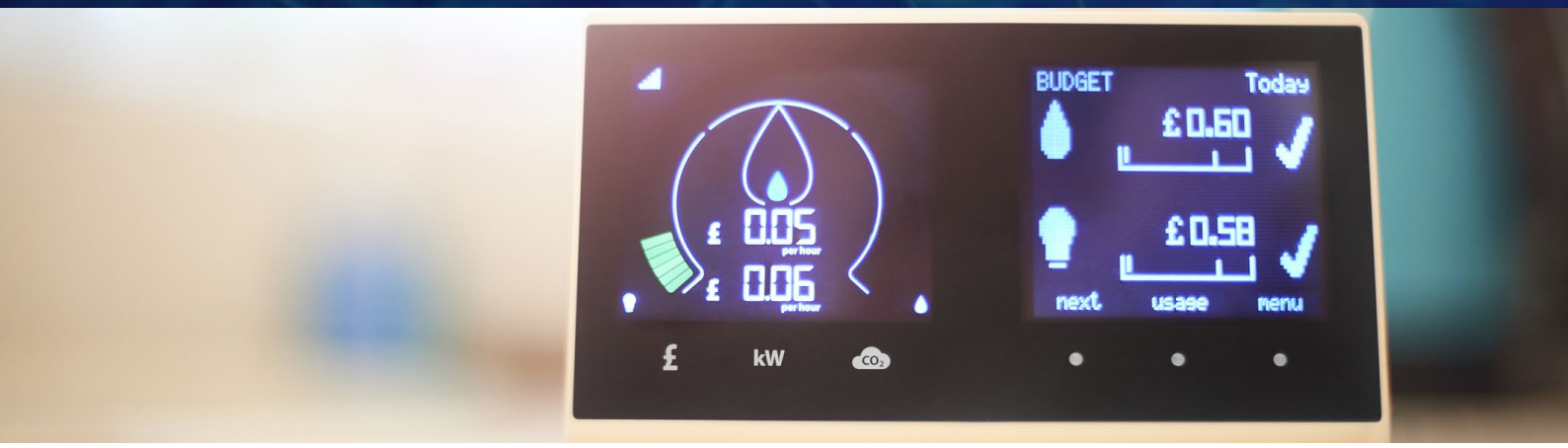
Data Service as a Platform

Researching the thermal character of UK dwellings



The challenge

In order to make effective energy policy, governments are interested in gaining an accurate and cost-effective understanding of energy demand in households. The current system is to use Energy Performance Certificates (EPCs), but these require on-site visits, which are expensive, and don't always provide accurate results.

Energy epidemiology is a relatively new area of research that aims to help reduce energy demand and contribute to effective and sustainable energy policy. At the heart of this practice is data, from which researchers develop empirically-grounded statistical models to build up a picture of underlying drivers of energy demand in UK dwellings.

The UK Data Service's Data Services as a Platform (DSaap) provides an integrated, safe environment where they can analyse smart meter data and weather data together to build new and innovative models.

> The UK Data Service provides a safe environment in which researchers can analyse smart meter and weather data together to build innovative new models to study household energy demand.

Research example

At the RCUK Centre for Energy Epidemiology, based at UCL Energy Institute, a research project by Tadj Oreszczyn and Jonathon Chambers sought to develop improved techniques for using smart meter data to measure energy efficiency in UK homes. They aimed to apply big data methods to smart meter data to generate empirical models of dwellings based on their energy consumption profiles and their related weather data.

Data and data issues

The introduction of smart meters in the UK has provided high quality, high-resolution electricity and gas consumption data across millions of homes across the UK. These datasets offer enormous potential, but this cannot be exploited by using traditional 'desktop' methods of analysis; older data processing frameworks struggle to cope with the volume of these data, especially when combined with computationally intensive analytics.

Researchers were spending most of their time trying to process the data, rather than addressing important research questions.

The availability of time series data, including energy consumption series from smart meters, local weather series, as well as contextual information dwellings, were crucial to this project. A summary of the key datasets used in this project is provided below.

UK Data Service

Dataset	Format	Size	Details	Access	User license
Climate Forecast Reanalysis System Owner: National Center for Atmospheric Research	.nc	10GB	Global, high-resolution gridded weather dataset drawn from sophisticated numerical weather prediction (NWP) models. Large temporal and spatial range with continuous weather data over a 31-year period from 1979 to 2010	Safe guarded (registration)	Data sharing agreement
Energy Demand Research Project: Early Smart Meter Trials, 2007-2010 (EDRP) Owner: Department of Energy and Climate Change (DECC)	.csv	12GB	Smart meter data from a 2007-10 trial to see how consumers react to improved information about their energy consumption, and to test smart metering infrastructure prior to and during the UK smart meter rollout. N = 18,000+ dwellings	Safe guarded (registration)	UKDS End User Agreement
EDF Smart Meter data Owner: EDF Energy	.csv	5GB	EDF Energy provided their subset of the EDRP smart meter data (above) with more detail	On request	Data sharing agreement
Solid Wall Insulation Field Trial (SWIFT) Owner: DECC / Energy Savings Trust	.csv, .xlsx, .pdf	10GB	Monitoring data (2010 to 2012) trials from 90 solid wall dwellings with pre and post-insulation monitoring to investigate the performance of various forms of solid wall insulation. Data comprised heat flux measurements, temperature monitoring, energy meter data, pressure test and thermal imagery.	On request	Data sharing agreement

Each dataset came in a different format, presenting different processing challenges. Much of the data suffered from general issues such as missing values and gaps, duplicates, inconsistent formatting, as well as more specific problems such as understanding timestamps and adjusting for daylight savings. The weather data may not account for local variations in weather, such as changes in wind speed and direction due to the configuration of nearby buildings. Errors in the energy time series data were generally poorly documented, and overall there was limited contextual and metadata available. Considerable work went into consolidating input and the data extraction processes thus required a bespoke extraction, transformation, and load workflow for each dataset. See Fig. 1:

> The use of the UKDS Data Services as a Platform made a real difference to this project. The increased performance and scalability allowed UCL researchers to refine and test computationally intensive models at scale, while the suite of tools facilitated interactive exploratory analysis in much shortened timeframes. In effect, near real-time analysis on billions of data points is now possible while operating in a trusted, secure environment.
>
> Professor Tadj Oreszczyn, Director of UCL Centre for Energy Epidemiology

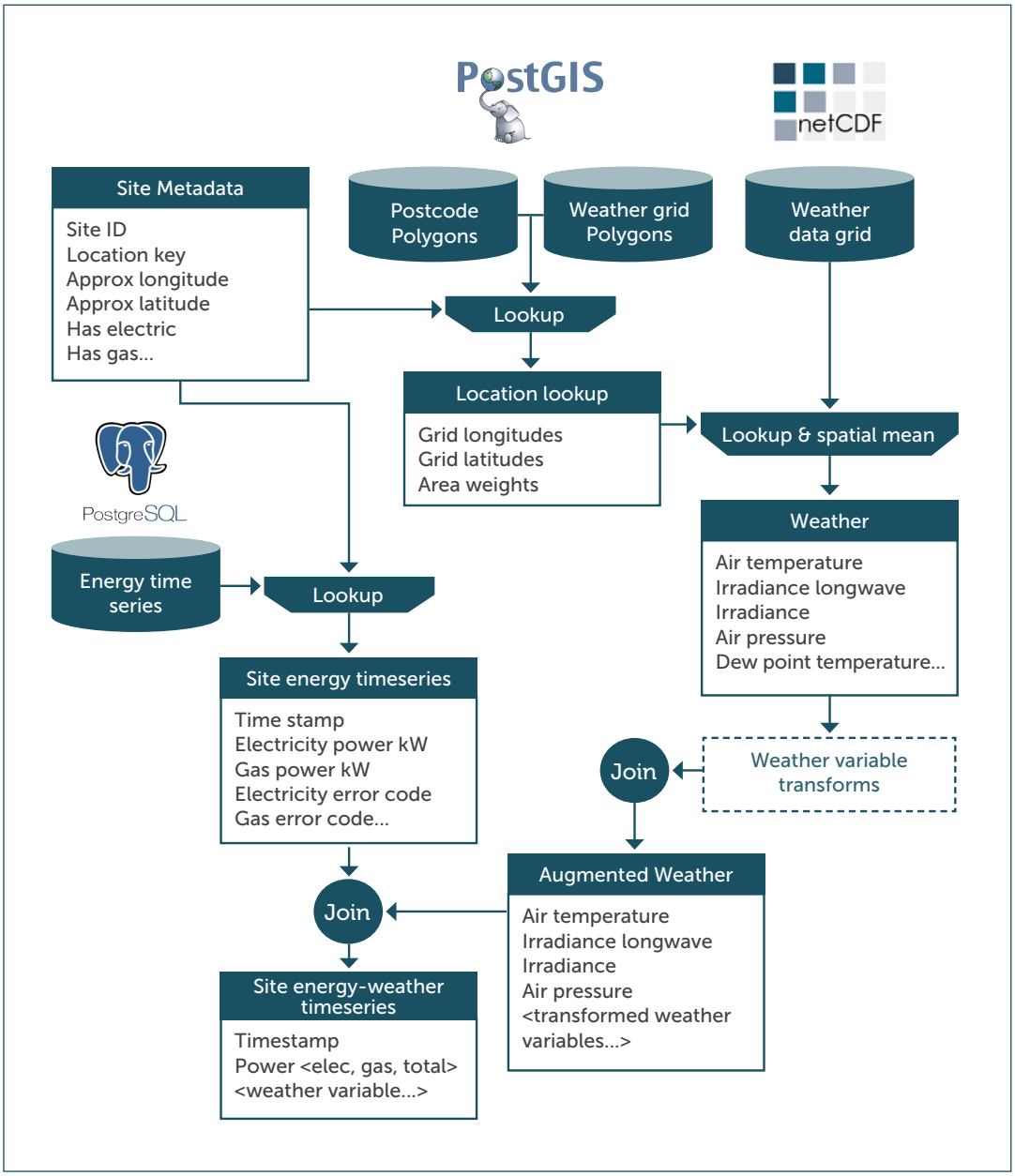


Figure 1. Original research workflow. Source: Chambers PhD Thesis

Our solution

The UK Data Service offers a solution to cope with the volume and complexity of the datasets involved, as well as the complexity of the analysis involved and data security concerns. Our Data Services as a Platform (DSaaP) initiative uses Hadoop software to store and process large data collections.

The personal datasets were safely stored in the UK Data Archive's ISO-certified, trusted digital repository, supported by a comprehensive framework of policies and procedures with legal, physical, and technical safeguards for information and risk management. DSaaP's dataflow management platform, based on the US National Security Agency's Apache NiFi, allowed us to move the data securely from disparate sources into DSaaP's Hadoop data clusters.

The Hadoop platform can process hundreds of millions of records in seconds or minutes instead of hours or days. The existing codebase, written in the Python programming language, was re-implemented in Spark's Python API, PySpark, to make sure the code is maintainable, and take advantage of the substantial performance benefits of Spark's distributed computation engine. The analytics platform in DSaaP provided data exploration and analysis tools, including Apache Spark, a high performance distributed computation engine, and Jupyter Notebooks, an integrated environment for interactive data analysis and visualisation. Examples of visualisations of annual energy consumption of households in the UK Energy Demand Research Project (EDRP) are shown in Figures 2a and 2b.

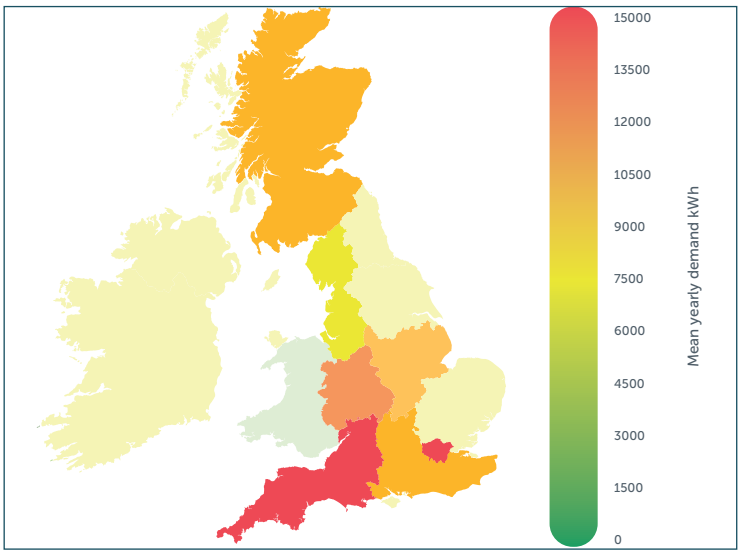


Figure 2a. Regional distribution of annual energy consumption of UK EDRP dwellings

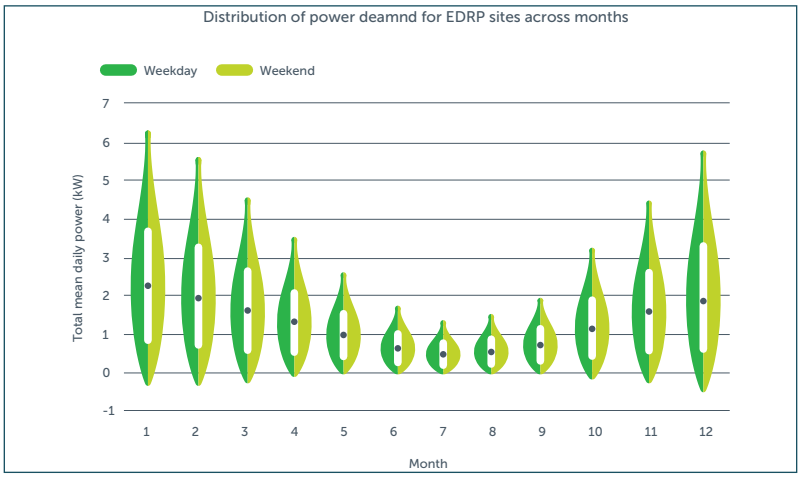


Figure 2b. Weekday vs. weekend distribution of annual energy consumption of UK EDRP dwellings

Summary of enhancements using big data platform and tools

Performance	Reduced query times over 2.5 billion observations from four hours to 15 seconds. For example, the mean yearly energy demand could be computed across all sites and visualised, as shown in the two figures above.
Productivity	Reduced the development time of new models from approximately two weeks to 20 minutes, by dramatically reducing the time spent on tackling the challenges associated with processing high volumes of data.
Scalability	Scaled out the analysis from a single site to over 8,000 sites, with up to 2.5 billion observations per query, which led to better model design and error detection. For example, applying an outlier detection filter on a single site on local desktop ran for an hour then crashed. With DSaaP, the filter could be run on over 8,000 sites in five minutes.
Simplicity	Provided a single point of access to disparate data sources and interactive data analysis environments, allowing researchers to easily carry out a wide variety of analyses at scale.

See our case studies on:

- Utilising smart meter data to enable household energy demand research
- Research with household energy data at scale
- Scaling up: digital data services for the social sciences

Authors:

Jonathon Chambers, Centre for Energy Epidemiology, UCL

Chris Park and Louise Corti, UK Data Service