Research with household energy data at scale

The challenge

It is only in the past decade that 'big data' or 'new forms of data' from sources like social media, digital sensors, financial and administrative transactions have become available as data commodities for the social scientist. The panacea of big data has led to a focus on developing solutions for powerful analytics rather than sustainable and safe solutions for data curation and delivery. The UK Data Service has expertise on data provenance and trustworthiness, ethical and legal entitlements, structure and quality, and is working to deliver solutions across the whole of the big data life cycle. A first set of case studies has focussed on household energy research, an area that demands huge amounts of data and computationally challenging storage and analyses, as well as increased disclosure risks that come with data linkage. Solutions of scale for researchers in this field are needed for data exploration, analysis, visualisation and data linkage.

"The UK Data Service has expertise on data provenance and trustworthiness, ethical and legal entitlements, structure and quality, and is working to deliver solutions across the whole of the big data life cycle.

Research examples

Readings of household energy consumption, such as electricity or gas consumption, collected at frequent intervals by electronic or 'smart' meters from just a few hundred houses, can accumulate hundreds of millions of data points, growing to terabytes in size in just a few years. This scale presents technical challenges in storing, managing, analysing, and sharing the data. As part of the Smarter Household Energy Data project, we have been liaising with UK energy researchers to investigate and plan for data management, infrastructure and access solutions to some of the technical challenges and limitations imposed by local systems and software. There is a longer-term need for household energy data to be made available to the wider research community beyond the duration of specific research projects, but this also has sensitivities.

The researchers we spoke to are at the forefront of policy-relevant interdisciplinary research, investigating energy demand and carbon emissions reduction, energy savings and mitigation of fuel poverty. They work with key stakeholders including government departments, Research Councils, regulatory authorities, energy companies and Distribution Network Operators.

Powerful research on household energy consumption is enabled through the use of various data types:

- Smart meters, building energy performance certificates and surveys of appliances that record consumption;
- Social surveys that report on demographic, attitudinal and behavioural information;
- Intervention studies that gather information from installing control devices, providing information and education through community-based initiatives, or retrofit and structural improvements such as insulation upgrades.

Typical analyses conducted on smart meter data include quantifying, monitoring and assessing the effectiveness of interventions in homes. Linking the data to socio-economic and demographic context, dwelling types, and external data sources, such as weather-related and Census data, enriches contextual information.

UK Data Service

Data and data issues

Many projects in this domain collect fine-grained data from gas and electricity readings at one or ten second intervals. Data at a sufficiently high resolution like this can even detect specific appliances (requires data at less than five minute intervals). As household-level data are typically held as separate files for each household, data storage and processing challenges are quickly accumulated. Often, researchers will want to aggregate, subset and download data for analysis with local systems and software, but are unable to open such large files resulting from these extractions, even using R, due to the memory of the average laptop or PC. Thus specialised systems and tools are needed to be able to analyse and visualise the data at scale.

Finally, since a key requirement for using smart meter data is to be able to link it to external data sources to gain greater research insights, this brings with it data processing and other concerns around consent and privacy.

Specific projects challenges

Project 1	Temperature and Energy Monitoring Post and prE Retrofit (TEMPER)
PIs and funder:	Leeds Beckett Univerity's Sustainability Institute funded by the department for Business, Energy and Industrial Strategy (BEIS)
Aim and potential benefits:	To collect data to quantify the reduction in fuel bills achieved by domestic retrofits in the UK.
Data challenge:	Researchers find MS Excel to be a useful exploratory data analysis tool to graph and visualise data and to spot anomalies, but Excel's graphics soon slow down with a high number of data points. Using R or Excel, the researchers often produce daily averages for the more fine-grained data for each household file, then aggregate the data up to a more manageable dataset. Models enable the identification of poor control of heating e.g. when people are putting the heating on when it's 20 degrees outside, which could identify erroneous data or poor heating control systems (e.g. thermostat not well calibrated).

Project 2	Solent Achieving Value from Efficiency (SAVE)
PIs and funder:	University of Southampton's Sustainable Energy Research Group (SERG) funded by Ofgem and Scottish and Southern Electricity Power Distribution (SSEPD)
Aim and potential benefits:	To understand the nature of temporal electricity demand and household responses to interventions and incentive, and help Distribution Network Operators determine what to do where in their networks and whether to target infrastructure or behaviour intervention in specific areas.
Website:	SAVE: Solent Achieving Value from Efficiency
Data challenge:	Data on household profiles is collected from a nationally representative sample of thousands of households in the Solent region, including survey data, time-use diaries, and fine-grained smart meter energy consumption data at fifteen minute intervals and power data sampled every ten seconds. The researchers undertake spatial micro simulation through combining Census Lower Super Output Area level outputs with household microdata to predict intervention results if rolled out across the whole population. Datasets are large and simulation modelling is computationally intensive.

Project 3	Role of Community-Based Initiatives in Energy Saving, 2010-2014 (CBIES)
PIs and funder:	University of Southampton's Sustainable Energy Research Group (SERG) funded the Economic and Social Research Council (ESRC)
Aim and potential benefits:	To gain a better understanding of the role that community-based initiatives can play in fostering net energy savings in UK households. Using fine-grained data, appliances can be detected by creating a 'signature' for each appliance. This could help predict whether and when an appliance will fail by analysing its energy cycles and provide informed intervention that could help load-shift from key peak times for the National Grid.
Website:	The role of community-based initiatives in energy saving
Data challenge:	Data analysed for the project is gathered from hundreds of households in the Hampshire area, derived from: an energy and spending survey; a social network survey; qualitative interviews; and gas and electricity meter data collected at one second intervals. Datasets with such fine-grained information are huge and data handling and modelling quickly becomes computationally intensive.

Quality of smart meter data

Smart meter data offers researchers access to raw, unadjusted data but suffer from some common data quality issues such as:

- Sources to link to are often heavily anonymized
- Data collection can be limited with biased samples, for example, using recruitment-based studies
- Poor provenance information with little standardised documentation
- Complex governance
- Limited reproducibility
- Often missing and duplicate observations with lack of standardized coding, for example NA, NULL, 99, '99', etc.; timestamp formats inconsistent with different combinations of date, time, and date and time columns, and handling of time zones and daylight saving

Centres of expertise in curation like the UK Data Service can help to define data, documentation and metadata standards and contribute to methods of standardising how energy data is stored and shared to encourage reproducibility. A unified and secure interface to smart meter data that can help researchers and policymakers is already being planned in the UK as part of a funded Smart Meter Research Portal project.

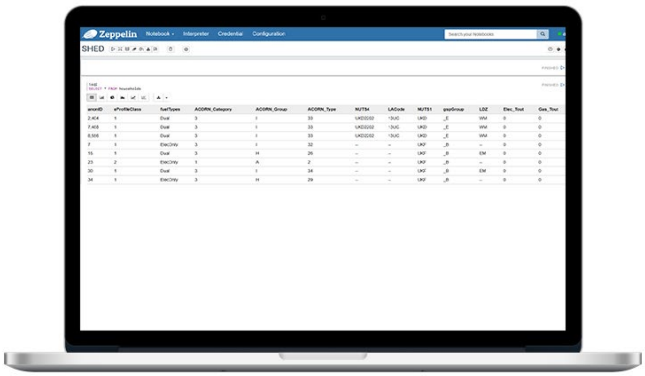
The technical solution

As Data Archives diversify their holdings to include big data, such as high-volume energy data, big data platforms like Apache Hadoop can offer increased computational and storage capabilities. Hadoop sits at the core of the UK Data Service's new data infrastructure, known as our Data Service as a Platform (DSaaP). The large multi-faceted energy datasets we have described can be ingested into the Hadoop system, facilitating easier storage, management, analysis and dynamic visualisation of the data.

The input from energy researchers has helped us to plan how we can best store and access data, think about how our security framework is applied, and consider what tools and views of the data users will be most useful.

Smart meter data: data manipulation and visualisation

Analytical tools integrated into the Hadoop system, such as Apache Zeppelin, support multiple querying with R, Python, Scala, and Structured Query Language (SQL), and provide powerful visualisation tools through a standard web browser. Data are loaded as data frames which present a view of data similar to a spreadsheet, SPSS, or database, where variables can be dynamically graphically explored and visualised, with drag and drop. Data frames also support methods for slicing and dicing data and are a standard way of working for data science, routinely used in packages such as R and Python.



Data table view

Using data from the Energy Demand Research Project archived at the UK Data Service, as an example, this has 18 million rows of smart meter data resulting from smart meter trials by energy suppliers from 2007-2011. Distributions of the variables can be rapidly and dynamically visualised using 'out of the box' features. Using Zeppelin, the data can be rapidly aggregated data from the two linked tables: household energy consumption from smart meter readings; and limited household attributes, linked via an anonymised household identifier. Figure 1a shows a plot of the count of fuel-type households (duel fuel and electricity only) by geographical region (NUTS1); and Figure 1b graphs mean household electricity usage over an average twenty-four hour period in December 2009 for fuel-type households, showing that dual fuel usage households consume less electricity on a winter's day than electricity-only households. Figure 1c shows a line plot of mean daily temperature superimposed on a bar chart of mean electricity consumption, and is based on spatial and temporal aggregations of energy consumption linked to open data from an external source, the UK Meteorological Office.

Any of the outputs produced including derived data tables, descriptive statistics, and visualisations can be easily stored on the Hadoop platform and, authorisation permitting, downloaded for use or reproduction elsewhere.

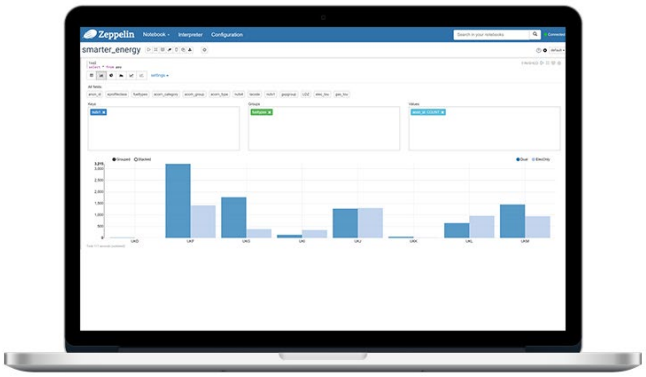


Figure 1a. Cross-tabulation comparing fuel-type households by region. Source: EDRP, 2007-2010

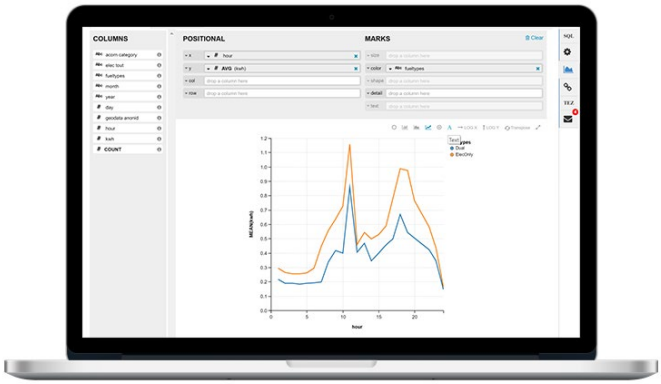


Figure 1b. Mean energy consumption by fuel-type household in one day in December 2009. Source: EDRP, 2007-2010

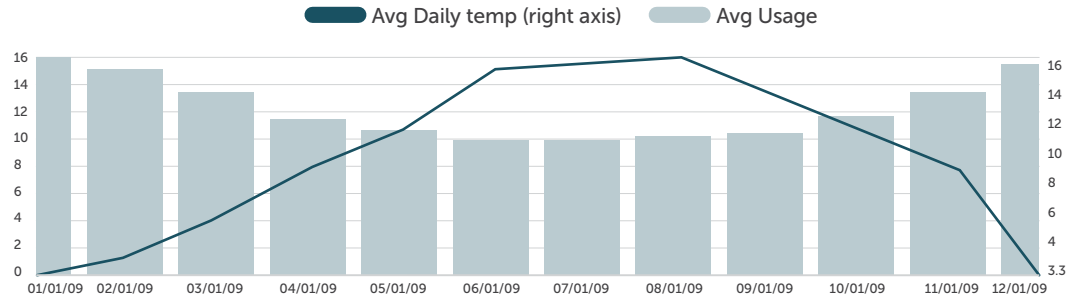


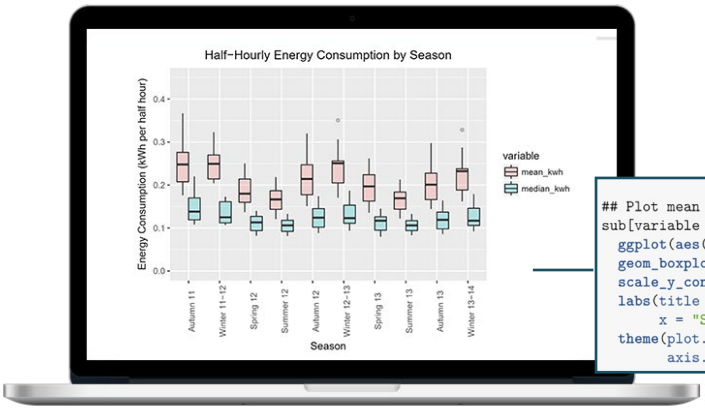
Figure 1c. Line plot of daily mean temperature over bar chart of mean energy consumption for a region

Smart meter data: data extraction, evaluation and cleaning

R is a useful tool for working with smart meter data. Open source data from Low Carbon London (LCL) based on meter readings from a broadly representative sample of Greater London of over 5,500 households between 2011-15 are used to show how to:

- evaluate data looking at the sampling rate, for any inconsistent time stamps, missing values etc. (Notebook 1)
- clean a large (11GB) dataset using the 'data.table' package, which provides an efficient implementation of R's data frame structure (Notebook 2).

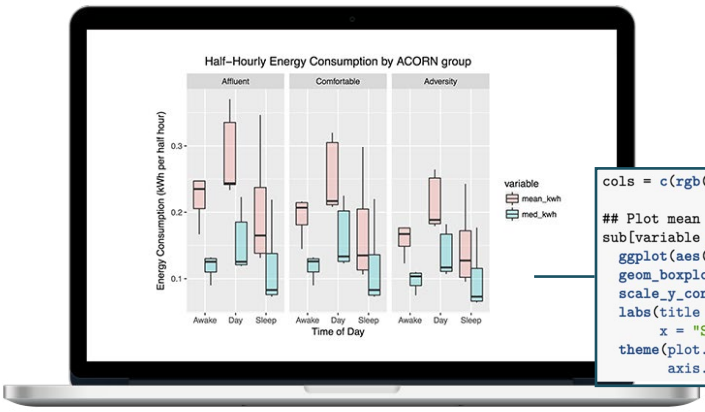
Once data are cleaned they can be explored in more detail, for example exploring seasonal effects (Figure 3a), or social class differences (Figure 3b) in consumption energy across all households.




```
## Plot mean and median kwh_hh by season.
sub[variable != "iqr_kwh"] %>%
  ggplot(aes(seas, value)) +
  geom_boxplot(aes(fill = variable), alpha = 0.25) +
  scale_y_continuous(limit = c(0, .45)) +
  labs(title = "Half-Hourly Energy Consumption by Season",
       x = "Season", y = "Energy Consumption (kWh per half hour)") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 90))
```

R code for plot in Figure 3a

Figure 3a. Plot of mean and median energy consumption by season




```
cols = c(rgb(1, 0, 0, .3), rgb(0, 0, 1, .3), rgb(0, 1, 0, .3))

## Plot mean and median kwh_hh by season.
sub[variable != "iqr_kwh"] %>%
  ggplot(aes(seas, value)) +
  geom_boxplot(aes(fill = variable), alpha = 0.25) +
  scale_y_continuous(limit = c(0, .45)) +
  labs(title = "Half-Hourly Energy Consumption by Season",
       x = "Season", y = "Energy Consumption (kWh per half hour)") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 90))
```

R code for plot in Figure 3b

Figure 3b. Plot of mean energy consumption by social class (ACORN group)

Dashboards for energy research

Dashboards that can monitor the status of ongoing trials, detect real-time errors and anomalies are useful research tools for projects that are receiving real-time data. Hadoop can be used to ingest, route and process real-time events in an efficient and scalable way. Researchers can choose the framework for building such a dashboard, using R packages like shiny, html widgets and so on, which would integrate well with the Service's data platform.

See our case studies on:

- Scaling up: digital data services for the social sciences
- Utilising smart meter data to enable energy demand research

Authors:

Louise Corti, Karen Dennison and Chris Park, UK Data Service.