEnergyEff)*) were associated with larger intercept (+11.3 kWh/day, *p*=0.049) and steeper *hdd\_c* slope (+1.57 kWh/day/°C per day, *p*=0.048) than buildings with ‘good’ hot water efficiency. Those with ‘very poor’ hot water efficiency were associated with smaller intercept (-12.73 kWh/day, *p*=0.033) and shallower *hdd\_c* slope (-2.09 kWh/day/°C per day, *p*=0.012) than those with ‘good’ efficiency.
  + **External walls:** Buildings with ‘very poor’ wall energy efficiency (*C(wallEnergyEff)*) were associated with lower intercept (-6.27 kWh/day, *p*=0.037) than buildings with ‘good or very good’ wall energy efficiency. Note: this variable has high VIF (6.8).
  + **Windows:** Buildings with ‘poor’ windows energy efficiency (*C(windowsEnergyEff)*) were associated with higher intercept (+6.04 kWh/day, *p*=0.041) than buildings with ‘average’ windows energy efficiency.
  + **Overall efficiency rating (Building SAP Energy Efficiency Rating *SAPEER*) :** Difference in building inefficiency from the mean (*negativeCurrentEnergyEfficiency\_c = (100- SAPEER)*) was associated with a positive effect on *hdd\_c* slope (+0.07 kWh/day/°C per day, *p*=0.021). Note: this variable has high VIF (14.6).
* **Standalone heat sources:** Presence standalone non-gas and non-electricity heat sources (e.g. fuelled by wood, coal, etc.) and where the use of these during cold weather ‘varies’ (*C(A8\_use\_own)*) was associated with a lower intercept (-10.7 kWh/day, *p*=0.026) compared to households without such heaters.
* **Open fireplaces:** Difference in number of open fireplaces from the mean (*numberOpenFireplaces\_c*) was associated with a reduction in both intercept (-4.76 kWh/day per open fireplace, *p*=0.001) and *hdd\_c* slope (-0.68 kWh/day per open fireplace/°C per day, *p*=0.001).

### Sociodemographic characteristics

The following sociodemographic characteristics were associated with increased consumption and found to have *p*-values <0.05:

* A higher **number of adults** (>=16 years old) occupants (*occupants\_adult\_c*) was associated with positive effect on both intercept (+5.52 kWh/day per adult, *p*=0.000) and *hdd\_c* slope (+0.34 kWh/day per adult/°C per day, *p*=0.018).
* A higher total **number of children** (<16 years old) occupants (*occupants\_0\_15\_c*) was associated with positive effect on intercept (+4.02 kWh/day per child, *p*=0.001).
* A higher average **age of adult occupants** (*average\_age\_adult\_c*) was associated with a positive effect on the intercept (+0.16 kWh/day per year, *p*=0.023).
* A lower proportion of **adults with a qualification** (*education\_c*) was associated with a positive effect on the intercept (+3.86 kWh/day, *p*=0.04).

### Behavioural factors

The following behavioural factors were found to have *p*-values <0.05:

* **Temperature set-point:** Households with lower heating temperature set-points (*C(A5\_setpoint)*) than households with the reference category ‘>20C & <=22.0C’ are associated with lower intercepts (-6.33 kWh/day for ‘<19C’ households *p*=0.002).
* **Use of clothes vs heating:** Households who ‘never’ put more clothes on when feeling cold rather than putting the heating on or turning it up (*C(A13\_often\_clothes)*), or where this question was not applicable, were associated with a steeper *hdd\_c* slope (+1.9 kWh/day/°C per day, *p*=0.021) compared to households who did this ‘very often’.
* **Smart meter:** Question A2 asked whether having a smart meter had affected the way energy is used in the home (*C(A2)*) was associated with a significant effect on consumption. Compared to the reference category ‘Not at all’, households who responded ‘a little’ were associated with smaller intercept (-4.7 kWh/day, *p*=0.009) and shallower *hdd\_c* slope (-0.6 kWh/day/°C per day, *p*=0.011).
* **Effort to reduce consumption:** Households who make ‘a great deal of effort’ to limit or reduce their energy consumption (*C(A14)*) were associated with a smaller intercept (-6.5 kWh/day, p=0.001) and shallower *hdd\_c* slope (-0.9 kWh/day/°C per day, *p*=0.001) compared to households who make ‘some effort’.

The estimates of the coefficients for the individual variables included in the models with fewer data are provided in the Supplementary Data but are not described here. Variables are ordered alphabetically and only shown if they were significant at 90% level or above. Those significant at 95% level are highlighted in bold.

# Discussion

## Comparison with previous studies

Returning to our first research question, we find that approximately two thirds of the variation in daily household total energy consumption is explained by the SERL Observatory variables. Overall, our results compare favourably with those found in the literature, with better goodness-of-fit indicators (higher marginal R2) and smaller errors relative to the standard deviation of the dependent variable. Anderson et al. [62] found a marginal R2 of 0.20 and conditional R2 of 0.81 for daily mean electricity demand, while we find marginal R2 of 0.68 and conditional R2 of 0.83 for the model with all data. Our marginal R2 is considerably higher, however we note that Anderson et al. used only a small number of contextual variables (number of residents, income band, number of children, and employment status) compared to our model. The conditional R2 is similar, indicating that a similar effect on explanatory power associated with including random effects in the model. Other studies in the literature (see Table 1) report adjusted R2 of 0.29-0.44 for daily demand, though are not directly comparable to the marginal R2 reported here as they are calculated differently.

While we report errors for the models, we note that none of the cited studies report comparable errors with the exception of [15,18]. Direct comparison with these is complicated because they log-transform the dependent variable and report prediction errors while we focus on inference instead. Nonetheless, [18] reports errors that are approximately 20% smaller than the standard deviation of the dependent variable. The standard deviation of our dependent variable is 42.74 kWh/day (89% of the mean). Our model has a RMSE error (17.7 kWh/day) which is 59% smaller than the standard deviation. We note, however, that prediction errors for the models should be expected to be considerably higher.

We now return to our second research question and consider what individual variables observed in the SERL Observatory data explain household-level daily energy consumption, grouping these into the following categories: building physical characteristics, appliances, sociodemographic characteristics and behavioural factors.

## Building physical characteristics

Our results agree with a number of existing findings regarding the association between building physical characteristics and energy consumption [14,38]: buildings with larger floor area, have more rooms, are older, and that experience colder or less sunny weather are associated with increased energy consumption.

Buildings with lower overall energy efficiency ratings were associated with increased energy consumption, as were those with poor window energy efficiency. However, counterintuitive results were found for walls and hot water efficiency: with ‘very poor’ efficiency associated with *reduced* demand. We note that spurious results are to be expected with linear regression containing large numbers of co-variates (as here) and the presence of multicollinearity can be expected to introduce instability to estimates. Future analysis will seek to reduce the impact of both these issues to see if these counter-intuitive results are reproduced.

Contrary to several previous studies [14,38], we do not find a significant association between level of building detachedness, or number of bedrooms and energy consumption. Possible reasons are high VIF for detachedness variables, and that in previous studies number of bedrooms may have acted as a proxy for characteristics that we explicitly capture such as the number of occupants.

Finally, the results indicate the effect on metered energy consumption of heat sources such as open fireplaces and solid fuel stoves even where, as here, these are in addition to gas heating. Presence and use of these were associated with a large and statistically significant reduction in (metered) energy consumption, although clearly this does not reflect differences in overall gross energy consumption which may indeed be higher for these dwellings.

## Appliances

Unlike previous studies [14], we do not find any significant association between presence of appliances and demand, though we note several were significant at 90% level. A possible explanation is that while previous studies have tended to focus on *electricity* consumption only, which is more likely to be affected by appliances, we are analysing *total* energy consumption, which is dominated by space and water heating, and as such unlikely to be significantly affected by electrical appliances such as laptops, dishwashers etc. Furthermore, some of these variables showed high VIFs and low sample sizes for treatment groups (e.g. only 17 households reported presence of air-conditioning unit). We recommend a further study to test these variables by investigating electricity consumption independently, as well as increasing the sample size.

## Sociodemographic characteristics

Our results confirm a number of existing findings regarding the effect of sociodemographic characteristics on energy consumption [14,38]: households with more adult occupants, more children, and with older adult occupants, are associated with increased energy consumption.

Previous studies report mixed results for the effect of tenure, and education on energy consumption. We find no significant effect associated with tenure, and found that education (a higher proportion of adults with qualifications) was associated with reduced demand.

## Behavioural factors

Behavioural factors can include energy conservation behaviour in the form of ‘purchasing’ activities or ‘habitual’ actions and are less well studied than the previous categories of factors [18]. Nonetheless some previous studies report an association between habitual energy saving behaviours and reduced consumption [18,63,64]. We found that households that set lower heating temperature set-points consumed less than those that set higher set-points. Households who ‘never’ put more clothes on when feeling cold rather than putting the heating on or turning it up were associated with increased demand on colder days than those who did this ‘very often’. Households who made ‘a great deal of effort’ to limit or reduce their energy consumption were associated with lower consumption than those who made ‘some effort’. Households who said that having a smart meter had affected the way they use energy ‘a little’ were associated with lower demand compared to those who said ‘not at all’. Finally, we found that households that ‘always’ open windows during typically cold or warm weather were associated with increased demand compared to households that did this less.

Overall, if we consider the size of the coefficients for the factors discussed (Figure 1) we see that ‘very poor’ hot water energy efficiency, solar radiation, use of non-mains powered standalone heaters and bank holidays have the largest relative effects associated with reductions in energy use, while (old) building age, ‘poor’ window energy efficiency, ‘very good’ hot water energy efficiency, number of occupants, and ‘never’ putting on more clothes rather than turning on the heating have the largest relative effects associated with increases in energy use. The large confidence intervals indicate the very high uncertainty in point estimates and we note the presence of counter-intuitive results associated with hot water efficiency which indicates the possibility of spurious results which should be further investigated using appropriate methods (see future work below).

Table

Description automatically generated

Figure . Size of coefficients for the statistically significant variables (p<0.05) showing those with largest relative effect on intercept (upper) and hdd slope (lower). Those with negative effect are shown in red, those with positive effect shown in blue. Error bars show 95% confidence interval of the estimate.

## SERL Observatory: a new national data resource for energy demand research

Returning to the final research question, we have shown that the EPC data and SERL survey data, when included alongside the basic data, are similar in terms of explanatory power. Moreover, the SERL survey data is much less affected by multicollinearity and has higher data availability for the SERL Observatory. Future researchers using similar techniques may wish to opt for a balance of maximising sample size and explanatory power by not requiring complete EPC data for their analytic sample. We believe these results demonstrate the value of SERL survey as a tool for collecting useful contextual data with relatively low participant burden, and note the complementarity of the SERL survey with EPC data which is nonetheless widely available for UK dwellings.

Overall, our results demonstrate that a large amount of (within-sample) variation can be explained by data collected within the SERL Observatory. The results largely support existing theory and add to the empirical evidence base that building physical characteristics, household sociodemographic information, and household behavioural factors all explain aspects of demand, across a wide range of contexts. Considering the complexity of the subject under investigation (daily residential energy consumption), the simplicity of the approach to data selection used here, and the relatively low burden on participants for data collection, we believe this is a promising result that demonstrates the value of the SERL Observatory dataset as a data resource for improving the understanding of energy demand in residential buildings. The final (third) wave of SERL participant recruitment was completed in March 2021 and over 8,000 further participants were recruited, bringing the total participant number to over 13,000. We therefore encourage future energy demand projects involving surveys to harmonise with the SERL survey to support greater interpretation, reproducibility and cross-validation between research findings [65].

## Future work

This paper presents a first initial step in a larger programme of research by multiple organisations using SERL Observatory data. We have started with simple but limited analysis; for example, using a fixed degree-day base to account for variations in heating of buildings, whereas it is possible that the temperatures at which heating is turned on is much more complex and interrelated with many of the variables. We also present models employing more co-variates than would be usual and which, as we have noted, display multicollinearity and instability as a result. We therefore plan more sophisticated analysis of the above data using weightings to produce population estimates, applying non-linear methods for inference and predictive models, and using variable selection methods to identify the most important individual factors. Further, we plan to use the full 13,000 observatory release 3 data to give greater statistical power and investigate the impact of coronavirus on energy demand, analyse both daily and half-hourly data. We also plan to analyse gas and electricity use separately to improve our understandings of the factors that correlate with each. There is considerable scope for research using the SERL Observatory data and we encourage UK academic researchers to submit proposals to access the data. More information about how to do this can be found on the SERL website ([www.serl.ac.uk](http://www.serl.ac.uk)) and UKDS data catalogue [10].

# Conclusions

This paper presents analysis of the SERL Observatory: a dataset of linked smart meter data and socio-technical contextual data for a representative sample of over 13,000 GB households. Here we analyse data from a sub-sample (N=617) of the first two recruitment waves (initial sample N=4716) and for the pre-coronavirus period (taken to be before March 2020).

The first aim was to quantify how much of the variation in total energy consumption can be explained by different combinations of SERL Observatory variables: ‘basic’ (e.g. local weather, region, date), EPC (where available), and the SERL survey (questions relating to the dwelling and occupants). As multiple observations were available per participant, linear mixed effects models were used to regress household-level daily total energy consumption against successive levels of contextual data over time to reveal the relationship between energy use and static (constant) and temporally changing variables (basic: weather, region, IMD and date; EPC; SERL survey; all data combined).

The explanatory power of the models was quantified using marginal R2 and root mean squared error (RMSE). Across all models, increasing contextual data increased explanatory power, from a starting marginal R2 of 0.36 for basic data only, to with a final marginal R2 of 0.68. RMSE for all models was similar 17.7kWh/day as all include random effects parameters. Adding EPC data or SERL survey to basic data increased explanatory power similarly (marginal R2 0.63 and 0.65 respectively), while SERL survey data is less affected by multicollinearity. Overall, our results compare favourably with those found in the literature, however we note that these results are based on within sample estimates; prediction errors should be expected to be higher.

The second aim was to identify statistically significant variables (*p*-value<0.05) observed in SERL Observatory data that are strongly associated with variation in household-level residential energy consumption. Buildings with larger floor area, that are older, have more rooms, have lower energy efficiency, and experience colder or less sunny weather were associated with increased energy consumption. Households with more occupants, more children, with older adult occupants, and with fewer adults with qualifications were also associated with increased energy consumption. Energy consumption in households was found to be lower in households that set lower heating temperature setpoints, that tried to save energy, and that put on more clothes rather than turning the heating on.

# Acknowledgements

This work has been funded by EPSRC through grant EP/P032761/1. 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 and to SERL Consortium Manager James O’Toole at UCL.

We acknowledge support from the SERL Independent Advisory Board, Data Governance Board and Research Programme Board which played critical rolee in the establishment and ethical operation of SERL.

The SERL Observatory includes European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 data. Neither the European Commission nor the European Centre for Medium-Range Weather Forecasts is responsible for any use that may be made of the Copernicus information or data it contains.

# Author contributions

**Eoghan McKenna**: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original Draft.

**Jessica Few**: Methodology, Writing - Original Draft, Visualization, Software, Formal analysis.

**Ellen Webborn**: Methodology, Data Curation, Software, Investigation, Writing - Review & Editing.

**Ben Anderson**: Methodology, Writing - Review & Editing.

**Simon Elam**: Data Curation, Writing - Review & Editing, Supervision, Project administration, Funding acquisition, Investigation.

**David Shipworth**: Investigation, Writing - Review & Editing, Funding acquisition.

**Martin Pullinger**: Investigation, Writing - Review & Editing.

**Adam Cooper**: Investigation, Writing - Review & Editing.

**Tadj Oreszczyn**: Investigation, Writing - Review & Editing, Supervision, Funding acquisition.

# References

[1] IEA, Net Zero by 2050, Paris, 2021. https://www.iea.org/reports/net-zero-by-2050.

[2] J. Wachsmuth, V. Duscha, Achievability of the Paris targets in the EU—the role of demand-side-driven mitigation in different types of scenarios, Energy Effic. 12 (2019) 403–421. https://doi.org/10.1007/s12053-018-9670-4.

[3] F. Creutzig, J. Roy, W.F. Lamb, I.M.L. Azevedo, W. Bruine De Bruin, H. Dalkmann, O.Y. Edelenbosch, F.W. Geels, A. Grubler, C. Hepburn, E.G. Hertwich, R. Khosla, L. Mattauch, J.C. Minx, A. Ramakrishnan, N.D. Rao, J.K. Steinberger, M. Tavoni, D. Ürge-Vorsatz, E.U. Weber, Towards demand-side solutions for mitigating climate change, Nat. Clim. Chang. 8 (2018) 268–271. https://doi.org/10.1038/s41558-018-0121-1.

[4] A.C.G. Cooper, Building physics into the social: Enhancing the policy impact of energy studies and energy social science research, Energy Res. Soc. Sci. 26 (2017) 80–86. https://doi.org/10.1016/j.erss.2017.01.013.

[5] J. Love, A.C.G. Cooper, From social and technical to socio-technical: Designing integrated research on domestic energy use, Indoor Built Environ. 24 (2015) 986–998. https://doi.org/10.1177/1420326X15601722.

[6] A.C.G. Cooper, Evaluating energy efficiency policy: understanding the ‘energy policy epistemology’ may explain the lack of demand for randomised controlled trials, Energy Effic. 11 (2018) 997–1008. https://doi.org/10.1007/s12053-018-9618-8.

[7] A. Cooper, D. Shipworth, A. Humphrey, UK Energy Lab: A feasibility study for a longitudinal, nationally representative sociotechnical survey of energy use, London, 2014. https://www.ucl.ac.uk/steapp/sites/steapp/files/synthesis.pdf (accessed May 20, 2021).

[8] G.M. Huebner, I. Hamilton, Z. Chalabi, D. Shipworth, T. Oreszczyn, Explaining domestic energy consumption - The comparative contribution of building factors, socio-demographics, behaviours and attitudes, Appl. Energy. 159 (2015) 589–600. https://doi.org/10.1016/j.apenergy.2015.09.028.

[9] E. Webborn, T. Oreszczyn, Champion the energy data revolution, Nat. Energy 2019 48. 4 (2019) 624–626. https://doi.org/10.1038/s41560-019-0432-0.

[10] S. Elam, E. Webborn, E. McKenna, T. Oreszczyn, B. Anderson, Ministry of Housing Communities & Local Government, European Centre for Medium-Range Weather Forecasts, Royal Mail Group Limited, Smart Energy Research Lab Observatory Data, 2019-2020: Secure Access, (2020). https://doi.org/http://doi.org/10.5255/UKDA-SN-8666-1.

[11] E. Shove, M. Pantzar, M. Watson, The dynamics of social practice: Everyday life and how it changes, Sage, Los Angeles, Calif. ; London, 2012.

[12] K. Gram-Hanssen, Efficient Technologies or User Behaviour, Which Is the More Important When Reducing Households’ Energy Consumption?, Energy Effic. 6 (2013) 447–457. https://doi.org/10.1007/s12053-012-9184-4.

[13] M. Hand, E. Shove, D. Southerton, Explaining showering: a discussion of the material, conventional, and temporal dimensions of practice, Sociol. Res. Online. 10 (2005). http://www.socresonline.org.uk/10/2/hand.html.

[14] R. V. Jones, A. Fuertes, K.J. Lomas, The socio-economic, dwelling and appliance related factors affecting electricity consumption in domestic buildings, Renew. Sustain. Energy Rev. 43 (2015) 901–917. https://doi.org/10.1016/j.rser.2014.11.084.

[15] A. Satre-Meloy, M. Diakonova, P. Grünewald, Daily life and demand: an analysis of intra-day variations in residential electricity consumption with time-use data, Energy Effic. 13 (2020) 433–458. https://doi.org/10.1007/s12053-019-09791-1.

[16] S. Wei, R. Jones, P. de Wilde, Driving Factors for Occupant-Controlled Space Heating in Residential Buildings, Energy Build. 70 (2014) 36–44. https://doi.org/10.1016/j.enbuild.2013.11.001.

[17] L.G. Swan, V.I. Ugursal, Modeling of end-use energy consumption in the residential sector: A review of modeling techniques, Renew. Sustain. Energy Rev. 13 (2009) 1819–1835. https://doi.org/10.1016/j.rser.2008.09.033.

[18] A. Satre-Meloy, Investigating structural and occupant drivers of annual residential electricity consumption using regularization in regression models, Energy. 174 (2019) 148–168. https://doi.org/10.1016/j.energy.2019.01.157.

[19] A. Foucquier, S. Robert, F. Suard, L. Stéphan, A. Jay, State of the art in building modelling and energy performances prediction: A review, Renew. Sustain. Energy Rev. 23 (2013) 272–288. https://doi.org/10.1016/j.rser.2013.03.004.

[20] A.T. Nguyen, S. Reiter, P. Rigo, A review on simulation-based optimization methods applied to building performance analysis, Appl. Energy. 113 (2014) 1043–1058. https://doi.org/10.1016/j.apenergy.2013.08.061.

[21] K. Amasyali, N.M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renew. Sustain. Energy Rev. 81 (2018) 1192–1205. https://doi.org/10.1016/j.rser.2017.04.095.

[22] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, X. Zhao, A review of data-driven approaches for prediction and classification of building energy consumption, Renew. Sustain. Energy Rev. 82 (2018) 1027–1047. https://doi.org/10.1016/j.rser.2017.09.108.

[23] Y. Iwafune, Y. Yagita, High-resolution determinant analysis of Japanese residential electricity consumption using home energy management system data, Energy Build. 116 (2016) 274–284. https://doi.org/10.1016/j.enbuild.2016.01.017.

[24] M.J. Kim, Understanding the determinants on household electricity consumption in Korea: OLS regression and quantile regression, Electr. J. 33 (2020) 106802. https://doi.org/10.1016/j.tej.2020.106802.

[25] F. McLoughlin, A. Duffy, M. Conlon, Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study, Energy Build. 48 (2012) 240–248. https://doi.org/10.1016/j.enbuild.2012.01.037.

[26] B. Anderson, S. Lin, A. Newing, A.B. Bahaj, P. James, Electricity consumption and household characteristics: Implications for census-taking in a smart metered future, Comput. Environ. Urban Syst. 63 (2017) 58–67. https://doi.org/10.1016/j.compenvurbsys.2016.06.003.

[27] E.W. Frees, Longitudinal and Panel Data, Cambridge University Press, 2004. https://doi.org/10.1017/cbo9780511790928.

[28] G. James, D. Witten, T. Hastie, R. Tibshirani, An introduction to statistical learning, Springer, 2017. https://link.springer.com/content/pdf/10.1007/978-1-4614-7138-7.pdf (accessed February 5, 2021).

[29] G. James, D. Witten, T. Hastie, R. Tibshirani, An introduction to Statistical Learning, 2000. https://doi.org/10.1007/978-1-4614-7138-7.

[30] J. Friedman, T. Hastie, R. Tibshirani, The elements of statistical learning, 2001. http://statweb.stanford.edu/~tibs/book/preface.ps (accessed February 5, 2021).

[31] A. Kavousian, R. Rajagopal, M. Fischer, Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants’ behavior, Energy. 55 (2013) 184–194. https://doi.org/10.1016/j.energy.2013.03.086.

[32] MHCLG, English Housing Survey 2017 to 2018: energy, 2019. https://www.gov.uk/government/statistics/english-housing-survey-2017-to-2018-energy.

[33] ACER, Annual report on the results of monitoring the internal electricity and natural gas markets in 2017, 2018.

[34] R. V. Jones, K.J. Lomas, Determinants of high electrical energy demand in UK homes: Appliance ownership and use, Energy Build. 117 (2016) 71–82. https://doi.org/10.1016/j.enbuild.2016.02.020.

[35] H. Fan, I.F. MacGill, A.B. Sproul, Statistical analysis of driving factors of residential energy demand in the greater Sydney region, Australia, Energy Build. 105 (2015) 9–25. https://doi.org/10.1016/j.enbuild.2015.07.030.

[36] G. Huebner, D. Shipworth, I. Hamilton, Z. Chalabi, T. Oreszczyn, Understanding electricity consumption: A comparative contribution of building factors, socio-demographics, appliances, behaviours and attitudes, Appl. Energy. 177 (2016) 692–702. https://doi.org/10.1016/j.apenergy.2016.04.075.

[37] H. Fan, I.F. MacGill, A.B. Sproul, Statistical analysis of drivers of residential peak electricity demand, Energy Build. 141 (2017) 205–217. https://doi.org/10.1016/j.enbuild.2017.02.030.

[38] BEIS, NEED Annex D: Determinants of household gas use, 2019.

[39] T. Snijders, R. Bosker, Multilevel analysis: An introduction to basic and advanced multilevel modeling, SAGE, 2012.

[40] E. Webborn, S. Elam, E. McKenna, Utilising Smart Meter Data for Research and Innovation in the UK (forthcoming), in: Proc. Eur. Counc. an Energy Effic. Econ. Summer Study, 2019.

[41] E. Webborn, E.J. McKenna, S. Elam, B. Anderson, A. Cooper, T. Oreszczyn, Increasing response rates and reducing bias: Learnings from the Smart Energy Research Lab pilot study, (n.d.). https://doi.org/10.31219/OSF.IO/F82B7.

[42] E. Webborn, S. Elam, E. McKenna, T. Oreszczyn, Utilising smart meter data for research and innovation in the UK, ECEEE Summer Study Proc. (2019) 1387–1396.

[43] E. Webborn, E.J. McKenna, S. Elam, B. Anderson, A. Cooper, T. Oreszczyn, Increasing response rates and reducing bias: Learnings from the Smart Energy Research Lab pilot study, OSF Prepr. (2021). https://doi.org/10.31219/OSF.IO/F82B7.

[44] J. Crawley, E. McKenna, V. Gori, T. Oreszczyn, Creating Domestic Building Thermal Performance Ratings Using Smart Meter Data, Build. Cities. 1 (2020) 1–13. https://doi.org/10.5334/BC.7.

[45] MHCLG, Energy Performance of Buildings Data: England and Wales, (2020). https://epc.opendatacommunities.org/.

[46] J. Crawley, P. Biddulph, P.J. Northrop, J. Wingfield, T. Oreszczyn, C. Elwell, Quantifying the Measurement Error on England and Wales EPC Ratings, Energies. 12 (2019).

[47] H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, J.-N. Thépaut, ERA5 hourly data on single levels from 1979 to present., (2018). https://doi.org/10.24381/cds.adbb2d47.

[48] DECC, Energy Trends: December 2014, special feature article - Energy usage in household with solar PV installations, 2014. https://www.gov.uk/government/statistics/energy-trends-december-2014-special-feature-article-energy-usage-in-household-with-solar-pv-installations.

[49] Brook Lyndhurst, Uptake of Ultra Low Emission Vehicles in the UK, 2015. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\_data/file/464763/uptake-of-ulev-uk.pdf.

[50] J. Spinoni, J. Vogt, P. Barbosa, European degree-day climatologies and trends for the period 1951-2011, Int. J. Climatol. 35 (2015) 25–36. https://doi.org/10.1002/joc.3959.

[51] D. Iacobucci, M.J. Schneider, D.L. Popovich, G.A. Bakamitsos, Mean centering helps alleviate “micro” but not “macro” multicollinearity, Behav. Res. Methods. 48 (2016) 1308–1317. https://doi.org/10.3758/S13428-015-0624-X.

[52] M. Wissmann, H. Toutenburg, Role of categorical variables in multicollinearity in the linear regression model, 2007. https://epub.ub.uni-muenchen.de/2081 (accessed August 23, 2021).

[53] MHCLG, English Housing Survey 2018 to 2019: headline report, 2020. https://www.gov.uk/government/statistics/english-housing-survey-2018-to-2019-headline-report (accessed May 26, 2021).

[54] BEIS, Energy consumption in the UK - GOV.UK, 2021. https://www.gov.uk/government/statistics/energy-consumption-in-the-uk (accessed May 26, 2021).

[55] J. Wooldridge, Introductory econometrics: A modern approach, Cengage learning, 2015. https://books.google.com/books?hl=en&lr=&id=wUF4BwAAQBAJ&oi=fnd&pg=PR3&dq=wooldridge+introductory+econometrics&ots=cATyYDlngo&sig=AkalfyXzQggN67iYhrU5UKaKCH0 (accessed September 10, 2021).

[56] S. Nakagawa, H. Schielzeth, A general and simple method for obtaining *R* 2 from generalized linear mixed-effects models, Methods Ecol. Evol. 4 (2013) 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x.

[57] P.C.D. Johnson, Extension of Nakagawa &amp; Schielzeth’s *R* 2 GLMM to random slopes models, Methods Ecol. Evol. 5 (2014) 944–946. https://doi.org/10.1111/2041-210X.12225.

[58] J. Miles, R -Squared, Adjusted R -Squared , Encycl. Stat. Behav. Sci. (2005). https://doi.org/10.1002/0470013192.BSA526.

[59] T. pandas development Team, Pandas, (2020). https://doi.org/10.5281/zenodo.3509134.

[60] W. McKinney, Data Structures for Statistical Computing in Python, in: S. van der Walt, J. Millman (Eds.), Proc. 9th Python Sci. Conf., 2010: pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.

[61] S. Seabold, J. Perktold, Statsmodels: Econometric and statistical modeling with python, in: Proc. 9th Python Sci. Conf., 2010: p. 92.

[62] B. Anderson, J. Torriti, Explaining shifts in UK electricity demand using time use data from 1974 to 2014, Energy Policy. 123 (2018) 544–557. https://doi.org/10.1016/j.enpol.2018.09.025.

[63] H. Wallis, M. Nachreiner, E. Matthies, Adolescents and electricity consumption; Investigating sociodemographic, economic, and behavioural influences on electricity consumption in households, Energy Policy. 94 (2016) 224–234. https://doi.org/10.1016/j.enpol.2016.03.046.

[64] K. Steemers, G.Y. Yun, Household energy consumption: A study of the role of occupants, Build. Res. Inf. 37 (2009) 625–637. https://doi.org/10.1080/09613210903186661.

[65] G. Huebner, M. Fell, N. Watson, Improving energy research practices: guidance for transparency, reproducibility and quality, Build. Cities. 2 (2021) 1–20. https://doi.org/10.5334/bc.67.

1. Government statistical estimate of relative deprivation in small areas – see <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019> [↑](#footnote-ref-2)
2. Note that only SMETS2 meters record daily readings, but all record half-hourly. [↑](#footnote-ref-3)
3. The Smart Data Communications Company (DCC) is the central communications infrastructure for the GB smart meter network. [↑](#footnote-ref-4)
4. SERL retains the most recent version. [↑](#footnote-ref-5)
5. Second-generation GB smart meters ("SMETS2”) can retain up to 13 months historic half-hourly consumption data. [↑](#footnote-ref-6)
6. From the ONS Address Base, excluding Scotland as the SERL Observatory does not currently contain Scottish EPC data. [↑](#footnote-ref-7)
7. Population proportions are for England for 2018-2019 and are taken from [53] [↑](#footnote-ref-8)
8. The exception was that the descriptive variable for secondary heating was included as the variable describing its energy efficiency had no data. [↑](#footnote-ref-9)