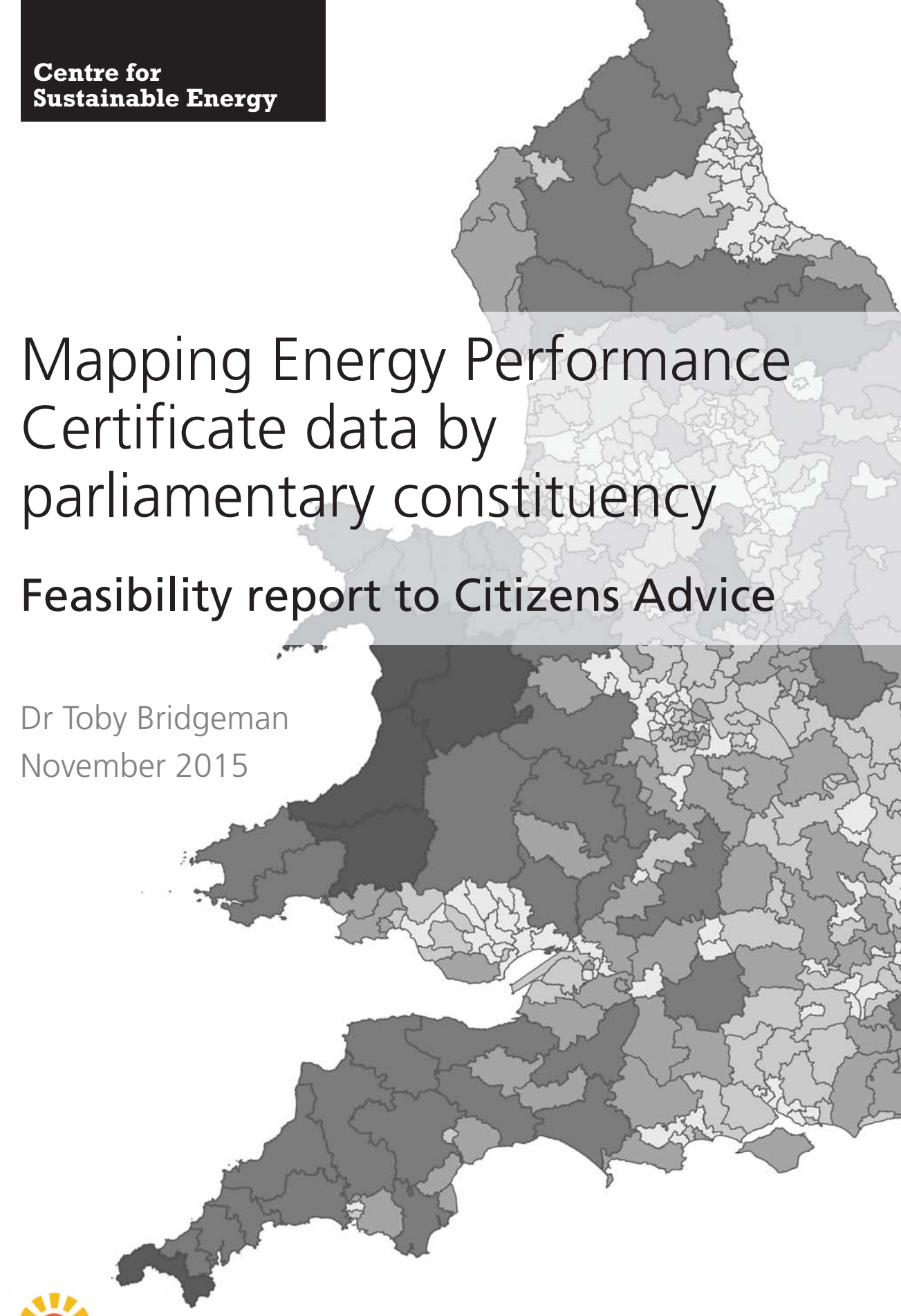
Centre for Sustainable Energy



Mapping Energy Performance Certificate data by parliamentary constituency

Feasibility report to Citizens Advice

Dr Toby Bridgeman

November 2015

Centre for Sustainable Energy

3 St Peter's Court
Bedminster Parade
Bristol BS3 4AQ

0117 934 1400
www.cse.org.uk
info@cse.org.uk

We are a national charity that shares our knowledge and experience to help people change the way they think and act on energy.

(Charity 298740)

Sign up to our monthly enews at www.cse.org.uk or follow us on Twitter at @cse_bristol

Acronyms

Acronym	Definition
BRE	Building Research Establishment
BREDEM	Building Research Establishment Domestic Energy Model
CERT	Carbon Emissions Reduction Target
CESP	Community Energy Saving Programme
CHAID	Chi-square Automatic Interaction Detection
CIGA	Cavity Insulation Guarantee Agency
COA	Census Output Area
CSE	Centre for Sustainable Energy
DCLG	Department for Communities and Local Government
DECC	Department of Energy and Climate Change
DPA	Data Protection Act
EBR	Energy Performance of Buildings Register
ECO	Energy Company Obligation
EHS	English Housing Survey
EPB	Energy Performance of Buildings
EPC	Energy Performance Certificate
EST	Energy Saving Trust
EUL	End User License File
FENSA	Fenestration Self-Assessment Scheme
FOI	Freedom of Information
HEC	Home Energy Check
HEED	Homes Energy Efficiency Database
ICO	Information Commissioners Office
IMD	Index of Multiple Deprivation
LIHC	Low Income High Costs fuel-poor definition
LSOA	Lower Level Super Output Area
MSOA	Middle Level Super Output Area
NEED	National Energy Efficiency Data-Framework
NHM	National Household Model
NLPG	National Land and Property Gazetteer
OAC	Output Area Classification
ONS	Office for National Statistics
ONSPD	Office for National Statistics Postcode Directory
OSG	One Scotland Gazetteer
PUF	Public Use File
RDSAP	Reduced Data Standard Assessment Procedure
SAP	Standard Assessment Procedure
SHCS	Scottish Housing Condition Survey
SIMD	Scottish Index of Multiple Deprivation
VOA	Valuation Office Agency

Contents

1 Executive summary

Statistics and data on fuel poverty and the energy efficiency standards of the UK housing stock are typically not accessible at regional, sub-regional and small area levels. They are either unpublished or in a format that neither allows detailed analysis of the underlying data nor can be easily understood by those outside the energy sector.

For example, data on Energy Performance Certificate (EPC) bands is already published by the Department for Communities and Local Government (DCLG) and the Scottish Government at aggregate local authority level. This does not correspond to parliamentary constituencies and therefore while politicians may understand the issues at a strategic level, they have limited means of assessing the extent of the problem in their constituency compared with other areas. Furthermore this does not go down to a local level sufficient to be of useful practical purpose or to be meaningful to local residents or the public.

Making housing energy efficiency data available at small area levels would allow a particular local area to be compared against national targets and policies. This would allow those concerned with energy efficiency and fuel poverty to highlight the extent of homes with poor energy performance to MPs. It would also provide a valuable resource to those wanting to better target local energy efficiency and fuel poverty programmes.

Citizens Advice commissioned the Centre for Sustainable Energy (CSE) to investigate the options for mapping energy efficiency statistics at small area level, primarily parliamentary constituency. The ultimate vision for this project is to produce a web-based, interactive mapping tool that can be used to:

1. highlight problems of fuel poverty and poor home energy standards in parliamentary constituencies to MPs; and
2. aid delivery agents to better target their energy efficiency programmes by providing constituency and other small area data.

The value of the tool could be increased by including options to filter the information by other indicators, for example tenure, low income and fuel poverty.

This project was split into two stages:

Stage 1 (November 2014–April 2015): Evaluate the data options for producing an energy efficiency database showing energy efficiency of domestic buildings at parliamentary constituency area (and other geographical areas) for mainland Great Britain.

Stage 2 (subject to the findings of stage 1): Construct the energy efficiency database identified as the best option in stage 1 (if technically feasible at reasonable cost) and develop an online mapping tool.

This report originally intended to only present the results and recommendations of Stage 1. However, further progress was made during Stage 1 of the research than originally envisaged; during investigations carried out for the research, DECC provided a detailed EPC dataset that was derived

from its National Energy Efficiency Data framework (NEED). However, this only includes data up to 2012 and is only representative of EPC data[1].

We therefore continued to investigate options for providing better quality data that is more up to date and representative. Nevertheless, the NEED-derived data still gives a useful indication of the variation in EPC standards across parliamentary constituencies. We have therefore mapped the data and provided an excel spreadsheet of parliamentary constituency data to accompany this report.

In total, six options were evaluated as part of the feasibility work in stage 1:

- **Option 1:** Purchase all Great Britain EPC records and use this data as a representative of all dwellings in Great Britain.
- **Option 2:** Purchase Experian household and housing data for Great Britain and model energy efficiency ratings using a housing energy model.
- **Option 3:** Purchase EPC data and Experian data and 'gap-fill' missing EPC records by using housing characteristics available in Experian.
- **Option 4:** Predict energy efficiency ratings by applying a statistical prediction model based on housing condition survey data (for example English Housing Survey (EHS) or Scottish Housing Condition Survey (SHCS) to Census data at Census Output Area (COA) level.
- **Option 5:** Use the National Energy Efficiency Data-framework (NEED) and the energy efficiency data contained within.
- **Option 6:** Made a bespoke request to the DCLG to produce the EPC data at parliamentary constituency area.

The full feasibility report may also help other organisations construct their own databases for smaller geographical areas. Furthermore, organisations holding non-publicly available data or with resources to construct the more expensive options evaluated are likely to find the report of interest.

1.1 Methodology

The six options were evaluated against several key criteria and several secondary criteria, shown in Table 1.1. Although options were evaluated against cost they were not ruled out on this basis alone; other organisations may have sufficient resources to fund these options.

1.2 Data option evaluation: Results and recommendations

A range of options including the direct use of energy efficiency data, options to model energy efficiency ratings and a combination of the two were evaluated. The likely accuracy of the different approaches was examined using exploratory analysis and statistical modelling. Table 1.2 provides a summary of the evaluation results.

Table 1.1: Criteria for evaluating options

[1] The data provided by DECC is for EPC data only and is not representative of the entire housing stock, only those dwellings which have received an EPC rating since 2008.

Criteria	Description	Considerations
Key criteria		
Coverage	Parliamentary constituencies in England, Scotland and Wales	Data set coverage Data set completeness
Accuracy	Representative of actual housing	Comparison against EPC ratings (considered as highest accuracy) (see section **3.2**)
Cost (costs with significant contribution and difference between options)	Purchasing and data acquisition costs Staff and resource costs for major data processing	Received quotations or where cost per record, number of records Estimation of staff time, assuming high level of staff expertise
Secondary criteria		
Technical requirements	Data-handling requirements	Summary of software, skills, technical ability and statistical modelling knowledge required to create database
Data licensing	Purchasing of commercial or personal data can have data licensing issues	Summary of restrictions and conditions
Updating the data set	Longevity of approach depends on ease and expense of annual updating	Summary of difficulty and cost

Following an investigation of the NEED option (**Option 5**), a request was made to the Department of Energy and Climate Change (DECC) for detailed EPC data. DECC responded very positively by providing a data set of the numbers of properties in each EPC band in each Lower Level Super Output Area (LSOA) in England and Wales. This was derived from a version of the NEED EUL file, itself produced by matching 4 million representative EPC records collected between 2008 and 2012 to the main NEED data framework (owned by DECC). Each record was weighted so that the data set represented all dwellings with EPCs and then EPC data was aggregated to LSOA level. While data at LSOA level should be treated with caution, aggregations to higher levels of geography, for example parliamentary constituency or government office region (as shown in Figure 1.1), should provide a valid representation of energy efficiency standards based on existing EPC records. We have therefore mapped this data, by parliamentary constituency, across England and Wales and provided an accompanying database. However, it is important to note the limitations of this data: it is based only on EPC records collected up to 2012, and is therefore not representative of the entire housing stock.

The evaluation also found that three other options are worth considering, given that they are likely to produce a more accurate and comprehensive data set than **Option 5**. The options listed below would use more recently collected EPC data and allow for coverage of the whole of Great Britain (that is, including Scotland):

- **Option 1:** Purchase EPC data and assuming these are representative of all dwellings in Great Britain.
- **Option 3:** Purchase EPC and Experian data and 'gap-fill missing EPC records by using housing characteristics available in Experian.

- **Option 6:** A bespoke request to the DCLG to make the EPC data available for free under a researcher licence or to produce aggregated EPC data at parliamentary constituency area.

Decisions on whether to proceed with these options are dependent on the funding available. On all other accounts they remain workable and realistic solutions, and all three involve the use of EPC data in one form or another. Furthermore, if the DCLG agreed to provide the EPC data free of charge under a researcher licence then this would substantially reduce the costs of both **Option 1** and **Option 3**.

Option 1 would use actual EPC records from the Energy Performance of Buildings Register (EBR). Assuming that the data needs to be purchased, the main limitations with this approach would be:

a) the coverage of data, which is approximately 50 per cent of all dwellings in Great Britain, and the resulting uncertainties regarding the ability for these records to represent all dwellings
b) the costs of purchasing the data, which would be in excess of £123,000 for Great Britain.

Table 1.2: Summary evaluation of the different data options

Data option	Cost of data	Accuracy rating	Coverage of all GB addresses	Technical requirements	Overall evaluation comments
Option 1: EPC data	**High** >£120,000	**High**	**~50%** (accurate determination requires access and analysis of the complete data set)	**Moderate**	✓ Uses results from address level energy assessments. Although coverage issues, the data set would reflect all current EPC data captured.
Option 2: Experian data	**Medium** £32,000–£45,000	**Low–medium**	**100%**	**High**	✗ Accuracy issues
Option 3: EPC plus Experian	**Very high** £153,000+	**High**	**100%**	**High**	✓ Likely to be the most accurate and has full address coverage, but high costs
Option 4: Census data and EHS	**Free**	**Low**	Uses information compiled from every address in the country, presented at COA	**High**	✗ Prediction models using Census data unlikely to accurately calculate energy efficiency rating
Option 5a: NEED End User Licence (EUL) file	**Free**	**n/a**	**4m** records weighted to represent all of England and Wales	**Moderate**	✗ No geographical information below English region or devolved nation
Option 5b: Derived NEED data supplied by DECC	**Free**	**Medium–high**	**97%** of addresses across all but 12 LSOAs in England and Wales	**Low**	✓ Data is free, with good coverage in England and Wales, and based on EPC data in NEED, but only up to 2012
Option 6: FOI request to DCLG/ EST Scotland	**Low** (Admin costs only)	**High**	**~50%**	**Low**	✓ If successful could produce the same or similar information as **Option 1**

Nevertheless, this is an attractive option as it uses data collected at address level, reflecting measures actually installed in each dwelling, and the subsequent dwelling level EPC rating. Furthermore, geographically representing the coverage of EPCs would represent a useful subset of information on a map, highlighting areas that have experienced low levels of EPC coverage.

Option 3 would increase the coverage of EPC to all addresses by purchasing address level data from Experian. This would be used to match addresses with EPC records to those without records by similar property characteristics – a version of this method is used by the Home Analytics tool in Scotland (see Section 4.8.2). While this is likely to increase the accuracy and coverage of the data, it comes at an increased cost and it is difficult to predict the increased level of accuracy provided.

If access to EPC data under a researcher licence is not possible (**Option 6**), it may be possible to access aggregated data from the EPC register data via a Freedom of Information (FOI) request. This would substantially reduce the costs (there are likely to be administrations costs to cover the production of the statistics) as well as bypass some of the data processing required to aggregate the data to the desired geography. However, there is no certainty that this attempt would be successful.

Finally, one method of reducing the costs could be to scale down the geographical coverage of any resulting maps, for example mapping only one English region or a devolved nation of Great Britain. Further details of this approach are given in the full report.

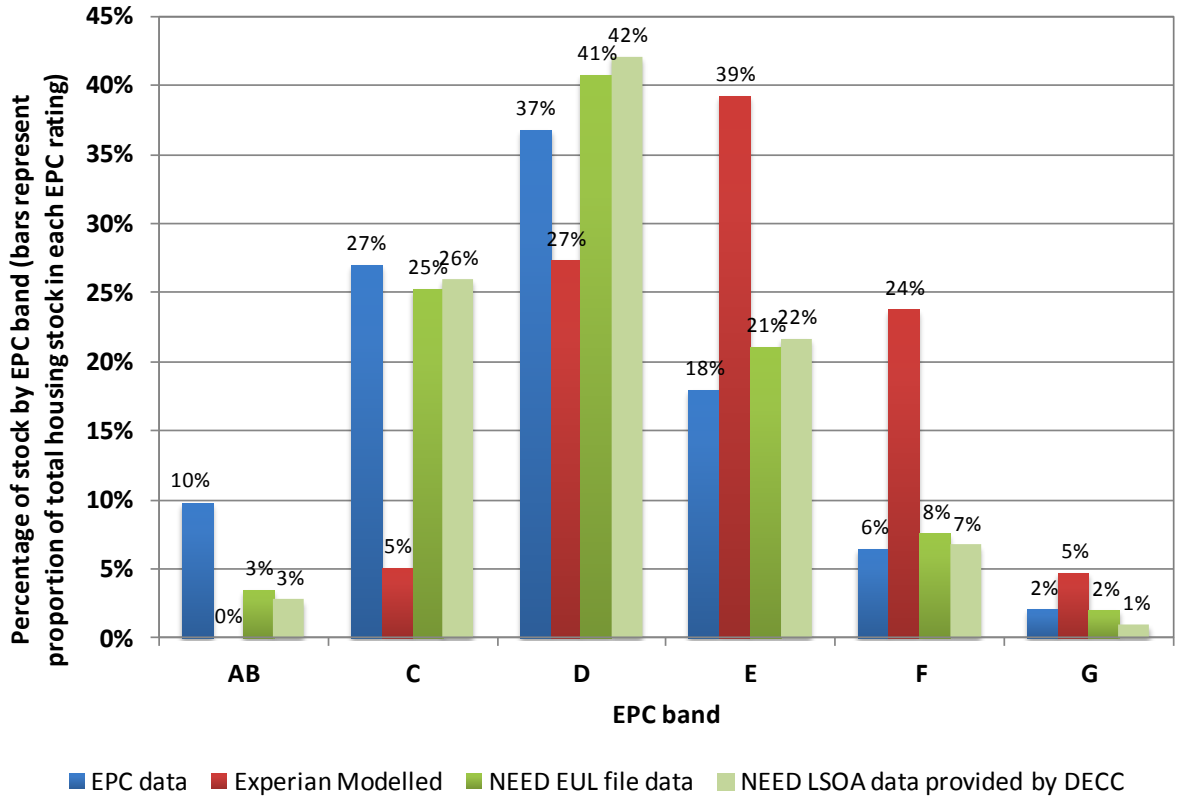


Figure 1.1: Comparison of EPC data for the South West of England from four data sets: 1) EPC register, 2) Experian modelled energy efficiency rating, 3) the NEED EUL file and 4) EPC data provided by DECC at LSOA level from the NEED EUL file

Two options were ruled out due to accuracy concerns and data access issues. Modelling energy efficiency from Experian data (**Option 2**) can provide address level energy efficiency values, from which an aggregated energy efficiency data set could be produced. However, comparing the modelled results with EPC data at different geographical levels (address level to regional totals) reveals a significant variety in the accuracy of these modelled values. This variety in accuracy is sufficiently large and unpredictable to offer any confidence in pursuing such an option, as illustrated in Figure 1.1.

Modelling energy efficiency using Census 2011 data at COA (**Option 4**) also yields unsatisfactory prediction accuracy of energy efficiency rating. For more details see Section 4.5.

1.3 Conclusion

The research found that it was possible to produce a reasonably accurate EPC database at parliamentary constituency level at modest cost through use of DECC's new NEED database, from which maps could be produced. The research went on to compile this database and constituency maps for England and Wales are included in the full report.

This option has some limitations, such as the data being slightly out of date. A more accurate database could be produced through purchase of the EPC register and the use of Experian data to model ratings for homes without EPC values. However, this option entails considerable expense. This cost would be reduced considerably should the DCLG agree to make the data available free under a researcher licence. This is already effectively the case in Scotland.

2 Introduction

Statistics and data on fuel poverty and the energy efficiency standards of the UK housing stock are often referenced in general terms (for example at national level) and used to underpin the strategic case for home energy efficiency policies and programmes. However, at regional, sub-regional and small area levels data is much less accessible. Even when data is published at regional or sub-regional level (for example maps of fuel poverty and energy efficiency uptake at local authority area) the format does not allow for detailed analysis of the underlying data and is not presented in a way that can be easily understood by members of the public, media or those outside the energy sector. Politicians may understand the issues at a strategic level, but have limited means of assessing the extent of the problem in their constituency compared with other areas.

EPCs are now embedded within the housing sector with housing professionals and the public is increasingly aware of their value in rating home energy performance. EPC bands are also now used routinely to underpin national energy efficiency policies and proposals for minimum housing standards and targets. Data on EPC bands is already published by the DCLG and the Scottish government at aggregate local authority level. However, this does not correspond directly to parliamentary constituencies. Additionally, it does not go down to a sufficiently disaggregated level to be of practical use to local residents, managing agents or politicians.

As such, making data more readily available on energy efficiency levels of housing in a particular local area would provide an instant comparator of housing in that area against national targets and policies. It would enable those concerned with energy efficiency and fuel poverty to highlight the issue of poor energy efficiency standards to MPs. It would also provide a valuable resource to those wanting to better target local energy efficiency and fuel poverty programmes.

Citizens Advice therefore commissioned the CSE to investigate options for producing energy efficiency statistics at small area level, primarily parliamentary constituency. The ultimate vision for this project is to produce a web-based interactive mapping and data tool that would enable energy efficiency and fuel poverty organisations to quickly and clearly identify areas containing the most inefficient housing. A key priority of the project is to highlight this information by parliamentary constituency to highlight the problems of fuel poverty and poor home energy standards to politicians and others concerned with tackling fuel poverty. The project also aims to provide data at other local geographies to facilitate the targeting of energy efficiency programmes.

In addition to the core focus of mapping energy efficiency data, the value of the tool could be increased by including capability to filter the information by other indicators. In particular, tenure of property (owner occupied, private rented and social rented status), low income status and vulnerability of households would be particularly helpful. For example, using this information it would be possible to identify areas likely to contain a high proportion of fuel poor households under the LIHC definition in England. A secondary consideration to the study is therefore to consider this possibility to include additional data and indicators to the core data required.

2.1 Data download options

As well as presenting the data in map form, the small area EPCs data – for example at LSOA – could also be made available for download from the web tool for use in conjunction with a range of other

socio-demographic data. For example the EPC data could be used alongside additional information from the 2011 Census – or other small area data sources such as the Index of Multiple Deprivation (IMD) – to target specific demographic groups which are living in the least efficient homes. Census data provides information on the numbers of older people or families with children under five years of age by COA, which can be aggregated to LSOA level. This could be used with the EPC data outputs to identify areas likely to have the most inefficient housing containing high numbers of households with older people or young children. (For information on Census geography area see Section 3.3)

However, it is important to clarify the aims of this project with regard to making data available and any issues regarding personal data protection and the Data Protection Act (DPA). In line with the previous position Consumer Focus has taken to privacy issues and publication of EPC data[2], and statements by the DCLG supported by the Information Commissioners Office (ICO), the project did not explore the release of EPC data at address level since this is categorised by the ICO as personal data. However, the ICO may consider it acceptable to provide anonymised aggregated data at COA level; CSE will seek guidance to clarify this. An anonymised aggregated data set available for smaller geographical areas than currently published would be useful to many organisations seeking to run energy efficiency schemes and improve the condition of housing while still protecting the privacy of individuals and households.

2.2 Project overview

The project was split into two stages:

Stage 1: Evaluate the data options for producing an energy efficiency database showing energy efficiency of domestic buildings at parliamentary constituency area (and other geographical areas).

Stage 2 (subject to the findings of stage 1): Construct the energy efficiency database identified as the best option in stage 1 (if technically feasible at reasonable cost) and develop an online mapping tool.

This report originally intended to only present the results and recommendations of Stage 1. However, further progress was made during Stage 1 of the research than originally envisaged; during investigations carried out for the research, DECC provided a detailed EPC dataset that was derived from its National Energy Efficiency Data framework (NEED). This only includes data up to 2012 and is only representative of EPC data[3].

We therefore continued to investigate options for providing better quality data that is more up to date and representative. However, the NEED-derived data still gives a useful indication of the variation in EPC standards across parliamentary constituencies. We have therefore mapped the data and provided an excel spreadsheet of parliamentary constituency data to accompany this report.

Below is a list of the possible options that were evaluated as part of the feasibility work in stage 1. For each option, the implications in terms of cost, accuracy and geographical coverage were considered:

[2] www.consumerfocus.org.uk/files/2009/06/Consumer-Focus-response-public-use-of-EPC-data.pdf
[3] The data provided by DECC is for EPC data only and is not representative of the entire housing stock, only those dwellings which have received an EPC rating since 2008.

- **Option 1:** Purchasing all GB EPC records and using this data as a representative of all dwellings in Great Britain.

- **Option 2:** Purchasing Experian household for Great Britain and housing data and modelling energy efficiency rating using a housing energy model.

- **Option 3:** Purchasing EPC data and Experian data and 'gap-filling' missing EPC records by using housing characteristics available in Experian.

- **Option 4:** The use of Census data on housing at COA and subsequent prediction of energy efficiency ratings of properties using a statistical prediction model based on housing condition survey data (for example the EHS or SHCS).

- **Option 5:** The use of the NEED framework and the energy efficiency data contained within.

- **Option 6:** A bespoke request to the DCLG to produce the EPC data at parliamentary constituency area.

This report has been written in a clear and transparent manner so that can provide a useful resource for other researchers in its own right. This resource could help other organisations construct their own databases for smaller geographical areas, particularly those using specifically held non-publicly available data or with resources to construct the more expensive options evaluated.

3 Methodology

The specific methodology used to evaluate the option varies for each type of data considered and is summarised in each respective evaluation in Section 4. However, for all options several key evaluation criteria were considered. These are described in detail below.

3.1 Options

3.1.1 Coverage

The geographical focus of the required tool is mainland Great Britain, covering all parliamentary constituencies in England, Scotland and Wales. The evaluation will identify if any of the data options considered are unable to represent this entire area. Alternatively, some address level sources will not have data for all dwellings, and so any solution would therefore be based on an incomplete data set, in terms of addresses covered. These data sets were investigated to determine whether there could be a skew or bias in the geographical representation of energy efficiency statistics.

3.1.2 Accuracy

Any data that is mapped needs to accurately reflect housing on the ground. The purpose is primarily to provide information on the constituencies with the most inefficient housing in the country, but the tool would ultimately be useful to a wide range of practitioners, including those targeting energy efficiency schemes. Therefore, in evaluating each approach the ability of the data to accurately represent the reality on the ground was considered. Although there are issues regarding the methodology and measurement of EPC ratings (detailed below in Section 3.2), these remain the best indicator of energy efficiency (or inefficiency) and will therefore be considered to provide the best level of accuracy. Other options evaluated here have been compared with known EPC ratings where possible.

3.1.3 Cost

The costs included and discussed here cover two aspects of building a database:

1. **Purchasing and data acquisition costs**: Data purchase costs using direct quotations from commercial organisations or based on number of records where data is priced per record.
2. **Staff and resource costs for major data processing tasks**: These are estimations of staff time requirements and a high level of staff expertise and are intended as indicatory only.

However, the evaluation of each option does not consider the costs of producing an aggregated small area stats database. These will fundamentally be the same for all options and represent a relatively minor cost in the wider scheme of database production.

While the costs of some options are prohibitively expensive for this project, other organisations may have access to sufficient resources to fund these options. No option was therefore specifically ruled out on cost grounds alone, as this is potentially the easiest barrier to overcome.

3.1.4 Other criteria

There are three other additional criteria that are important to consider when evaluating the data options:

- **Technical requirements**: Handling large amounts of data (many millions of records) will require specialist data management software; any required skills, technical ability and statistical modelling knowledge of those undertaking the proposed options was taken into account.
- **Data licensing issues:** Purchasing of commercial or personal data is likely to come with some restrictions and conditions, which are summarised where relevant.
- **Updating the data set:** The longevity of each approach will depend on the ease and expense of annually updating data used to represent energy efficiency. This is summarised and costed where possible.

3.2 Caveats associated with the use of EPC ratings

CSE is aware of both the limitations of EPC ratings as a measurement of overall energy efficiency and the variation in quality of their measurement, which is based on the Building Research Establishment Domestic Energy Model (BREDEM). These include the following:

- The model assumes that occupants heat their houses to 21°C (living rooms) and 18°C (other rooms). However, many households are likely to heat their homes to different temperatures and so the energy consumption can be overestimated (and in some cases underestimated).
- Similarly, assumptions for the electrical and energy demand and hot water requirements are made using simplified equations relating to the number of people in a household. This also is unlikely to give an accurate reflection of water use and electrical consumption from appliances.
- Rooms are assumed to be heated evenly and there is no allowance for unique heating behaviours such as heating rooms used during occupied hours.
- It is difficult to accurately model air infiltration and ventilation heat loss paths. Air infiltration around windows and doors is input as a percentage. Ventilation openings often differ in design.
- There are limitations with the modelled u-values[4] used in the calculation. Furthermore, u-values are never calculated across one fabric material as assumed by the algorithm.
- BREDEM is also less useful for 'alternative' methods of construction, such as system builds, non-traditional construction types and older housing such as cob walled dwellings.

Nevertheless, EPCs remain a useful indicator for understanding the efficiency of a given property or group of properties and they remain a key focus for a number of national policies and targets, such as minimum energy efficiency ratings of social homes in Scotland and the government's new fuel poverty targets for England. The EPC methodology is also used as part of the current Green Deal assessment procedure. Thus, for the purposes of this project, EPC ratings remain the best indicator for highlighting the extent to which local housing stocks meet national standards, and emphasising the level of improvement required in specific areas.

Nevertheless, any publicly available EPC summary data and web tools that are supported by EPC data would need to have visible and clearly worded caveats associated with them.

[4] A 'u value' is a measure of a heat loss, typically from a wall. The lower the u value, the better the insulation provided by the material used in construction.

3.3 Office for National Statistics (ONS) geographic areas

A summary of the main geographical areas discussed in this report is provided below. The two types of areas considered are the Census areas more traditionally associated with statistical outputs, summarised in Table 3.1, and the electoral/administrative areas, with details provided in Table 3.2.

For Census area geography, Middle Level Super Output Areas (MSOAs) comprise LSOAs, which in turn comprise COAs. For electoral boundaries, electoral wards are used to designate parliamentary constituency boundaries which are typically set to contain approximately 70,000 eligible voters. Typically, local authorities (in the form of metropolitan districts, non-metropolitan districts, unitary authorities and London boroughs) also comprise electoral wards. However, local authority and parliamentary constituency boundaries are rarely contiguous and vary significantly in terms of number of dwellings and population.

With regard to the aggregation of data, it is important to note that not all data in certain geographical areas can be aggregated to other areas. While having an address level database will allow aggregation of data to the majority of the different areas specified here using ONS lookups, it is not always as straightforward to convert from Census geography to administrative/electoral geography if postcode level data is not available.

Table 3.1: Census geography areas in the UK

Area type	Population	Number of households	Total number in Great Britain
COA	100–625	40–250	England: 171,372 Wales: 10,036 Scotland: 46,351
LSOA/Data zone (Scotland)	1,000–3,000	400–1,200	England: 32,844 Wales: 1,909 Scotland: 6,505
MSOA/Intermediate zone (Scotland)	5,000–15,000	2,000–6,000	England: 6,791 Wales: 410 Scotland: 1,235

Table 3.2: Electoral and administrative areas in Great Britain (2011)

Area type	Average population		Total number in Great Britain
Electoral ward	~5,500	8,941	England: 7,707 Wales: 881 Scotland: 353
Westminster parliamentary constituency	~70,000 electors (90,000 total population)	632	England: 533 Wales: 40 Scotland: 59
Local authority[5]	• Average: ~150,000 (Range: 2,200–980,000)	380	England: 326 Wales: 22 Scotland: 32

[5] These comprise metropolitan districts, non-metropolitan districts, unitary authorities and London boroughs (including the City of London Corporation).

It is also important to note that while the DCLG currently produces live tables on EPC statistics by local authority, the variation in surface area, population and number of dwellings is significant. Therefore, presenting data and mapping at local authority level makes comparisons with different areas more difficult and less meaningful.

4 Evaluation of data options

This section presents the results of a detailed evaluation of each data option listed in Section 1. Where relevant and useful some statistical and modelled analysis is provided to illustrate the merits or shortcomings of each option. In presenting the evaluation, each approach is considered and described in terms of:

- the data and its proposed use
- the costs of purchasing and processing the data
- considerations of this approach, including accuracy, coverage and risk.

However, these criteria are not applicable for the FOI request option, which warrants a separate style of evaluation.

A brief summary of the evaluation is given below, including a quick comparison matrix of all the options. A detailed evaluation of each option is given in the following sub-sections.

4.1 Headline evaluation results

Table 4.1 provides the headline evaluation results distilled from the findings in the following section. Overall, three options were ruled out as a result of accuracy concerns and data access issues. The accuracy issues concerning the use of Experian data are illustrated in Figure 4.1, which shows proportions of dwellings by EPC band from EPC data, NEED data and Experian modelled data in the South West of England.

Three options remain workable and realistic, but all involve the use of EPC data in one form or another. A discussion on these overall conclusions can be found in the final section of the report. Also included in the recommendations section is a presentation of an alternative option for a smaller area of the country that could satisfy any accuracy concerns while at the same time not be prohibitively expensive.

Table 4.1: Summary evaluation matrix of all options

Data option	Cost of data	Additional activities and costs	Technical requirements	Accuracy rating	Coverage of all GB addresses	Overall evaluation comments
1. EPC data	**England and Wales:** 0.01 per record; £120,000 **Scotland:** Unknown (but likely to be £1,000s)	Data processing and cleaning (removing duplicates) Est.: £4,000–£5,000	**Moderate**	**High**	**~50%** (accurate determination of coverage requires access and analysis of the complete data set)	✓ Uses results from address level energy assessments. Although coverage issues, the data set would reflect all current EPC data captured.
2. Experian	£32,000–£45,000	Data processing/cleaning; CHAID modelling; energy efficiency modelling: £10,000–£20,000 *(Optional: SPSS licence)*	**High**	**Low–medium**	**100%**	✗ Accuracy issues (see Figure 4.1)
3. EPC plus Experian	**England and Wales:** £153,000–£163,000 **Great Britain:** >£160,000	Data processing/cleaning; Energy efficiency data matching: £7,000–£12,000 *(Optional: SPSS licence)*	**High**	**High**	**100%**	✓ Likely to be the most accurate and has full address coverage, but very high costs
4. Census data and EHS	Free	Data accessing; Dev and run energy efficiency prediction model: £8,000–£16,000 *(Optional: SPSS licence)*	**High**	**Low:** prediction model inaccurate; self-reported; more than four years out of date	Uses information compiled from every address in the country, presented at COA	✗ Prediction models using Census data unlikely to accurately calculate energy efficiency rating
5. NEED: End User License (EUL) file	Free	**n/a** *(Optional: SPSS licence)*	**Moderate**	**n/a**	**4m** records weighted to represent all of England and Wales	✗ No geographical information below English region or devolved nation, significant data restrictions on accessing full NEED
5b. Derived NEED data supplied by DECC	Free	None	**Low**	**Medium–high**	**97%** of addresses across all but 12 LSOAs in England and Wales	✓ Data is free, with good coverage in England and Wales, and based on EPC data in NEED, but only up to 2012
6. FOI request to DCLG and EST Scotland	Likely to entail cost to cover staff time to produce data	Unknown until request made, but could include external IT consultancy fees for data production	**Low**	**High**	**~50%**	✓ If successful, could produce the same or similar information as **Option 1**

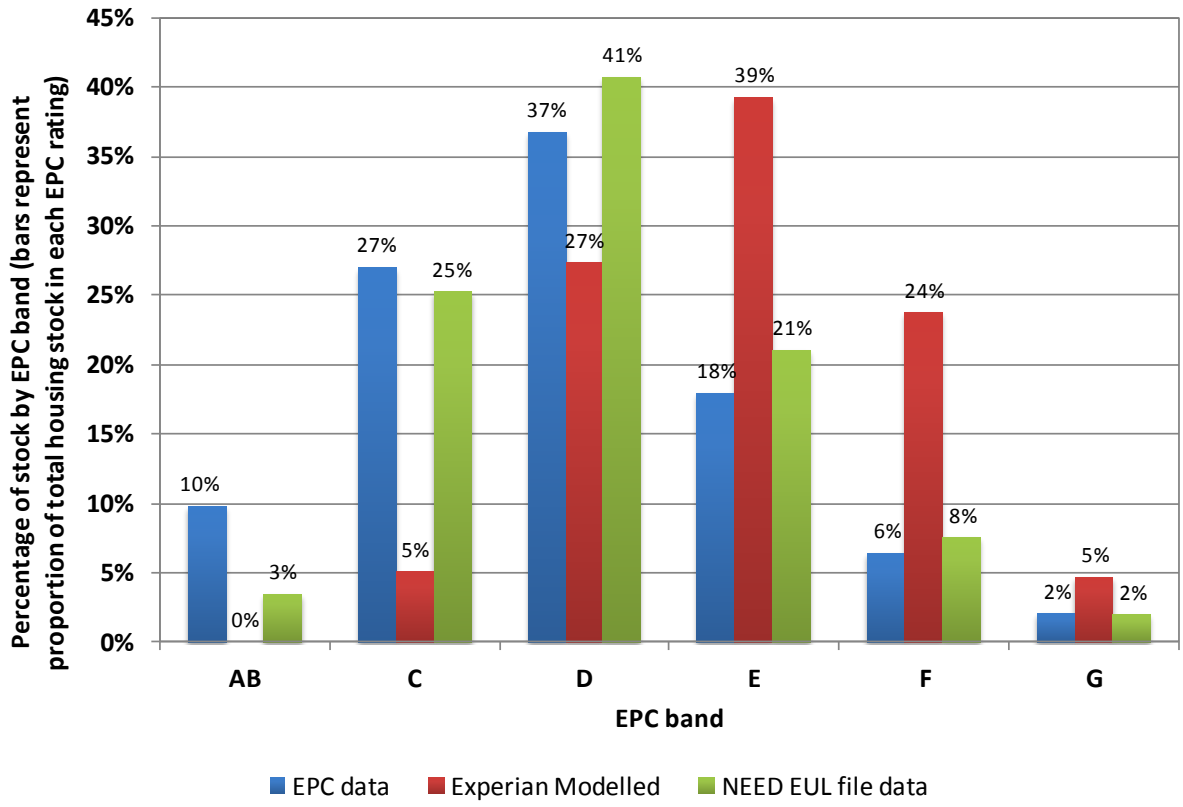


Figure 4.1: A comparison with EPC data for the South West of England contained in three separate data sets: 1) EPC register, 2) Experian modelled energy efficiency rating and 3) aggregated EPC bands in the NEED (NEED data contains some EPC data)

Figure 4.1 shows a comparison of the proportion of dwellings with ratings in each of the energy efficiency bands from three of the main data sets considered in this review. Overall, the EPC data describes the most efficient housing stock, although the EPC data coverage in the South West is around 45 per cent of all dwellings (see Appendix Table D). NEED data, which comprises EPC data, describes a slightly less efficient housing stock, with proportionally fewer A, B and C rated properties and high proportions of D, E and F rated dwellings. By comparison, the modelling of Experian data alone suggests a significantly less efficient stock overall.

4.2 Direct purchase and use of EPC data

Table 4.2: Summary details of constructing an energy efficiency database using EPC data

Data issue/requirement	Comment
Data owner	**England and Wales:** Landmark on behalf of the DCLG **Scotland:** EST Scotland on behalf of the Scottish Government
Number of records in database	**England and Wales:** ~12m records **Scotland:** 1.2m records
File format of data	Excel or csv
Size of data file	3–10GB (estimated and depending on size of data package)
Direct costs	**England and Wales**: £0.01–£0.10 per record; £123,000–£1.2m **Scotland:** No direct costs for data, but undisclosed fee to cover costs of producing data from the register (likely to require external IT consultants and associated fees)
Activities involving additional costs	Data cleaning and duplicate removal
Estimated additional costs	£4,000–£5,000
Typical timescale to obtain data if known (from date of request to receiving data)	Unknown
Update requirements to maintain database accuracy	Annual update agreement available with EPC (where new EPCs have been recorded at addresses which already have previous EPC records). New records added to the register will require an additional purchase at £0.01 a record.
Ownership constraints (e.g. licensing issues, T&Cs etc.)	(See Appendix for user credentials and specified intended use)
Technical requirements	Ability to handle and manipulate a large database, including search and removal of duplicate records and test this process
Accuracy of EPC data at aggregate geography	Strong to unknown accuracy, depending on the coverage of EPC in a given area, which ranges 20–75% across local authorities in England
Smallest geography of data	Address level, but not all GB addresses have data points
Coverage of data	Approximately 50% of addresses (but level of duplication in the database unknown). Coverage in England and Scotland ranges 20–76% in local authorities.

The EPC register contains a series of EPC data that is compiled from domestic energy assessments conducted at address level by trained individuals. Data has been collected since 2008. Using EPC data provides the most accurate option for illustrating the energy efficiency of domestic buildings. A number of other energy efficiency databases (for example NEED, which is discussed in Section 4.6) or tools (for example Home Analytics tool, which is described in Section 4.8.2) are based on EPC data.

The limitations of using EPC data are predominantly the high cost of the data itself, the less than complete coverage[6] and potentially uneven distribution of the records, geographically and by

[6] On 30 October 2014 there were 12,334,904 EPCs lodged on the register in England and Wales since 2008, although some will be duplicates. Therefore, over half the homes in England and Wales would not be covered by EPC data in the register.

tenure. Additionally, EPC data does not contain some of the additional information that would be of value for this project, such as tenure. However, this could be determined from other sources and added to the database to supplement EPC data.

A summary of the details and implications of using EPC data is provided in following section.

4.2.1 Summary of EPC data and proposed use

The energy efficiency rating is a measure of the overall efficiency of a building. This rating is based on the performance of the building and its fixed services (such as heating and lighting). The higher the rating the more energy efficient the home is. The energy efficiency ratings are expressed as bands ranging from A to G, where A is the most efficient and G is the least efficient.

For an energy assessment of a building to be valid, it must be lodged on one of the EBRs by an accredited energy assessor. EPCs for dwellings are held on the Domestic Energy Performance Certificate Register. The EBR for England and Wales is operated on the DCLG's behalf by Landmark Information Group. Scotland holds a separate register; EPC data for Scotland would have to be directly purchased through Energy Saving Trust (EST) Scotland from the Scottish Energy Performance Certificate Register[7].

There are certain restrictions regarding the eligibility of organisations to purchase EPC data from the register. There is a defined list of eligible organisations and a limitation on what an eligible organisation intends to use the data for. The type of organisations that qualify for eligibility include national and local authorities, government departments, energy/environmental charities, social housing providers, education providers, energy companies and Green Deal providers (the full list of eligible organisations is provided in the Appendix). Charities must be working in the field of research concerning the Earth's climate or environment, building use or design, or the promotion of energy efficiency. The CSE, for example, would therefore be considered an eligible charity.

In addition, the intended use of data must fall into one of a number of categories, which include:

- promoting energy efficiency improvements (as defined in Section 2 of the Energy Act 2011)
- conducting research into, or developing or analysing policy (or policy proposals) in relation to the energy efficiency of buildings
- conducting research into the effectiveness or impact of energy efficiency improvements
- identifying geographic areas where the energy efficiency of buildings is low relative to other areas, or conducting research into the extent, causes or consequences of such lower levels of efficiency.

A map showing energy efficiency by parliamentary constituency would fit with any of the above categories.

The data itself would be provided by the DCLG in csv format (the latest Excel version does not permit files with more than 1 million rows) and could be directly uploaded to a database engine. All data

[7] https://www.scottishepcregister.org.uk

requires some examining and cleaning, and the EPC data should primarily be checked for duplication and for outdated postcodes (some postcodes have changed in recent years).

Aggregate data on the numbers of EPC records by parliamentary constituency could be produced straight from the EPC data (constituency is included as a variable in the small data pack of EPC data – see Section 4.2.2 below) and used to produce a database that would sit behind any mapping tool. However, since the EPC data has been collected since 2008, it is unclear (without directly examining the data) whether the constituency details contained in the register are those that were redrawn for the 2010 election or a mixture of pre-2010 and post-2010 constituencies. It is therefore recommended that the EPC data is joined to the Office for National Statistics Postcode Directory (ONSPD) by postcode (hence the need to clean and update all existing postcodes) to ensure that the constituencies represent the current boundaries.

4.2.2 Costs of using the EPC data

There are three data options available for purchasing England and Wales EPC data from the register[8], with the following associated fees:

1. Small data pack: £0.01 per record.
2. Medium data pack: £0.05 per record.
3. Large data pack: £0.10 per record.

For this project it is likely that the small data pack would contain sufficient information. The list of variables include in this data are the:

- address of the property (including the postcode)
- energy performance indicator
- potential energy performance indicator
- energy efficiency
- property type
- inspection date
- region
- local authority area
- constituency
- county
- date on which the data was entered onto the register.

A full list of all available variables in the EPC data set from small medium and large data packs is provided in the Appendix.

For Scotland EPC data, the pricing is different and there is no fixed unit cost per record. EST Scotland merely charges an admin fee to cover the staff time involved in accessing and providing the data to a third party. However, to date the largest area covered by an EPC data request is a local authority area. The EPC team in the Communities and Local Government Directorate in Scotland would require some external IT consultancy to extract Scotland-wide data covering all EPC lodgements, which

[8] Accessing register data under the Energy Performance of Buildings (England and Wales) regulations 2012, Department for Communities and Local Government, April 2014.

would incur additional and higher fees. As this has apparently never been done before, the total cost of doing this was not available from the EPC team.

For the purposes of this work, the costs of purchasing data for England and Wales are likely to represent the bulk of the overall costs. These are presented below in Table 4.3. The total cost of purchasing a small pack of data covering all EPC records in England would amount to approximately £117,000, with Wales adding a further £6,000, a total of around £123,000.

An alternative option could be to make a specific request to the DCLG for free access to EPC data under a researcher licence agreement, with the intention of using the data to produce statistics of interest while not publicly releasing any address level EPC data itself. This is considered and discussed further in Section 0. If this was successful, the purchasing costs specified here would be reduced to zero and the only cost associated with this option would be the processing cost. However, it is not yet clear whether or not this is a likely outcome (to date, requests of this nature to the DCLG have not been successful).

Table 4.3: Costs of purchasing EPC data for English regions and Wales (small data pack)

Region	Number of lodgements	Estimated total dwelling stock	Percentage EPC coverage	Cost of approximate purchase costs
London	1,859,535	3,358,180	55.4%	£18,600
South East	1,859,300	3,694,420	50.3%	£18,600
North West	1,514,946	3,143,930	48.2%	£15,100
East of England	1,217,701	2,531,910	48.1%	£12,200
Yorkshire and The Humber	1,140,401	2,319,920	49.2%	£11,400
West Midlands	1,112,940	2,376,730	46.8%	£11,100
South West	1,090,102	2,401,300	45.4%	£10,900
East Midlands	989,878	1,971,520	50.2%	£9,900
North East	506,872	1,178,280	43.0%	£5,100
Wales	639,030	1,394,464	45.8%	£6,400
ENGLAND	11,695,874	22,976,190	50.9%	£117,000
ENGLAND AND WALES	12,334,904	24,370,654	50.6%	£123,300

Table 4.4: Additional data tasks and staff resource required

Task	Number of records	Estimated number of staff days required[9]	Estimated staff costs based on a rate of £500/day	Task descriptions
EPC data cleaning	~13.4m	4–5	£2,000–£2,500	e.g. Removing duplicates
EPC data to ONSPD matching (postcode)	~13m	4–5	£2,000–£2,500	Joining additional area level information by postcode from ONSPD
Total		**8–10 days**	**£4,000–£5,000**	

[9] Depending on the expertise level of the staff used and the exact nature of the data processing requirements

Additional costs, although minor in comparison to purchasing costs, would be required in order for the data to be cleaned and assigned to a geographical area for aggregation. This would mostly involve the removal of duplicates to ensure only one record for each address and ensuring address details for each record are complete and correct. The latter process is required in order to perform address matching with the ONS Postcode Directory in order to allocate a parliamentary constituency identifier (or Census area) to the EPC data. These tasks and an estimation of the resource requirements are detailed in Table 4.4.

4.2.3 EPC data analysis and considerations

The DCLG and its Scottish counterpart, the Communities and Local Government Directorate, have both published some EPC statistics. Both include the number of EPC lodgements by EPC band by local authority. The England and Wales data also includes additional information such as EPC lodgements by tenure and transaction type. Using additional external data sets, we have analysed the EPC statistics to determine their coverage by local authority, rurality and tenure. These are presented and summarised below.

There has been a previous attempt to visualise this data. The Open Data Communities website provides some access to it along with a mapping tool.[10] It is possible to download data from this in csv file format and use the map to visualise different aspects of the data on a map. Unfortunately, its functionality is limited and it does not provide a summary of all EPC data, only data collected in each quarter and by one EPC band at a time. It also has no scale showing the numbers of EPC records. As a result, it does not visualise anything of practical use neither does it allow identification of the most inefficient areas of housing in the country, nor at a geographic areas level small enough to be useful to practitioners or campaigners.

Table 4.5 shows the number of lodgements by transaction type for all domestic EPC data in the register. The majority of EPCs lodged are the result of property sales (40 per cent) or rentals (33 per cent). More recently, EPCs lodged as a result of Green Deal assessments have contributed a significant proportion to the annual number of EPCs. In 2013, EPCs lodged as a result of interaction with the Green Deal were responsible for 10 per cent of the total of 2 million.

Table 4.5: Number of EPC lodgements by transaction type in England and Wales (2008–2014)[11]

Transaction type[12]	Total number of EPCs registered between January 2008 and June 2014	Percentage of all transactions
Marketed sale	4,588,378	39.7%
Rental (not specified)	3,819,721	33.1%
Not recorded	927,221	8.0%
New dwelling	790,113	6.8%
None of the former	781,219	6.8%
Assessment for Green Deal	301,035	2.6%
Non marketed sale	199,286	1.7%
Feed in Tariff application	119,504	1.0%
Following Green Deal	19,195	0.2%
Rental (social)	3,467	0.03%
Rental (private)	917	0.01%
Total number of domestic EPCs lodged	**11,550,056**	**100.0%**

[10] opendatacommunities.org/data/building-standards-and-sustainability/energy-efficiency-and-performance/domestic/certifcates-lodged-by-number/by-energy-efficiency

[11] Series has been collected over a different timescale to other EPC statistics, hence the total EPC lodgement figure differs from Table 4.3.

[12] Transaction types are those provided by DCLG in the EPC data statistics.

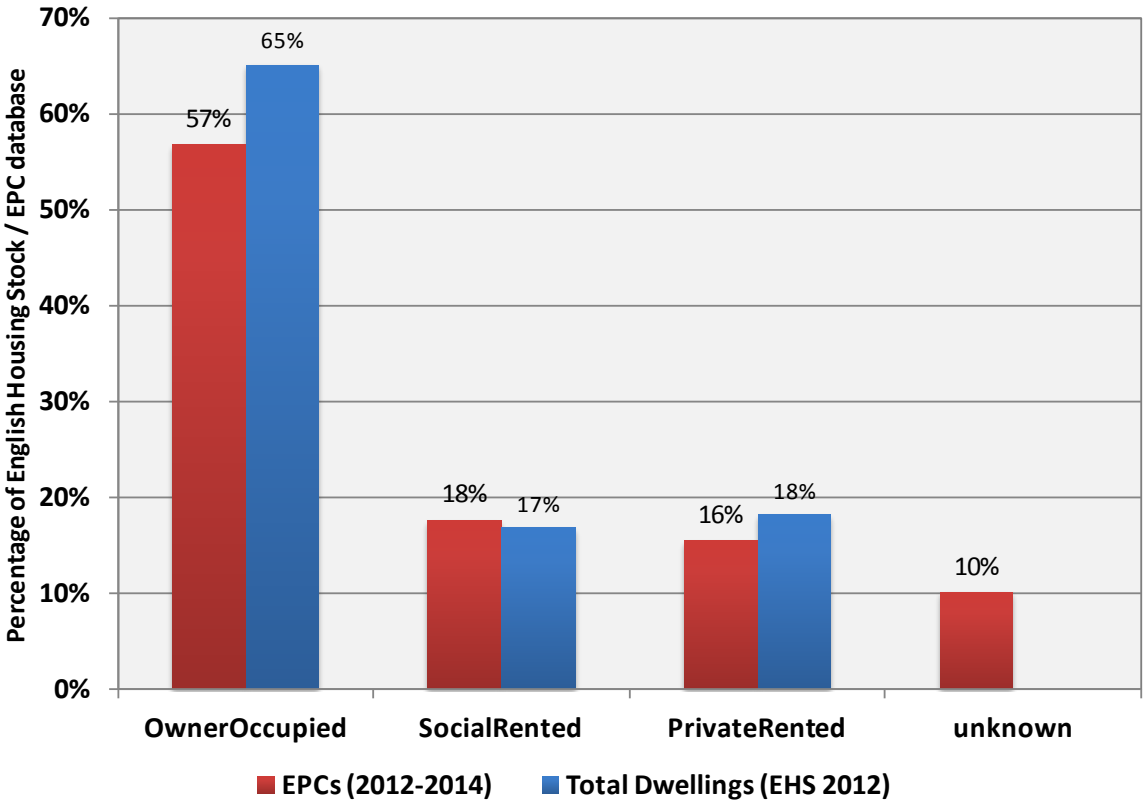


Figure 4.2: Proportion of EPC coverage by tenure compared with overall tenure coverage in England

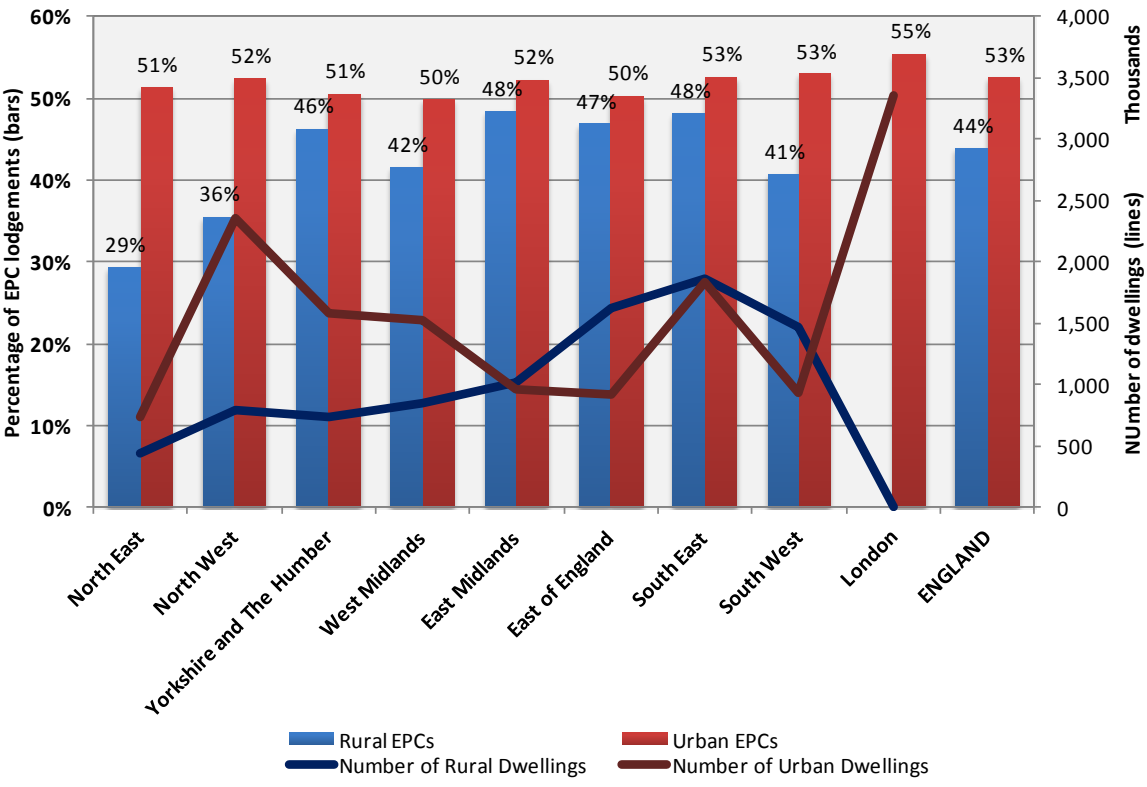


Figure 4.3: Coverage of EPCs by English region and rurality

Statistics by tenure are only currently available from 2012. These statistics are presented for England alongside tenure breakdowns for the EHS 2012 in Figure 4.2. Overall, there is a reasonable match between the proportions of EPCs taken for the different tenures and the overall tenure distribution. However, owner occupied housing is slightly under-represented. Owner occupied housing on average tends to be less efficient than social housing. Therefore aggregate EPC data could potentially slightly overestimate the energy efficiency of areas where owner occupied dwellings have not had proportional numbers of EPCs registered.

As part of the analysis, the rurality of each local authority in England (as specified by the ONS urban/rural local authority classification[13]) was used to investigate whether the coverage of EPCs differed between urban and rural areas.

Figure 4.3 shows a summary of the coverage by rural and urban local authorities in each English region. The results show that there is a distinctly higher coverage of urban dwellings than rural dwelling in most regions, and it is most pronounced in the South West, the North East and the North West. Overall in England, 44 per cent of dwellings in rural local authorities have had EPCs in contrast to 53 per cent of dwellings in urban local authorities. This suggests that, in general, rural areas are likely to be less well represented by EPC data than urban areas, although this is likely to vary across the country.

Additional analysis looked at the coverage of EPCs in each English, Welsh (not shown) and Scottish local authority to indicate the range of coverage across Great Britain at a smaller geography. The top 10 local authorities in England (in terms of percentage of all houses with EPCs) are presented in Table 4.6. Interestingly, eight of the ten local authorities are in London.[14] All those on the list are officially urban areas. The five local authorities in Scotland with the highest proportion of homes with EPC are shown in Table 4.7, and both Edinburgh and Glasgow feature in the list. Only one area is predominantly rural while the other four are urban.

Table 4.6: The 10 local authorities in England with the highest proportion of EPC coverage

Local authority	Number of lodgements	Total dwelling stock	Urban/rural classification	EPC coverage
City of London	4,179	5,510	Urban	76%
Tower Hamlets	74,746	105,380	Urban	71%
Harlow	24,743	35,720	Urban	69%
Newham	70,182	103,210	Urban	68%
Kensington and Chelsea	57,349	84,800	Urban	68%
Nottingham	87,666	131,190	Urban	67%
Salford	70,954	108,610	Urban	65%
Wandsworth	87,102	134,620	Urban	65%
Southwark	79,624	123,270	Urban	65%
Westminster	75,723	118,320	Urban	64%

[13] www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/rural-urban-definition-and-la/rural-urban-local-authority--la--classification--england-/index.html

[14] This could be a result of London having a very active housing market with a higher turnover of sales than other parts of the country.

Table 4.7: The five local authorities in Scotland with the highest proportion of EPC coverage

Local authority	Number of lodgements	Total dwelling stock	Urban/rural classification	EPC coverage
West Dunbartonshire	29,919	44,880	Urban	67%
Glasgow City	180,190	301,633	Urban	60%
Midlothian	20,479	37,503	Urban	55%
Clackmannanshire	12,549	23,894	Rural	53%
Edinburgh City	123,311	237,524	Urban	52%

The local authorities with the lowest coverage were also examined with the results provided in Table 4.8 (England) and Table 4.9 (Scotland). The lowest coverage in England is predominantly in rural county council authorities in various parts of the country. Cheshire has the lowest coverage of all local authorities in Great Britain with only 19 per cent of all dwellings having EPCs. The nine local authorities with the lowest coverage are all rural areas, and the lowest eight have less than 22 per cent of dwellings with EPCs. It is likely that a number of Lower Super Output Areas (LSOAs) in these local authorities would have no data at all.

Table 4.8: The 10 local authorities in England with the lowest proportion of EPC coverage

Local authority	Number of lodgements	Total dwelling stock	Urban/rural classification	EPC coverage
Cheshire East	31,214	166,240	Rural	19%
Cheshire West and Chester	28,845	147,570	Rural	20%
Wiltshire	40,666	201,990	Rural	20%
Shropshire	27,532	135,570	Rural	20%
Central Bedfordshire	22,446	108,690	Rural	21%
Isles of Scilly	289	1,390	Rural	21%
Northumberland	30,771	148,010	Rural	21%
Cornwall	53,988	258,880	Rural	21%
County Durham	68,512	233,490	Rural	29%
Castle Point	14,484	37,680	Urban	38%

Table 4.9: The five local authorities in Scotland with the lowest proportion of EPC coverage

Local authority	Number of lodgements	Total dwelling stock	Urban/rural classification	EPC coverage
Eilean Siar	3,288	14,490	Rural	23%
Shetland Islands	2,639	10,852	Rural	24%
Orkney Islands	3,940	10,717	Rural	37%
Dumfries and Galloway	27,259	73,555	Rural	37%
Highland	44,901	113,703	Rural	39%

In summary, EPC data has a variable coverage across Great Britain. While the coverage of different tenures of housing seems to reasonably match levels of tenure nationally, coverage in rural areas could be an issue when mapping this data. However, the analysis shows that in urban areas, and for the entire region of London, EPC coverage is typically over 50 per cent. Cost of purchasing EPC is likely to present a significant barrier for a number of organisations and projects.

The issues of cost and rural coverage of EPC data alone do not rule out the use of a mapping tool for this option. EPC data is the most accurate reflection of energy efficiency assessment measured and collected at address level held in a single data source, and for this reason should not be ruled out. However, if a map were to display EPC data at smaller areas than parliamentary constituency, it is suggested that some of these areas should display a 'not enough data available' label, for example where less than 25 per cent of dwellings had an EPC lodged on the register.

4.3 Estimation of energy efficiency using Experian data

Table 4.10: Summary details of constructing an energy efficiency database using Experian data and modelled inputs

Data issue/requirement	Comment
Data owner	Experian
Number of records in database	**England and Wales:** ~23m **Scotland:** ~3.4m
File format of data	csv
Size of data file	~5–6GB (estimated)
Direct costs	£32,000–£45,000
Activities involving additional costs	Predicting additional RDSAP variables using a CHAID model from the EHS, building a housing stock database for an energy model, modelling energy efficiency database, data matching
Estimated additional costs	£10,000–£20,000
Typical timescale to obtain data if known (from date of request to receiving data)	Within one week
Update requirements to maintain database accuracy	Renewal of licence agreement each year with the inclusion of any new addresses from new dwelling constructions
Ownership constraints (e.g. licensing issues, T&Cs etc.)	Data is provided by Experian on a non-exclusive, non-transferable licence. The data owner cannot sell, transfer, sub-license, distribute, commercially exploit or otherwise make available to any third party any Experian data.
Technical requirements	Database software able to process more than 20m records (Microsoft SQL server on a high spec PC or server)
Accuracy of EPC data at aggregate geography	Variable (poor to good but unpredictable)
Smallest geography of data	Address level
Coverage of data	All addresses in Great Britain (England, Wales and Scotland)

4.3.1 Summary of Experian data and proposed use

Experian owns household profile data at address level, which provides information about key property characteristics for every address in a given area and covers all of the UK, as well as the address details of each property. The key Experian data that would be purchased to model EPC data would be the following variables:

- Gas mains flag.
- Tenure.
- Property build year.
- Number of bedrooms.
- Residence type.
- Mosaic Public Sector (optional, but potentially useful for characterising households; further details of Mosaic are given below).

CSE has previously used Experian data successfully on a number of projects. It formed the basis of the Somerset Housing Assessment Database that was ultimately used to analyse the improvement options for seven local authorities in the Greater Somerset area.

Previously, we have used the data in conjunction with the EHS to create a Chi-square Automatic Interaction Detection (CHAID)[15] model that predicts other housing variables that are required as input data for a RDSAP[16] assessment of properties, such as wall type and heating system. (CHAID modelling selects a set of predictor variables and their interactions that optimally predict the variability in the dependent measure. The resulting CHAID model is a classification tree that allows the analyst to identify clusters of households described by similar categorical variables, typically driven by socio-demographics or the characteristics of the property.) A similar approach could be used to create an RDSAP input database from household level Experian data which could then be modelled in a Building Research Establishment Domestic Energy Model (BREDEM) based energy model capable of analysing housing stock databases and powerful enough to handle a data set representing the GB housing stock. This would produce an EPC rating for each property, which could then be banded by EPC band.

CSE would advise the use of the National Household Model (NHM) in this instance. The NHM is the DECC's new domestic energy policy modelling and analytical tool, which CSE were commissioned to build. It uses a domain-specific and highly flexible modelling language that enables analysts to create policy scenarios and explore the potential impacts on domestic energy demand (and associated bills and emissions) over time. A key component of the NHM is the 'energy calculator', which calculates energy use by fuel and energy service at household level, based on the BREDEM-8 (2001) and Standard Assessment Procedure (SAP) 2009 and 2012 algorithms. It therefore also has the capability to produce a SAP 2009 or SAP 2012 efficiency rating.

Experian data also contains information that could be used to report on additional indicators, including property type, off gas status and property tenure. Experian's Mosaic data[17] also provides data on household segmentation, which could be used to identify low income and other households likely to be more vulnerable as a result of living in inefficient properties. Below is a selection of some Mosaic categories that could be used to identify low income households:

- M56: Older people living on social housing estates with limited budgets.
- M57: Old people in flats subsisting on welfare payments.
- N66: Childless, low income tenants in high rise flats.
- O69: Vulnerable young parents needing substantial state support.

The output of the model (EPC band) could then be used in conjunction with the ONS Postcode Directory data to join parliamentary constituency information (as well as other small area geography data) to each address. This could then be used to produce an aggregated data set showing the number of modelled EPC bands in each parliamentary constituency (or other desired area).

The benefit of this approach is that every address in the country would be assigned an EPC rating, but the disadvantage is that every one of these results will have been modelled from a limited number of data points. A discussion on the accuracy of this approach is provided below in section 4.3.3.

[15] CHAID is a popular analytic technique for performing classification or segmentation analysis. It is an exploratory data analysis method used to study the relationship between a dependent variable and a set of predictor variables.
[16] Reduced Data Standard Assessment Procedure
[17] http://www.experian.co.uk/marketing-services/products/mosaic-uk.html

Box 4.1: A selection of terms and conditions for using Experian data

EXPERIAN DATA: TERMS AND CONDITIONS

Licence and use[1]

8.1 Experian grants the Client a non-exclusive, non-transferable licence to use any Experian Data and/or Experian Materials provided as part of the Services in the Territory for the Permitted Purpose on any licence terms identified in the Schedule. The licence granted under this Clause is made separately in respect of each individual element of the Experian Data and/or Experian Materials and commences on the day that each such element of the Experian Data and/or Experian Materials is first made available to the Client.

The Client agrees that it will:

8.4.1. use the Services, Experian Data and/or Experian Materials for the Permitted Purpose only and in accordance with any Documentation;

8.4.2. (subject to Clause 5) not sell, transfer, sublicense, distribute, commercially exploit or otherwise make available to, or use for the benefit of, any third party any of the Services, Experian Data and/or Experian Materials;

8.4.3. not (and will not allow any third party to) adapt, alter, modify, reverse engineer, de-compile or otherwise interfere with the Experian Data and/or Experian Materials without the prior written consent of Experian or as otherwise permitted by law;

8.4.4. Only take such copies of the Experian Data and/or Experian Materials as are reasonably required for the use of the Experian Data and/or Experian Materials in accordance with this Agreement;

Termination of contract

Upon termination of this Agreement (or the relevant elements of it):

13.3.1. The parties shall each promptly return the Confidential Information of the other party to its owner;

13.3.2. the Client shall, at Experian's request either return the Experian Materials and Experian Data to Experian or destroy such materials and, if destroyed, provide a certificate stating that such materials have been destroyed;

Experian data comes with some noteworthy terms and conditions. Obtaining Experian data involves purchasing a 12-month licence for the required data for a given area. Those who purchase a licence, defined as 'the Client' in the terms and conditions of the licence, must use the data for a specific pre-agreed purpose and are not entitled to distribute the data or make it available to any third party.

However, the Client may distribute data produced from the original Experian data provided that it is not possible to reverse engineer the published data to reveal the original Experian household data. As this project would only concern the publication of EPC data, and this itself is dependent on a significant number of inputs, it would not contravene these terms. Furthermore, the project proposes to release aggregate area data from which it would not be possible to ascertain individual

dwelling EPC data, and certainly impossible to determine the original Experian data that was used to calculate them.

A selection of clauses from the Experian terms and conditions deemed to be worth consideration is provided in Box 4.1. Of particular note is the requirement to delete Experian data on termination of the 12-month licence agreement. Should a data set be produced using Experian data and it needed updating after the licence agreement has been terminated, then an additional licence would need to be purchased. Alternatively, if this is foreseen during the initial purchase of a licence, it may be possible for the Client to agree to purchase a licence over a longer period of time and be awarded a discount as a result. This would have to be negotiated with Experian at the time and we are unable to provide any specifics to the particular discount that maybe applied. All costs stated in this section are for a 12-month licence.

4.3.2 Costs of using Experian data

The costs for a Great Britain Experian data set are shown below in Table 4.11, broken down by different nations. Increasing the size of data set from an England-wide data set to a data set that covers Wales and Scotland does not significantly increase the cost, due to the small proportional increase in numbers of households and the nature of the purchasing cost scales used by Experian. Mosaic can also be purchased at postcode level, rather than address level which would reduce the total costs.

In summary, a GB Experian household variable data set would cost £35,000, and adding Mosaic data would cost between £5,000 and £10,000 depending on whether address level or postcode level Mosaic data was purchased. For the purposes of this study, it is recommended that postcode level Mosaic data is used to determine low-income status of households.

Table 4.11: Costs of Experian household and Mosaic data at address level for Great Britain

Region covered	Option A: Household variables	Option B: Household variables plus address level Mosaic	Option C: Household variables plus postcode level Mosaic
England	£32,000	£42,000	£37,000
England and Wales	£33,000	£43,000	£38,000
Great Britain (England, Wales and Scotland)	£35,000	£45,000	£40,000

The costs presented here are likely to represent the bulk of the overall costs for producing an aggregate EPC data set at parliamentary constituency, although some significant staff resource and time will be required to conduct the following tasks:

- Access the latest EHS survey data (if not currently held).
- Produce and run a CHAID model using Experian and EHS data.

- Impute remaining variables required for a housing stock energy model and build a housing stock.[18]
- Run the model and extract the required results.
- Join modelling results, address details and ONSPD data and produce the aggregated EPC data set at parliamentary constituency level.

Overall, it is estimated that this approach would cost somewhere in the region of £10,000–£20,000 of staff resource.

4.3.3 Considerations of using Experian data

In order to examine the potential accuracy of modelled energy efficiency ratings, a selection of EPC records were accessed from urban areas in the South West, avoiding those local authorities with low coverage rates of EPC data. As discussed in Section 4.2.3, EPC coverage appears to be best in urban areas. Furthermore, CSE has previously modelled energy efficiency for all dwellings in the South West rating using an Experian based housing stock data set and a BREDEM based energy model[19].

We therefore chose postcodes which contained more than 10 addresses and had more than 50 per cent EPC record coverage on the Domestic Energy Performance Certificate Register website[20]. In total, just over 250 individual pdf documents were obtained from two urban areas. The EPC rating was noted and compared with the corresponding outputs from an Experian based data set and CSE's own modelling of energy efficiency of these properties. This was intended as a purely demonstrative and illustrative purpose and should at no point be considered a rigorous examination and comparison between the two approaches. Doing that would require access to a representative sample of EPC data, which would incur significant costs beyond those available for this feasibility study. Furthermore, obtaining a bulk amount of EPC data would likely negate the need for this particular approach, that is, modelled energy efficiency ratings from an alternative data set.

Nevertheless, a comparison of address level EPC records and modelled energy efficiency results for the data obtained is presented in Figure 4.4. It shows that for the records we examined there was some reasonable correlation between some addresses. As Table 4.12 shows, approximately 34 per cent of records in each data set were within 4 points of each other. This the error of margin allowed by different EPC ratings conducted by different assessors at any one dwelling. Conversely, 66 per cent of records are outside this margin of error. Furthermore, a number of records were significantly different, with almost 40 per cent of records having a difference of more than 10 EPC points and 12 per cent having a difference of over 20 points. However, aggregating results to higher levels of geography could reduce the margins of error.

[18] The NHM requires a total of ~100 input variables.
[19] CSE's in-house Housing Assessment Model (HAM)
[20] The conditions of the website state: "You may not create a link to this website from another website or document without Landmark Information Group's prior written consent." We did not feel that obtaining Landmark Information Group's prior written consent to reference the URL of the website here was good use of the project budget. We recommend the use of an internet search engine if you wish to navigate to the site.

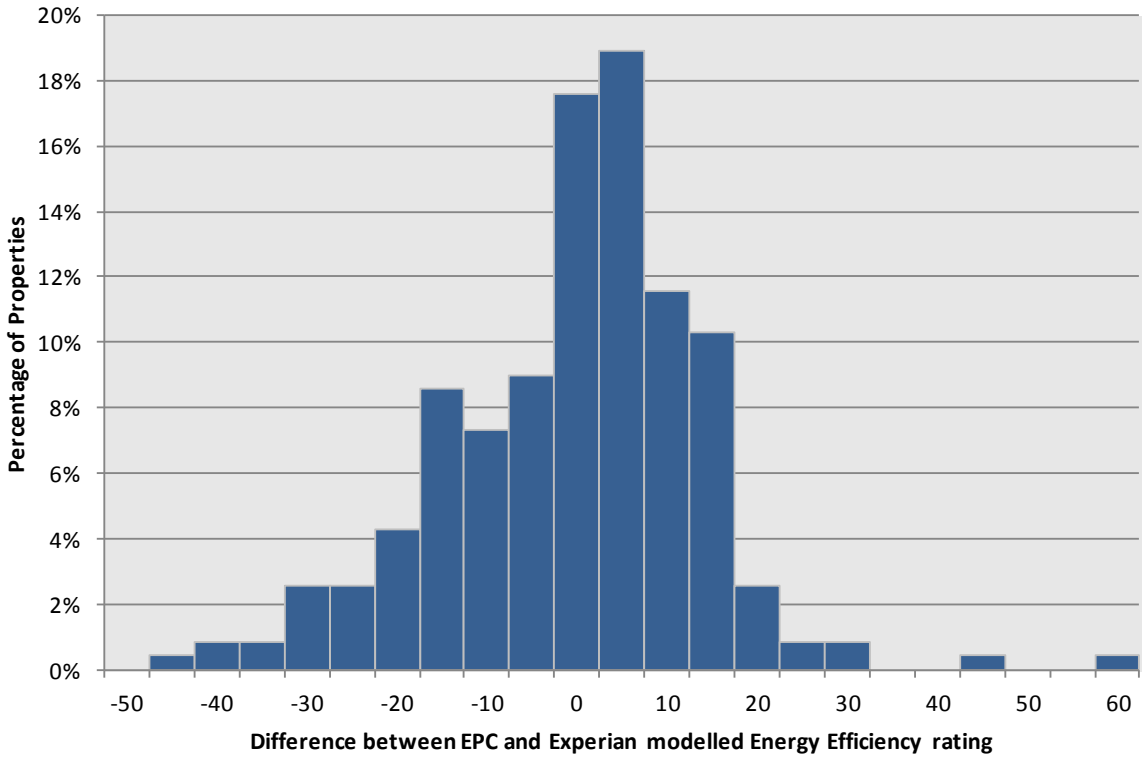


Figure 4.4: Comparison of EPC data with energy efficiency ratings calculated from modelling Experian household data in a randomly selected group of urban postcodes

Table 4.12: Difference between EPC data and Experian modelled energy efficiency

Difference between two data sets (rating values)	Proportion of records within range	Proportion of records outside range
+/-2	23%	77%
+/-4	34%	66%
+/-6	45%	55%
+/-10	61%	39%
+/-20	88%	12%

Therefore, a comparison of postcode level data was made, whereby the average EPC rating and corresponding modelled energy efficiency rating were calculated and compared for each postcode. The results are shown below in Table 4.13. Of the 14 postcodes examined, four had average EPC ratings and average modelled energy efficiency ratings within a +/-4 point range. The other eight postcode average differences ranged between -11.3 and +7.5 points. This suggests that aggregating to higher geography does remove some of the outliers in terms of difference between the two data sets, but that a significant difference remains for a number of postcodes.

Table 4.13: Comparison of average EPC ratings by postcode with average energy efficiency rating

Postcode	Average Experian rating	Average EPC rating	Difference	Count of addresses in postcode
TA1 1BU	51.9	42.5	-9.5	15
BA11 1EL	46.9	43.4	-3.6	11
BA11 1DA	54.9	43.6	-11.3	18
TA1 1DP	57.9	47.5	-10.3	11
TA1 1BD	57.3	50.7	-6.6	14
TA1 1EN	57.1	51.1	-6.0	10
TA1 1BA	49.6	52.5	2.8	11
TA1 1BB	61.7	56.0	-5.7	10
BA11 1ER	54.5	56.0	1.5	13
BA11 1HY	66.9	59.9	-7.0	27
BA11 1DE	65.2	66.7	1.5	19
BA11 1HW	66.5	74.0	7.5	14
TA1 1DA	74.6	75.2	0.6	11
TA1 1DX	74.8	77.4	2.6	22
Total	**61.2**	**58.0**	**-3.1**	**206**

Finally, a comparison at regional level was made for South West England. Figure 4.5 shows the proportion of records in the two data sets in each EPC band. The difference between the two is marked, and the Experian data set significantly underestimates the average efficiency of dwellings in comparison with EPC data.

One explanation is that the records currently in the EPC register have been conducted on properties that on average are more efficient than the housing stock as a whole. However, while it is certainly the case that urban properties – which tend to be the more efficient on average – have had a higher proportion of assessments in contrast to their rural counterparts (as shown in Section 4.2.3), it is unlikely that this would account for the difference seen here between the two data sets. Furthermore, including NEED data in this comparison (see Section 4.6) illustrates that the data obtained through an Experian based model is significantly different from other energy efficiency data on the housing stock.

One reason behind considering the use of an Experian based data set is to reduce some of the costs involved in directly purchasing EPC data. However, Experian data still incurs significant purchasing costs and requires substantial further processing and additional modelling. Although the total costs of this are expected to be less than half those for purchasing an EPC data set, the analysis here suggests that the reduction in accuracy and therefore certainty of the results is not balanced by the reduction in costs. Ultimately, a map showing the energy efficiency of housing in small geographical areas has to be an accurate reflection of the true picture. The evidence here suggests that an approach using Experian data would not guarantee this.

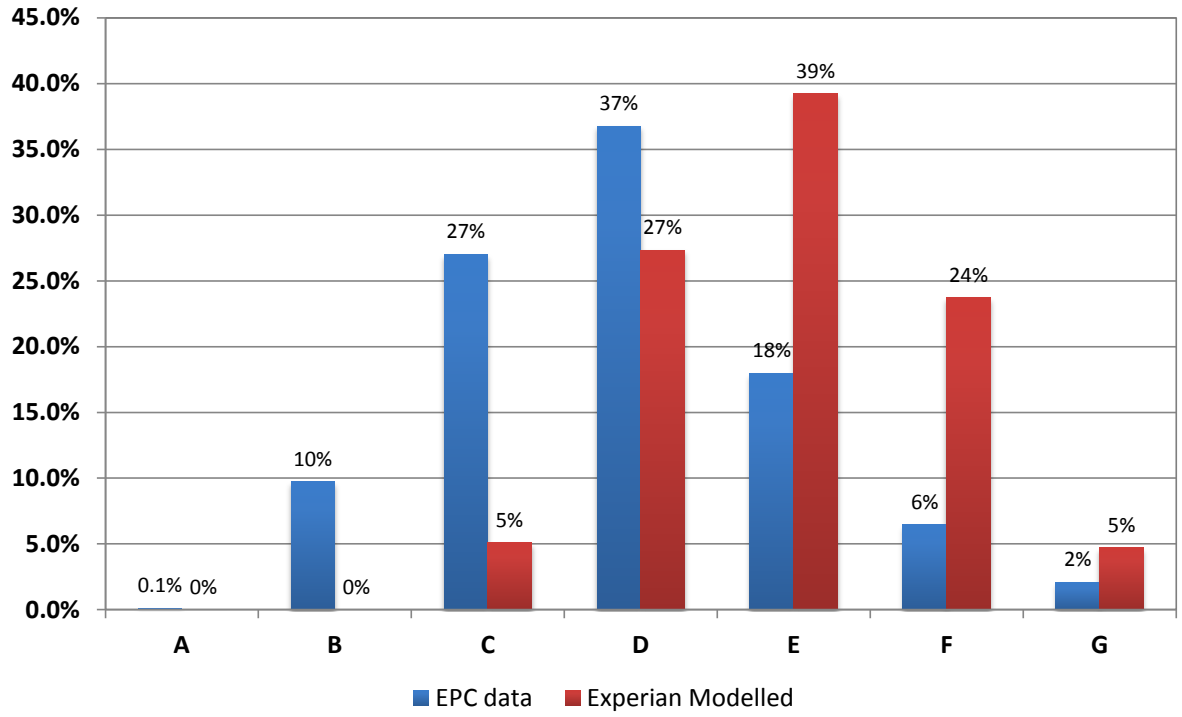


Figure 4.5: Proportion of dwellings in each EPC band for EPC data and modelled energy efficiency band in South West England

4.4 Use of EPC data combined with Experian household data

Table 4.14: Summary details of building a combined EPC and Experian data set

Data issue/requirement	EPC data	Experian
Data owner	**England and Wales:** Landmark on behalf of the DCLG **Scotland:** EST Scotland	Experian
Number of records in database	**England and Wales:** ~12m records **Scotland:** 1.2m	**England and Wales:** ~23m **Scotland:** ~3.4m
File format of data	csv	csv
Size of data file	3–10GB (estimated and depending on size of data package)	~5–6GB (estimated)
Direct costs	**England and Wales**: £0.01–£0.10 per record; £120,000–£1.2m **Scotland:** No direct costs for data, but undisclosed fee to cover costs of producing data from the register	£32,000–£45,000
Activities involving additional costs	Data cleaning and duplicate removal. Address matching to join both data sets. Completing missing EPC variables using a statistical prediction model or aggregated lookup table.	
Estimation of additional costs	£7,000–£12,000	
Typical timescale to obtain data	Unknown	Within one week
Update requirements to maintain database accuracy	Annual update agreement available with EPC (where new EPCs have been recorded at addresses which already have previous EPC records). New records added to the register will require an additional purchase at £0.01 a record.	Renewal of licence agreement each year with the inclusion of any new addresses from new dwelling constructions.
Ownership constraints (e.g. licensing issues, T&Cs etc.)	• (See Appendix A for user credentials and specified intended use)	Data is provided by Experian on a non-exclusive, non-transferable licence. The data owner cannot sell, transfer, sublicense, distribute, commercially exploit or otherwise make available to any third party any Experian data.
Technical requirements	Ability to handle and manipulate a large database, including search and removal of duplicate records and ability to test this process Database software able to process more than 20 million records	
Accuracy of data	High	
Smallest geography of data	Address level	
Coverage of data	All addresses in Great Britain (England, Wales and Scotland)	

4.4.1 Proposed usage of data

The primary aim of this approach is to produce a data set that represents every address in Great Britain. With EPC data representing around half of all addresses and with coverage appearing to range between 20 and 76 per cent in English local authorities, a map based solely on EPC data would suffer from a level of data inconsistency and be more reflective of the true situation in some areas in comparison with other areas, with rural areas likely to have the least amount of data behind them.

One way of ensuring all addresses have an efficiency rating (either an EPC or a modelled rating) would involve combining the EPC data set with the address level household data set available from Experian. This would involve some significant address matching, on the scale of that used for the National Heat Map[21], requiring some 12 million EPC records to be matched to a 26-million record Experian data set covering all addresses in Great Britain. The result would be an address level data set with all addresses having Experian data on built form, property age, number of bedrooms, current tenure and mains gas connection, and approximately half of all addresses with a corresponding EPC rating. This address matching step would represent one of the two major pieces of data processing involved in this option.

Table 4.15: Example of the contents of a lookup table that could be used to assign an average SAP rating to dwellings without EPC data based on their Experian data characteristics

Experian household variable					Average SAP rating (from records with EPC data)	Number of properties
Tenure 2011	Residence type 2011	Property age	Number of bedrooms	Mains gas flag		
Owner occupied	Detached house	pre-1919	1	No	13.85	826
				Yes	23.06	231
			2	No	14.22	4,237
				Yes	23.99	1,062
			3	No	16.12	28,740
				Yes	26.12	6,018
			4	No	17.76	21,865
				Yes	28.50	6,057
			5	No	18.51	29,676
				Yes	29.53	5,809
		1920–1944	1	No	19.18	168
				Yes	22.15	269
			2	No	15.28	807
				Yes	24.97	731
			3	No	17.07	4,104
				Yes	27.02	3,771
			4	No	28.06	2,980
				Yes	28.82	3,310
			5	No	25.54	2,855
				Yes	30.16	2,569
		1945–1965	1	No	*...etc*	

[21] http://tools.decc.gov.uk/en/content/cms/heatmap/about_map/about_map.aspx

The second data processing procedure required would involve allocating an energy efficiency rating to all addresses in the resulting data set that did not have a corresponding EPC rating. The suggested method to do this would be to produce an average EPC rating for each node option for the list of household variable options at a particular geographical area, such as English region/devolved nation or local authority level (lower than this is likely to produce too few records in the calculation of an aggregate energy efficiency rating). Rules would have to be created in order to ensure that each node was statistically robust and that the number of cases representing each node was sufficiently large. A possible option for this would be to use SPSS or similar statistical analysis software that would allow some regression analysis. CSE would recommend the use of CHAID modelling.

An alternative and more simple approach would be to produce and use a lookup table using Experian household variables and only records that had a corresponding EPC rating. This table could then be used to look up the average EPC rating of dwellings with matching property characteristics. An example of a potential lookup table is provide in Table 4.15 – in practice, a lookup table would also include region and rurality. (As an example, from such a table, all three-bedroom, owner-occupied detached houses, off the gas grid network and built before 1919 without an EPC rating, would be allocated an energy efficiency rating of 16.12.)

For both the address matching and statistical modelling, a powerful database programme such as SQL would be required due to the size of the files.

4.4.2 Costs of creating a combined EPC and Experian data set

The costs of producing an energy efficiency database with both EPC register and Experian data are separated below into data purchasing costs (Table 4.16) and data processing costs, with the former accounting for the bulk of costs (Table 4.17). Less than 10 per cent of costs would be accumulated through additional processing and modelling of the data. Nevertheless, this is an important and complex process. We have made estimates for the number of days required to perform the tasks assuming a high level of data processing expertise and some domain knowledge of the data sets at an assumed day rate of £500. This is obviously subject to significant variation based on the type of organisation conducting the work and these costs should be taken as purely indicative.

In total, the data required to produce the combined data set will cost between £158,000 and approximately £168,000, depending on the costs of accessing Scotland EPC data (Table 4.16). The additional processing costs will be in the order of approximately £7,000–£12,000 (Table 4.17). Overall total costs to build a Great Britain energy efficiency database using EPC data and Experian data would therefore be in the region of £165,000–£180,000, as shown in Table 4.18.

However, as with the first option considered (direct purchase and use of EPC data), there is a possibility that EPC data could be accessed for free under a researcher licence or similar agreement. If this were possible the costs of this option would alter significantly. This is noted in the table below. However, there is still some uncertainty about this possibility and to date, requests for access to EPC data have not been successful. Nevertheless, this situation may change in the near future and continuing requests for free access to EPC data should be considered.

Table 4.16: Data costs for a combined EPC and Experian data set for Great Britain

Data	Number of records	Min. costs	Max. costs	Comments
EPC data: England and Wales	~12.3m	£123,000 (or free if requests for researcher access to data are successful)	£123,000	These could reduce slightly depending on the number of duplicate EPC records.
EPC data: Scotland	~1.2m	*unknown*	*unknown*	Costs are subject to staff time required to access data and the cost of additional external IT consultancy fees.
Experian data: Great Britain	~26m	£35,000	£45,000	Dependent on whether Mosaic is purchased for low-income household ID.
Total data costs		£158,000	£168,000	
Total data costs (assuming free access to EPC data)		£35,000	£45,000	

Table 4.17: Additional data tasks and staff resource required for processing EPC and Experian data

Task	Number of records	Estimated number of staff days required[22]	Estimated staff costs based on a rate of £500/day	Task descriptions
Experian data cleaning	3,358,180	2–4	£1,000–£2,000	Removing duplicates
Experian to EPC address matching	3,358,180	4–6	£2,000–£3,000	Joining both data sets together
Experian and EPC data to ONSPD matching	3,358,180	2–4	£1,000–£2,000	Joining additional area level information by postcode from ONSPD
Modelling EPCs for Experian addresses with no corresponding data	1,498,645	6–10	£3,000–£5,000	Assigning a modelled EPC band to Experian records without any corresponding EPC data or where address matching has failed
Total		**14–24 days**	**£7,000–£12,000**	

[22] Depending on the expertise level of the staff used and the exact nature of the data processing requirements

Table 4.18: Summary of total costs for a Great Britain address level energy efficiency database compiled from EPC and Experian data

Item	Min.	Max.
Data	£158,000	£168,000
Staffing	£7,000	£12,000
Total (rounded to nearest £1,000)	**£165,000**	**£180,000**

4.4.3 Considerations of using a combined data set

While an aggregated data set would be based on an address level data, approximately half that would be modelled data based on average energy efficiency ratings. Unfortunately, without access to a significant amount of EPC data to test the model, it is not possible to state how true a reflection of the energy efficiency of dwellings this would be. In fact, because energy assessments have not (on the whole) been conducted on those addresses without EPC data, it would prove a difficult task to verify the modelled data. Furthermore, in areas where there was limited coverage of EPC data, the modelled data would reflect the bulk of the information used to produce aggregated data.

It is therefore not possible to quantify how much of an improvement the addition of Experian data and subsequent prediction of energy efficiency rating of properties without EPC data would be. Assuming that housing with similar characteristics (age, built form, number of bedrooms, tenure and gas connection) has similar energy efficiency, this option is likely to produce variable results in terms of accuracy, particularly as the variables used do not account for insulation levels or heating systems. Furthermore, using average energy efficiency ratings of similar types of housing could hide the extremes that exist and thereby potentially conceal areas of the most inefficient and most efficient housing. This point is raised as a potential data health warning rather than a certainty.

Nevertheless, this approach has been undertaken by other groups and departments with access to some of the data sets in question, notably the EST's Home Analytics tool (see Section 4.8.2). If cost was no barrier, this could be a favourable option. It would put in place the groundwork for a national address level data set, which could be augmented by a series of additional data sets, either commercially available ones or data available from local and national schemes. Section 4.8.3 describes some of the data that could be added to the data set over time.

However, while the accuracy of such an approach might be uncertain, what is less speculative is the cost associated with this option. The point to emphasise here is that it is uncertain how much additional initial benefit the increased cost and data modelling requirements would bring to the mapping. Furthermore, highlighting areas where there is low coverage of EPC data (which would be an additional output from using EPC data on its own) is a valuable exercise and useful output in itself.

In conclusion, this option is not dismissed outright but should be considered only if sufficient funds are available and there is a long-term plan to build a national address level housing efficiency data set. However, as at least two known parties are undertaking this work (DECC through the NEED

framework and the EST using the Home Analytics tool), this would replicate significant work that has already been carried out. For the purposes of this project the merits of this approach do not outweigh the increased costs, particularly when the costs only cover the use of the data for 12 months, with additional costs required after this point.

4.5 EHS and Census data (sub-regional EPC rating statistics)

Table 4.19: Summary details of modelling energy efficiency using Census data and the EHS

Data issue/requirement	Comment
Data owner	**Census data:** ONS **EHS:** DCLG
Number of records in database	**ONS:** 227,759 (England 171,372; Wales 10,036; Scotland 46,351) **EHS:** ~13,000
File format of data	Csv
Size of data file	Unknown
Direct costs	None (data is freely available to researchers)
Activities involving additional costs	Creating a regression model (including testing) Running the regression model and calibrating the results
Estimation of additional costs	£10,000–£15,000
Typical timescale to obtain data if known (from date of request to receiving data)	Within one week
Update requirements to maintain database accuracy	Latest ONS statistics on number of dwellings in a COA/LSOA
Ownership constraints (e.g. licensing issues, T&Cs etc.)	EHS data: Must not be distributed to any third parties
Technical requirements	Database software able to process more than 171,000 COAs in England alone (Microsoft SQL server on a high spec PC or server) Ability to conduct, test and run statistical regression models
Accuracy of EPC data at aggregate geography	Poor
Smallest geography of data	COA
Coverage of data	All COAs in Great Britain (England, Wales and Scotland)

4.5.1 Proposed usage of data

In the face of significant cost implications of an EPC data set and, to a lesser extent, an Experian based modelled energy efficiency database, this approach is put forward as one option that would utilise freely available data, and therefore incur significantly lower costs; the only costs incurred would be through staff time requirements of data acquisition, modelling and testing.

The data required is the full EHS[23], which would be used to produce a regression model, and Census data at COA, upon which that model would be run.

The EHS[24] is a continuous national survey commissioned by the DCLG. It collects information about people's housing circumstances and the condition and energy efficiency of housing in England. The EHS covers all housing tenures and provides valuable information that can be used to assess a wide amount of detail regarding the English housing stock and the households therein.

[23] At the time of writing, the latest version of the EHS is the 2012–2013 dataset. The 2013–2014 dataset should be available around July 2015.
[24] www.gov.uk/government/collections/english-housing-survey

Census data reports on a succession of detailed socio-demographic information and housing details collected during 2011. Much of the data is freely available through the ONS website at COA. Some of the variables that would be of interest for determining energy efficiencies of properties by COA include:

- tenure
- household composition
- dwelling type (built form), including number of converted flats (that is, those with lower than average levels of exposed surfaces)
- central heating system present and fuel type
- number of bedrooms
- total number of rooms
- rural and urban morphology
- English region, Scotland and Wales location
- number of vacant dwellings (also potentially useful as a reporting variable to highlight areas with a high number of vacant properties).

The EHS contains this information in a number of variables that could be matched to the Census metadata used to predict the categorical value of the EPC rating (that is, assign a rating band between A and G). The resulting model would then be run on Census data at output area level, with the results of the prediction producing the percentage of dwellings in each EPC band in each COA. This in turn would then be used to calculate the number of dwellings using the latest housing and population estimates for each COA. The information would then be aggregated to LSOA, MSOA and parliamentary constituency areas.

This option is based on a method used by the Building Research Establishment (BRE) that predicts numbers of fuel poor households by super output area (and other geographic areas) using similar data sets for DECC.

Fuel poverty itself is calculated by understanding the costs of heating and powering a dwelling and income information about the household inhabiting that dwelling. The energy costs themselves are an indication of how energy efficient (or inefficient) a property is, and the energy efficiency rating of a building is a function of:

- the costs of consuming energy in a dwelling to maintain a certain level of warmth
- the floor area of that dwelling.

The equation to calculate energy efficiency rating is calculated in a two-stage process:

$$\text{Energy cost factor (ECF)} = 0.47 \times \text{total cost}/(\text{total floor area} + 45) \quad (1)$$

Depending on the outcome of equation (1), the SAP rating is then calculated using either of the following equations:

$$\text{if ECF} \geq 3.5, \quad \text{SAP/EPC rating} = 117 - 121 \times \log_{10}(\text{ECF}) \quad (2)$$

or

$$\text{if ECF} < 3.5, \quad \text{SAP/EPC rating} = 100 - 13.95 \times \text{ECF} \quad (3)$$

It is therefore logical to conclude that if it is possible to produce national statistics on fuel poverty using a regression model on Census data, then a similar approach could be used to indicate levels of energy efficiency of housing in a given area using a similar approach.

However, it is important to note a couple of considerations. Firstly, there would be some distinct differences to the BRE methodology and what is proposed here. Most obviously, a calculation of the energy efficiency rating of dwellings is purely a function of the physical characteristics of the property and would not need the additional factors required in the BRE's regression model in order to determine household characteristics and likely income levels. Intuitively this would appear to simplify the problem somewhat.

However, the method applied by the BRE uses a data set that is augmented by several commercial data sets, most significantly of these is 'RESIDATA', now known as Experian data. Primarily this enables the addition of property age to the Census data, which is a key indicator (and therefore predictor) of the energy efficiency of properties. Including this in the list of data would incur some cost, approximately a fifth of total costs of Experian data specified in Sections 4.3 and 4.4 without Mosaic data.

However, the intention of this approach is to use freely available data. Purchasing commercial data may have some merits in improving a prediction method, but that would also reduce the benefits of an approach that seeks to minimise the costs of producing a housing energy efficiency database. Furthermore, and as is argued in previous sections, once the line of purchasing data has been crossed, the evidence suggests that obtaining actual EPC data provides the most satisfactory method of indicating energy efficiency levels of housing across the country. Therefore, the approach here evaluates the possibility of using freely available data only.

4.5.2 Costs of predicting energy efficiency from Census data

The data required for this approach is freely available. However, in order to access EHS data, an organisation must be registered for an account with the UK Data Archive and access the housing stock databases through the UK Data Service website. Registering for access through the UK Data Archive can be done here: www.data-archive.ac.uk/sign-up/credentials-application.

Bespoke Census 2011 data can be downloaded from the online ONS tool, Neighbourhood Statistics: www.neighbourhood.statistics.gov.uk. (No registration or user information is required to access Census data from this site.)

Where costs will be incurred through this approach is on staff time. In this case, it is difficult to estimate the amount of time required as it is very dependent on the type of software being used to store and handle the data, and to create, test and run regression models. Below are a list of tasks considered essential to this process and an indication of the amount of time required to perform them. A high level of staff expertise is assumed.

Table 4.20: Additional data tasks and staff resource required

Task	Number of records	Estimated number of staff days required[25]	Estimated staff costs based on a rate of £500/day	Task descriptions
Accessing EHS and Census data and processing	**EHS:** 13,000 **Census:** 227,759	2–4	£1,000–£2,000	
Producing and testing prediction model (EHS)	13,000	4–8	£2,000–£4,000	Use predictor variables in EHS to calculate energy efficiency band.
Running a prediction model (Census data)	227,759	5–10	£2,500–£5,000	Run statistical prediction model on Census data; produce numbers of dwellings in each EPC band in each COA.
Final processing and testing of outputs, including calibrating with National Statistics	227,759	5–10	£2,500–£5,000	Check and test outputs and ensure a match with national statistics through calibration of final results.
Total		**16–32**	**£8,000–£16,000**	

4.5.3 Considerations for using Census data and EHS based models

The most critical aspect of this option is successfully developing a reliable and accurate prediction model of energy efficiency. If this can be done satisfactorily, then there can be confidence in the accuracy of the predicted result when the model is applied to Census data. It is, however, worth noting that in the methodology document for the sub-regional fuel poverty statistics, the guidance states:

"Census output data [of fuel poverty] should be used with care, as this is a particularly small level of geography so data will be at its most approximate."

This could be considered a warning that predicting statistics using this type of methodology is not without significant error margins. In discussions with other fuel poverty practitioners, many believe that data in the sub-regional fuel poverty statistics series produced below local authority level is not considered to be reliable.

However, in order to understand how accurate an energy efficiency prediction model could be, we attempted to predict energy efficiency of all the records in the EHS 2012 using the CHAID

[25] Depending on the expertise level of the staff used and the exact nature of the data processing requirements

classification tree approach. CHAID modelling selects a set of predictor variables and their interactions that optimally predict the variability in the dependent measure. The resulting CHAID model is a classification tree that allows the analyst to identify clusters of households described by similar categorical variables, typically driven by socio-demographics or the characteristics of the property.

A model was created using predictor variables coded to match data contained in the Census 2011 at COA. The model was asked to use these to predict EPC band as defined by the EHS energy efficiency rating band variable '**EPceeb09e**'. The predictor variables included:

- built form
- tenure
- presence of central heating and fuel type
- total number of bedrooms
- total number of rooms
- English region
- rurality (an urban/rural classification).

The summary results comparing the original EPC band with the predicted value produced by the model are shown below in Table 4.21. The best CHAID model developed achieved an overall successful prediction rate of 54 per cent, and managed to predict 91 per cent of records rated at band D correctly. However, the model was unable to predict any dwellings at either end of the efficiency scale: no band A or band F or G dwellings were correctly assigned their efficiency band.

In fact, 86 per cent of all records were predicted to be a D band. Figure 4.6 shows these records, broken down by original EPC band, and provides a useful illustration of the limitations of this model.

Table 4.21: Summary CHAID results for predicting EPC band in the EHS

EPC band		Predicted						Percentage correct
		A/B	C	D	E	F	G	
Observed	**A/B**	0	85,052	45,549	4,495	0	0	0%
	C	0	1,319,320	2,508,373	27,391	0	0	34%
	D	0	917,745	10,251,467	143,095	0	0	91%
	E	0	214,633	4,964,801	196,243	0	0	4%
	F	0	61,630	826,716	138,288	0	0	0%
	G	0	10,764	173,868	45,809	0	0	0%
	Overall percentage	**0%**	**12%**	**86%**	**3%**	**0%**	**0%**	**54%**

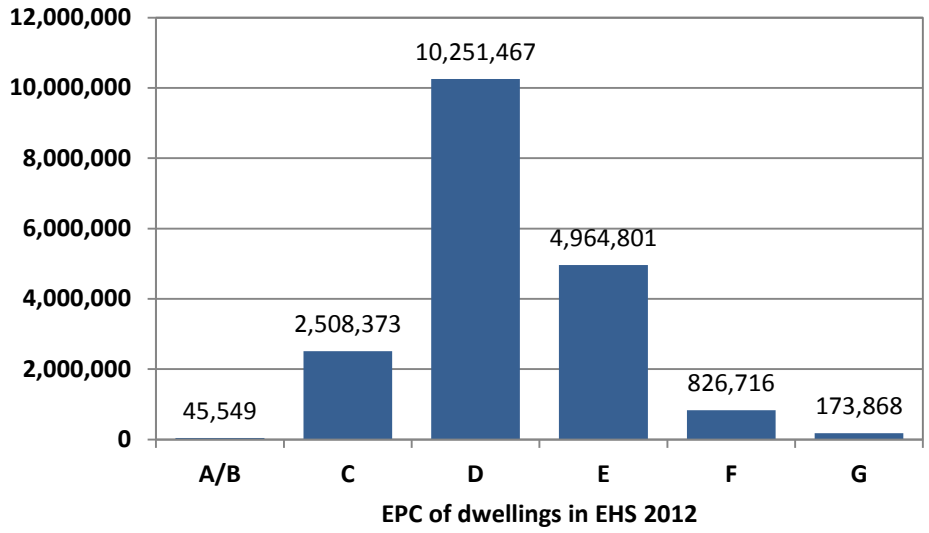


Figure 4.6: Original EPC band of dwellings predicted to have band D rating using a CHAID prediction model

Hence, it is clear that attempting to predict the energy efficiency band using these predictors is unlikely to yield meaningful results. Asking a model to predict a band from a choice of seven different categories might be complicating the situation, so an alternative modelling situation was considered whereby the model was asked to predict one of two outcomes: whether a dwelling has a efficiency rating of C or above, or D or below. A summary of the prediction success of this model is provided in Table 4.22. Initial examination of these results shows an increased level of success in predicting the outcome correctly, with 83 per cent of all outputs correctly assigning the simplified efficiency level. Overall, the model underestimates the numbers of dwellings in the band C or above category by approximately 2 million records, but is successful in predicting 95 per cent of dwellings with a SAP band of D or below.

The prediction model was examined further to look at the breakdown of different types of housing in each prediction category. Figure 4.7 shows the number of dwellings with an EPC rating of D or below from the EHS and a predicted by the CHAID model. These are broken down by built form. This suggests that there is a reasonable fit with these dwellings, although there does seem to be an over-prediction of houses (detached, semi-detached and terraced) with SAP band D or below.

Table 4.22: Summary CHAID results for predicting EPC band C or above or D and below in the EHS

		Predicted		Percentage correct
	EPC band	A to C rating	D or lower	
Observed	A to C rating	1,125,717	2,864,463	28%
	D or lower	817,302	17,127,757	95%
	Overall percentage	**9%**	**91%**	**83%**

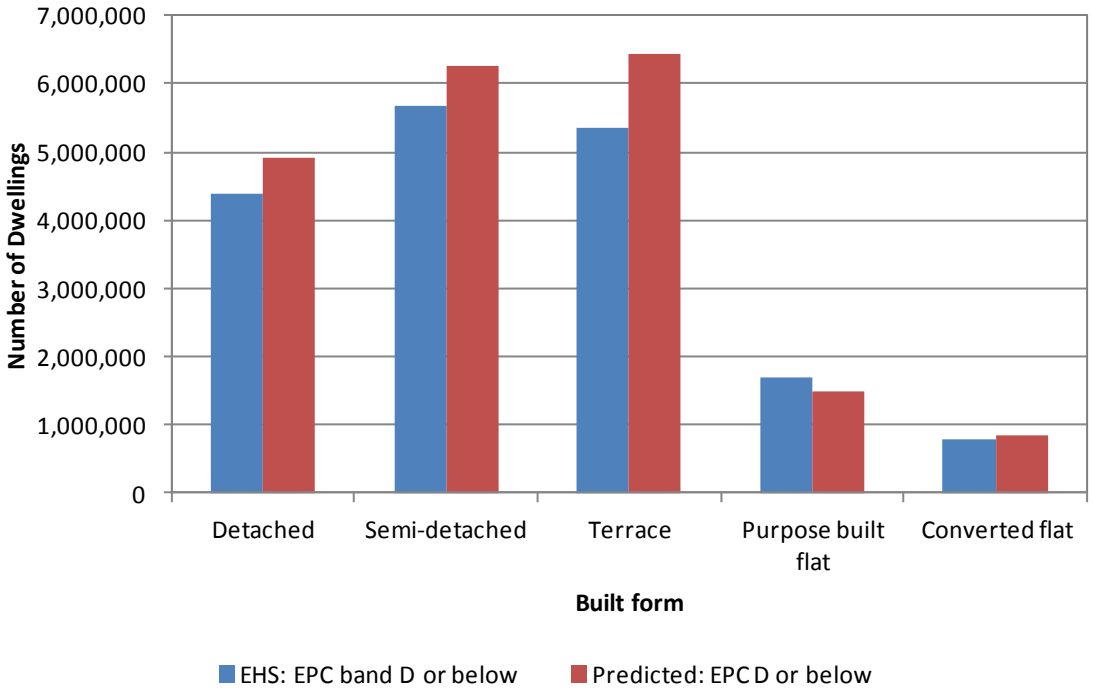


Figure 4.7: A comparison of the number of dwellings by built form for properties predicted to have energy efficiency of band D or below

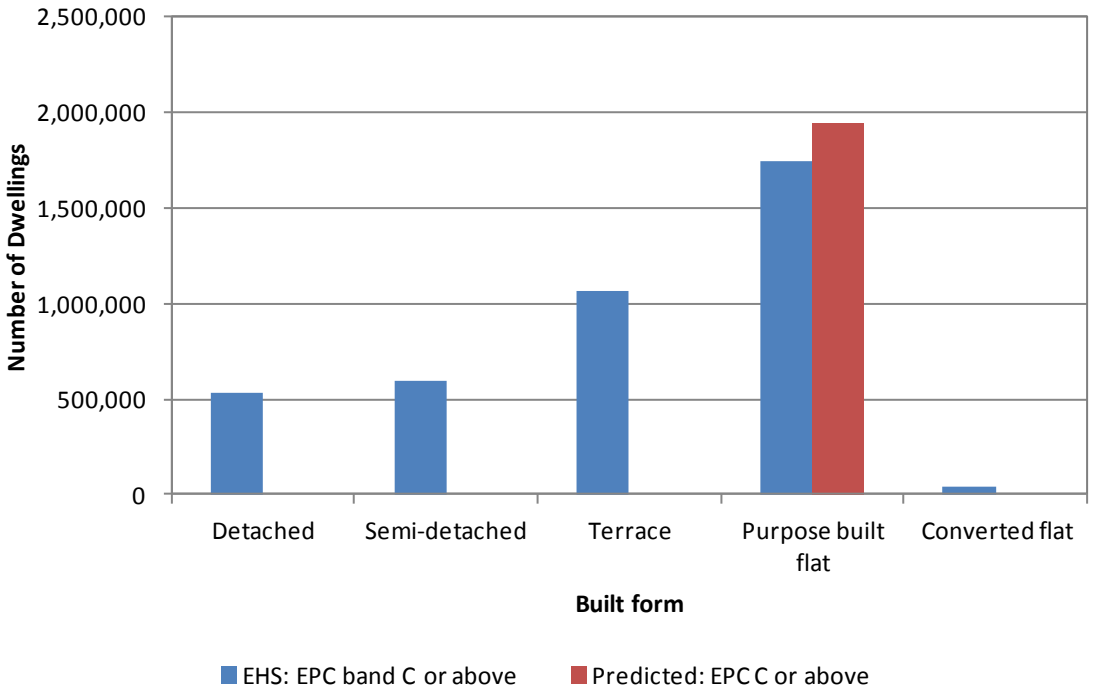


Figure 4.8: A comparison of the number of dwellings by built form for properties predicted to have an energy efficiency band C or above

However, when a similar analysis was performed on the model outputs for dwellings predicted to be band C or above, a significant problem was found with the model. Figure 4.8 shows how only purpose-built flats where predicted to have a band C or above. All housing and converted flats were predicted to be band D or below. This has some series implications when determining the energy efficiency rating of dwellings using Census data – only areas with purpose-build flats would register as having dwellings with efficiency ratings above band C. **These results strongly suggest that this approach is not a viable option for this project.**

As mentioned previously, this model could potentially be improved with the addition of other data, in particular adding age of properties to the Census data by purchasing this variable from Experian and aggregating and matching to each COA. Property age could then be used as a predictor variable in the CHAID model and may improve the prediction rate. However, there are also some data issues with Census data itself, and three particular points are worth making, if not the main considerations for ruling out this approach.

Firstly, Census data was collected in 2011 which make it almost four years out of data, while EHS data is around two years out of date. In comparison, while some EPC data will be further out of date than that (that is, lodged in 2008–2010), a significant amount of EPC data has been collected in recent years and, with the exception of a small number of properties that will have opted to be removed from the register, any dwelling receiving measures from the Energy Company Obligation (ECO) and the Green Deal will have had an EPC rating registered in the database.

Another commonly cited problem with using the Census is that the data itself has been collected through self-completed questionnaires. Any self-reported data should be treated with caution as its accuracy can vary significantly due to a variety in understanding and knowledge of questions being asked (indeed this is one of the reasons why property age is not available through the Neighbourhood statistics tool). The Census is no exception to this.

Finally, when considering update options for an energy efficiency database it should be pointed out that although some peripheral data suggested here will be updated annually, the next set of Census data will not be available until after the next Census in 2021. This has significant implications for any plans to update the data.

4.6 NEED framework

The NEED framework was first announced in the Heat and Energy Saving Strategy in 2009 and was developed by DECC, with support from the EST and gas and electricity suppliers. DECC's stated intention of developing the framework was to assist with its business plan priority to "save energy with the Green Deal and support vulnerable consumers". The data framework provides the largest source of data available for analysis of energy consumption statistics and the impacts of installing energy efficiency measures.

The framework is an address level data set pinned together by information from National Land and Property Gazetteer (NLPG) data. Additional information from public and commercial data sets is used to provide data on geographical location (small area geographies), energy consumption (from meter point data), improvement measures installed from various energy efficiency schemes, property attributes and household characteristics (obtained from a range of sources), energy efficiency information (from EPC data) and fuel poverty statistics.

The NEED covers both non-domestic buildings and domestic dwellings. The full list of data sources used to produce the domestic NEED and the data they contribute is provided in Table 4.23.

However, the full domestic data framework is not a publicly available data set. This is for a number of reasons, predominantly due to the fact that the several commercial data sets used in the framework (for example, Experian, NPLG, Valuation Office Agency (VOA)) are subject to stringent data sharing protocols and restrictions. (As mentioned in Section 4.3 above, licensees of Experian data are not permitted to "sell, transfer, sub-license, distribute, commercially exploit or otherwise make available to, or use for the benefit of, any third party any Experian data".)

However, DECC have explored making some data from the NEED framework publicly available. In July 2014, there were two publicly and freely available data sets:

- A Public Use File (PUF) consisting of 49,815 records selected to be representative of the housing stock, available at data.gov.uk/dataset/nationalenergyefficiencydataframeworkanonymiseddata.
- An End User License (EUL) file: Anonymised sample of 4,086,448 records available from the UK Data Archive (registration required) at discover.ukdataservice.ac.uk/catalogue/?sn=7518&type=Data%20catalogue.

These draw on data from a range of sources matched together at household level and anonymised to ensure no individual household can be identified from the data. The PUF serves merely as an example of the data for perusal, while the EUL file has been weighted to represent all housing in England and Wales.

However, EPC data is not incorporated into the core NEED dataset held by the Department of Energy and Climate Change (DECC). It is used by DECC for supplementary analysis only. In order to produce the EUL dataset available on the UK Data Archive, DECC used 4 million EPC records as a 'data spine'[26]. These EPC records were selected as representative of the wider EPC data (in terms of both

[26] The information regarding this methodology was obtained through correspondence with the NEED team at DECC.

energy efficiency profiles and geographical distribution of the nationwide data) and a selection of NEED variables appended to this.

Table 4.23: Data sources used in NEED

Category	Source	Description
Premises	National Land and Property Gazetteer (NLPG)	Contains a unique identifier for each address in England and Wales, which acts as the spine of NEED.
	One Scotland Gazetteer (OSG)	As above, for Scotland, provides a unique identifier for every address in Scotland.
Geography	Office for National Statistics Postcode Directory (ONSPD)	Provides information on postal and administrative boundaries (including local authority).
Energy consumption	Energy suppliers and Xoserve/Gemserv	Gas and electricity meter consumption data for all domestic and non-domestic meters in GB, 2004–2011 and meter profile number for electricity meters. Gas data is weather corrected.
Measures installed	Homes Energy Efficiency Database (HEED)	Information on energy efficiency measures installed through government schemes (including Energy Efficiency Commitment, Carbon Emissions Reduction Target (CERT) and Community Energy Saving Programme (CESP).
Property attributes	Valuation Office Agency (VOA)	VOA is an executive agency of HM Revenue and Customs, responsible for business rates and council tax. It collects property detail data to inform this function. The data includes floor area (domestic and non-domestic), property type and property age (domestic only).
	Experian	Modelled data for domestic property attributes such as number of bedrooms, property age and type. Purchased for 10% of households in England.
	Homes Energy Efficiency Database (HEED)	Includes wall construction and property type, but accuracy and coverage vary.
	Energy Performance Certificate (EPC)	**[not currently used in the main framework but sited for use in future NEED development work]** Information likely to be included: Energy efficiency rating information, wall type and main heating fuel (where not available from other sources).[27]
Households characteristics	Experian	Modelled data for household characteristics such as income and tenure. Purchased for 10% of households in England.
	Output Area Classification (OAC)	Categorises geographic areas each containing approximately 125 households into 21 socio-economic groups, based on the 2001 Census.
Fuel poverty statistics	Fuel poverty sub-regional statistics	(Public anonymised EUL data set only.) Allocation of fuel poverty status of households to one of five bands based on the estimate of the proportion of households in fuel poverty in the LSOA (2001) they are located in. Uses the 2011 estimates of fuel poverty low-income, high-cost definition.

[27] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/437164/Domestic_NEED_Methodology_June_2015.pdf

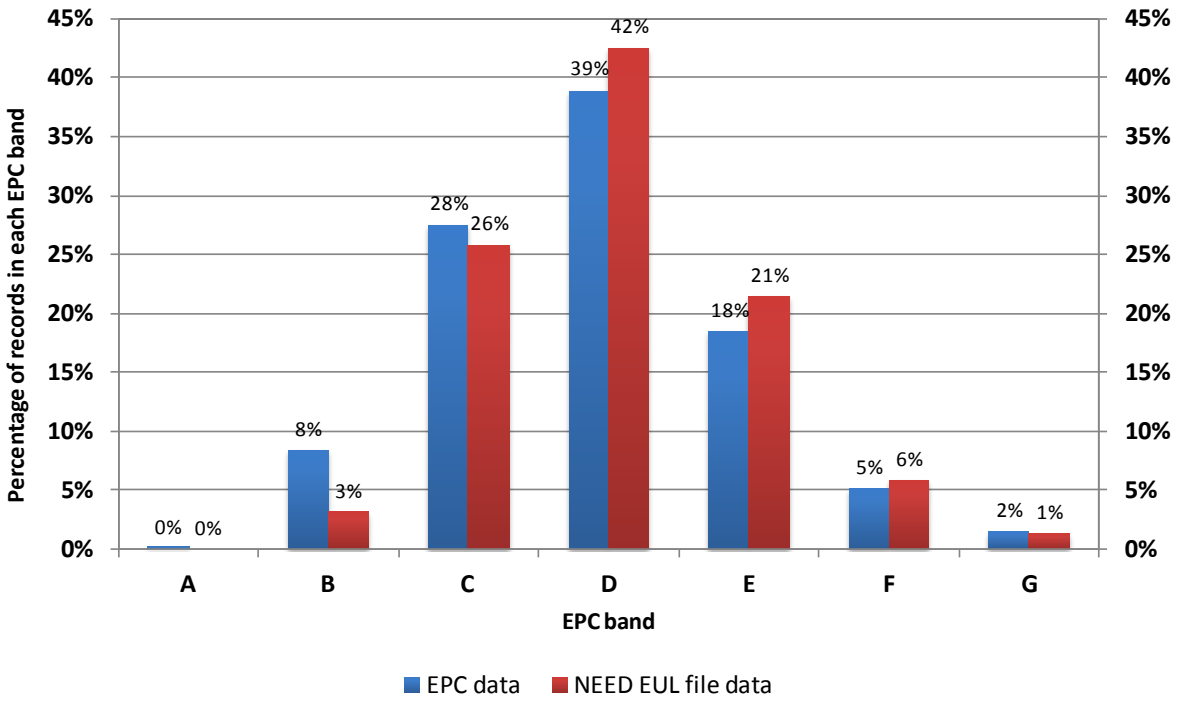


Figure 4.9: Proportions of dwellings in each EPC band in England and Wales, from EPC live tables data and the NEED EUL file

Table 4.24: Number of dwellings by EPC band and region from the NEED EUL file

Region	Band A/B	Band C	Band D	Band E	Band F	Band G	Total
North East	33,132	344,922	529,829	208,772	47,570	10,577	1,174,801
North West	96,054	838,716	1,330,882	632,243	156,482	38,026	3,092,402
Yorks/Humber	63,092	588,509	991,236	477,439	122,920	30,343	2,273,538
East Midlands	55,171	458,904	798,831	449,912	131,105	23,983	1,917,905
West Midlands	53,513	536,199	1,003,450	540,049	145,545	34,558	2,313,313
East of England	70,078	652,975	1,056,392	525,296	144,201	27,897	2,476,840
London	144,477	884,482	1,292,671	674,474	149,096	30,213	3,175,414
South East	114,161	936,370	1,555,292	748,232	202,618	42,087	3,598,759
South West	81,122	588,000	948,438	490,850	174,781	46,187	2,329,377
Wales	28,020	298,465	554,048	320,599	106,858	32,505	1,340,495
Total	**738,821**	**6,127,542**	**10,061,067**	**5,067,865**	**1,381,174**	**316,375**	**23,692,844**

As a result, there should be an element of caution used when using such data. According to staff at DECC, there is recognition that EPC data is somewhat biased towards more efficient properties and there have been some attempts to counter this in the energy efficiency information in the EUL file. In order to explore the overall reliability of the energy efficiency data in the EUL file, some analysis was conducted and compared with information available in the DCLG live tables on EPC data. These

are shown in Figure 4.9. Overall, the figure suggests that EPC data currently held on the register slightly overestimates the energy efficiency levels of housing across England and Wales, and that – in order to account for this – the NEED team have attempted to balance the energy efficiency ratings in the EUL file when selecting EPC records for the data spine and when weighting the dataset.

Unfortunately, the data in the EUL file is currently completely anonymised data, with the lowest geographic area variable being English region or Wales. Therefore a direct request was made to the NEED team at DECC to see if it was possible to include a lower level geography indicator in the EUL file, or, alternatively, a data set derived from the EUL file showing numbers of properties in each EPC band by LSOA. The latter request was fulfilled and DECC has subsequently provided a data set for England and Wales. This data has now also been made publicly available on the DECC website[28]. The successful release of this data meant that Stage 1 of the project advanced further than originally anticipated; it resulted in data being made available that could potentially be used to map energy efficiency levels across England and Wales. Therefore, the NEED-based EPC data at LSOA level was examined and analysed to determine whether it could be used to produce energy efficiency maps. This is summarised below.

Along with the limitations associated with basing the data on 4 million EPC records, the data was provided with the following caveats:

- Tables are based on the EUL data set. While the EUL weighting is applied, it is not intended to make the data set representative at LSOA level, and was produced to be representative at region level. Therefore, the results should be used with caution.
- The data is based on EPC data up to October 2012. There is strong bias in the EPC data with a lot more energy efficient properties having had an EPC rating compared with the general housing stock. The weighting helps improve the representation of the overall results. However, some LSOAs, for example those with a high proportion of new builds, are not likely to be representative.
- Any cells that are based on fewer than six records have been suppressed to protect against disclosure. Alternatively, there may be some LSOAs that did not appear in the anonymised data set sample, as the sample was not selected to be representative at LSOA level. As a result, no information is provided for these LSOAs.

In summary, the caveats above suggest that at LSOA level the data should be treated with caution. It is therefore sensible to aggregate this data to higher geographies in order to reduce the bias caused by new builds and missing data.

This data was provided in a useful and simple structure and requires little processing in order to use it for the production of maps. It includes data for 34,741 LSOA records across England and Wales meaning that just 12 LSOA are not included in the data set. Overall, the data provided by DECC contains around 800,000 fewer records than the EUL data set, mostly the result of suppressing data due to disclosure concerns (as outlined above in the data caveats). As Figure 4.10 suggests, this

[28] www.gov.uk/government/statistics/national-energy-efficiency-data-framework-need-ad-hoc-requests-2014

appears predominantly to affect properties in the highest and lowest bands where the number of properties are smallest. Overall at a national level, the data provided by DECC appears to match the EUL EPC data well. The main difference is that a slightly higher proportion of dwellings are in the C, D and E bands and slightly lower proportion in the A and B and the F and G bands (Figure 4.10).

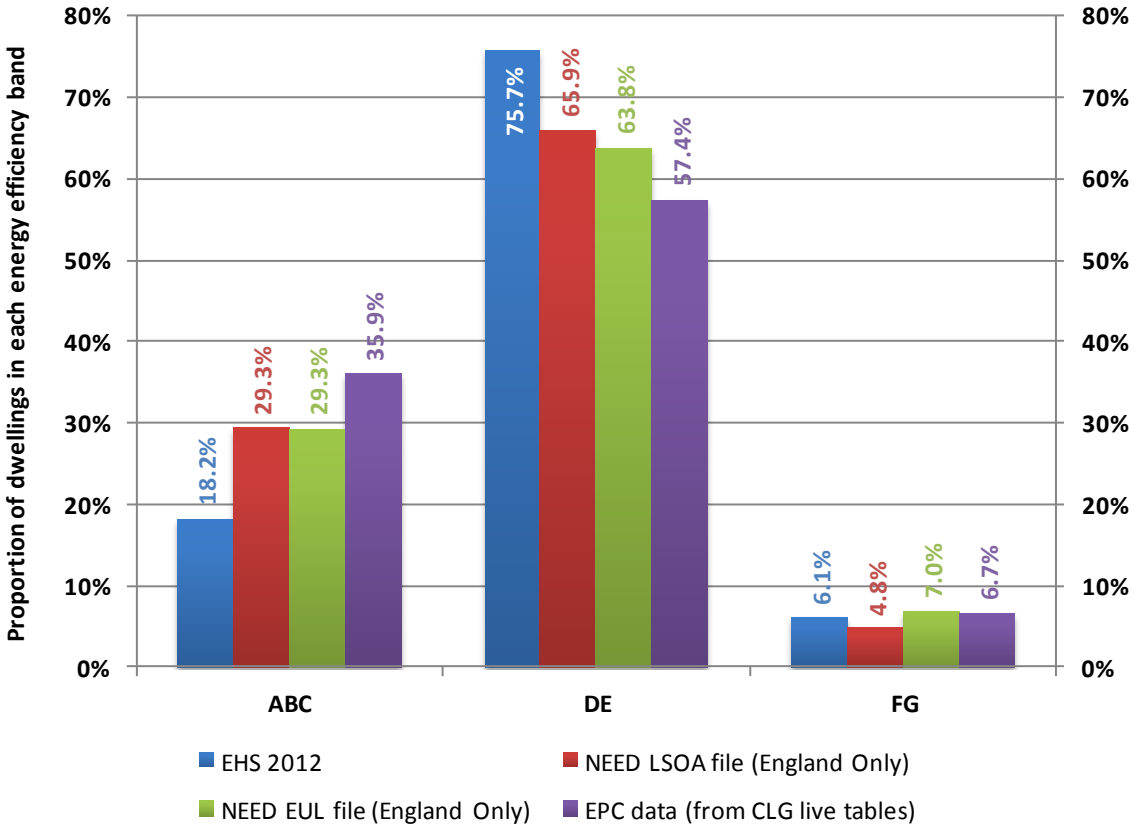


Figure 4.10: Proportions of dwellings in each EPC band in England and Wales, from English Housing Survey (EHS) 2012, EPC live tables data, the NEED EUL file and EPC data provided by DECC at LSOA level from the NEED EUL file

In conclusion, this data and the methodology used to produce it is less than perfect in terms of representing the energy efficient of the entire housing stock. However, providing that is it used cautiously and with recognition of varying levels of accuracy at certain geographies (e.g. LSOA), it provides a useful source of data with which to examine and spatially represent energy efficiency data for the English housing stock. This has been demonstrated below.

The LSOA data was examined and aggregated to higher levels of geography, including MLSOA and parliamentary constituency using the ONSPD. Using this latter data set, the 20 constituencies with the highest proportion of properties in C bands or higher were determined and are presented in Figure 4.11. Similar, the 20 constituencies containing the highest proportion of F and G properties were also examined. This is presented in Figure 4.12 below.

The full data set was also used to produce maps of England and Wales by parliamentary constituency. Figure 4.13 below shows the proportion of homes in each constituency that are rated

C or above, while Figure 4.14 shows the proportion of homes in each constituency that are rated F or G.

LSOA data can be matched to a range of other publicly available data sets at LSOA, such as the IMD series, fuel poverty statistics, health statistics and a rate of Census data, allowing for more specific identification of areas using different socio-demographic indicators.

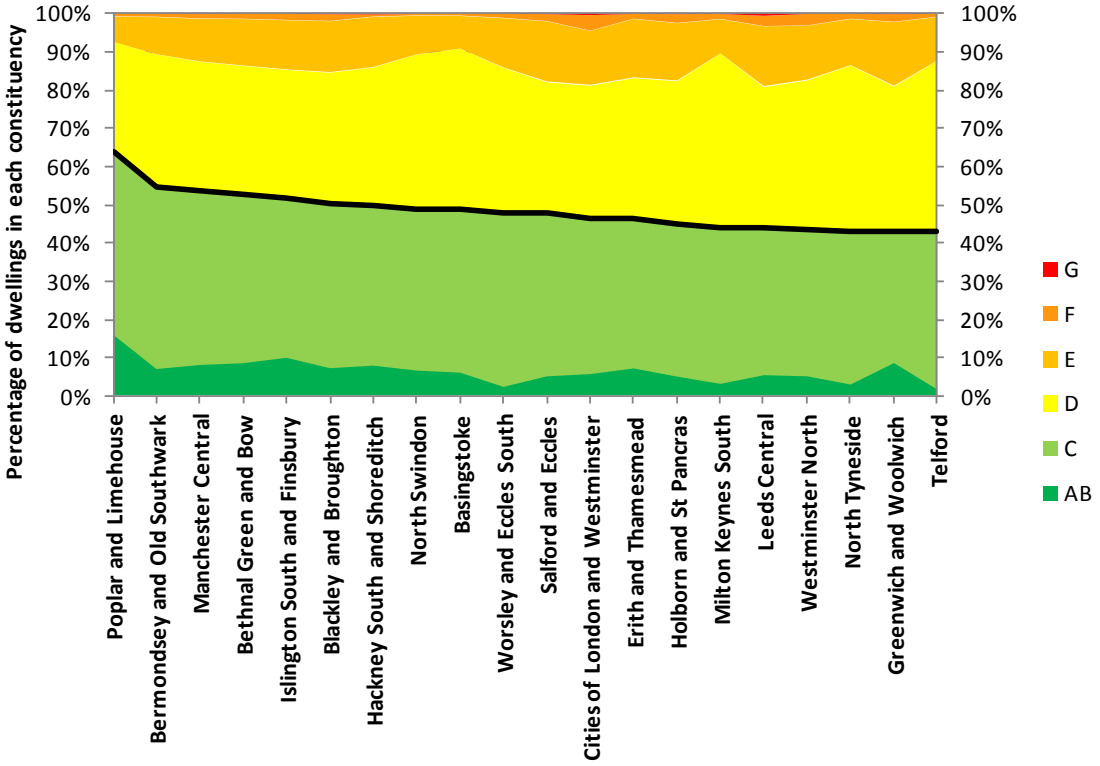


Figure 4.11: Proportions of dwellings in each EPC band for the 20 parliamentary constituencies with the highest proportions of dwellings in band C or above (England and Wales only)

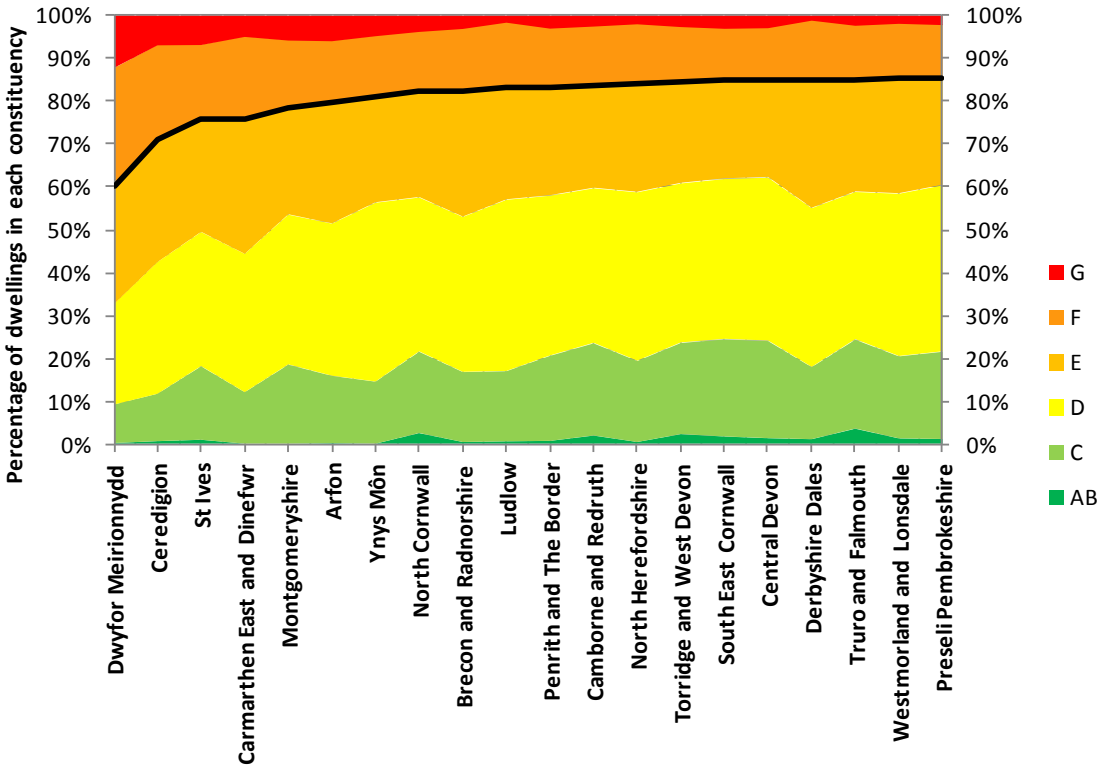


Figure 4.12: Proportions of dwellings in each EPC band for the 20 parliamentary constituencies with the highest proportions of dwellings in bands F or G (England and Wales only)

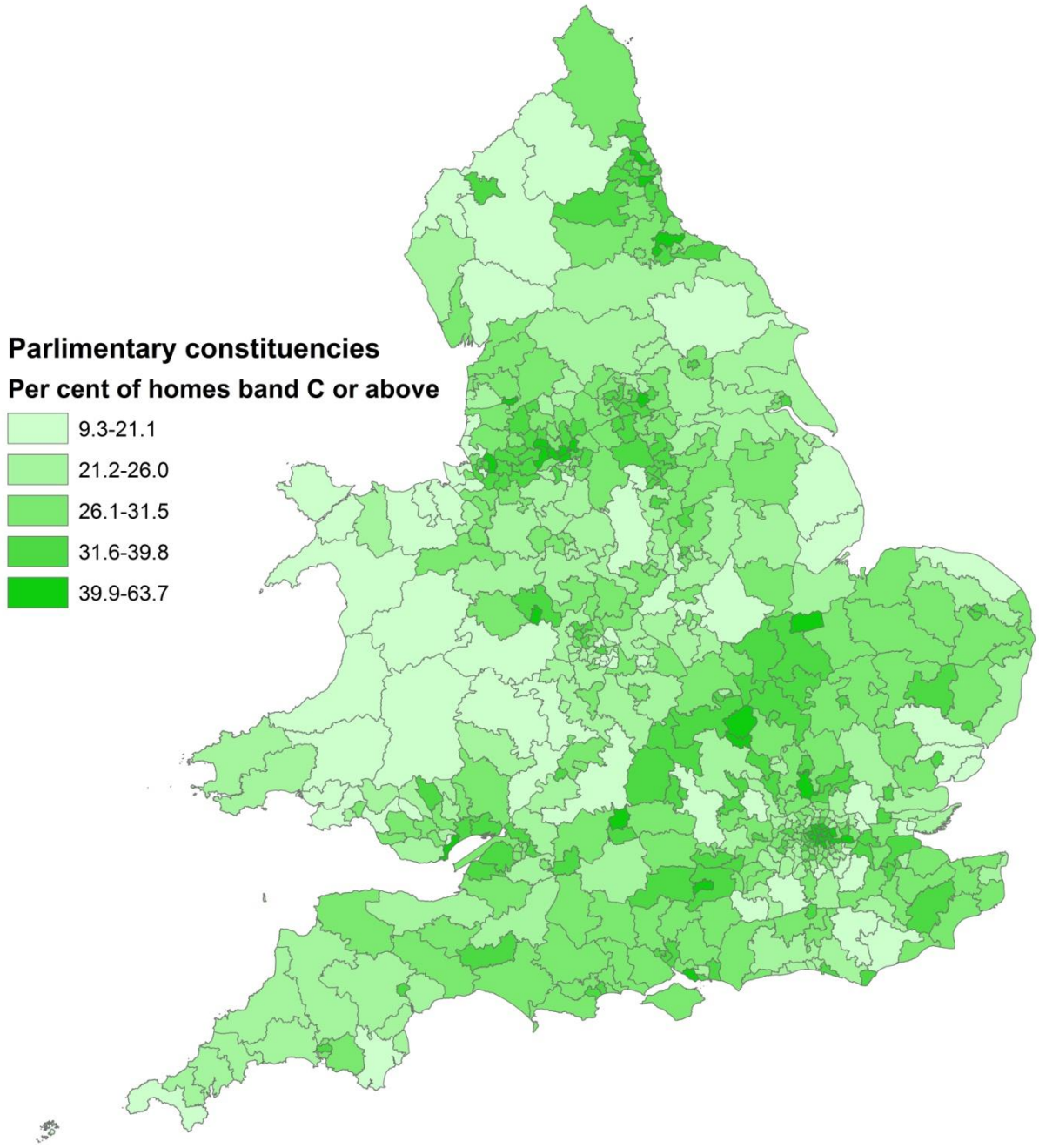


Figure 4.13: Proportion of dwellings with EPC ratings of A, B or C by parliamentary constituency from data provided by DECC using NEED

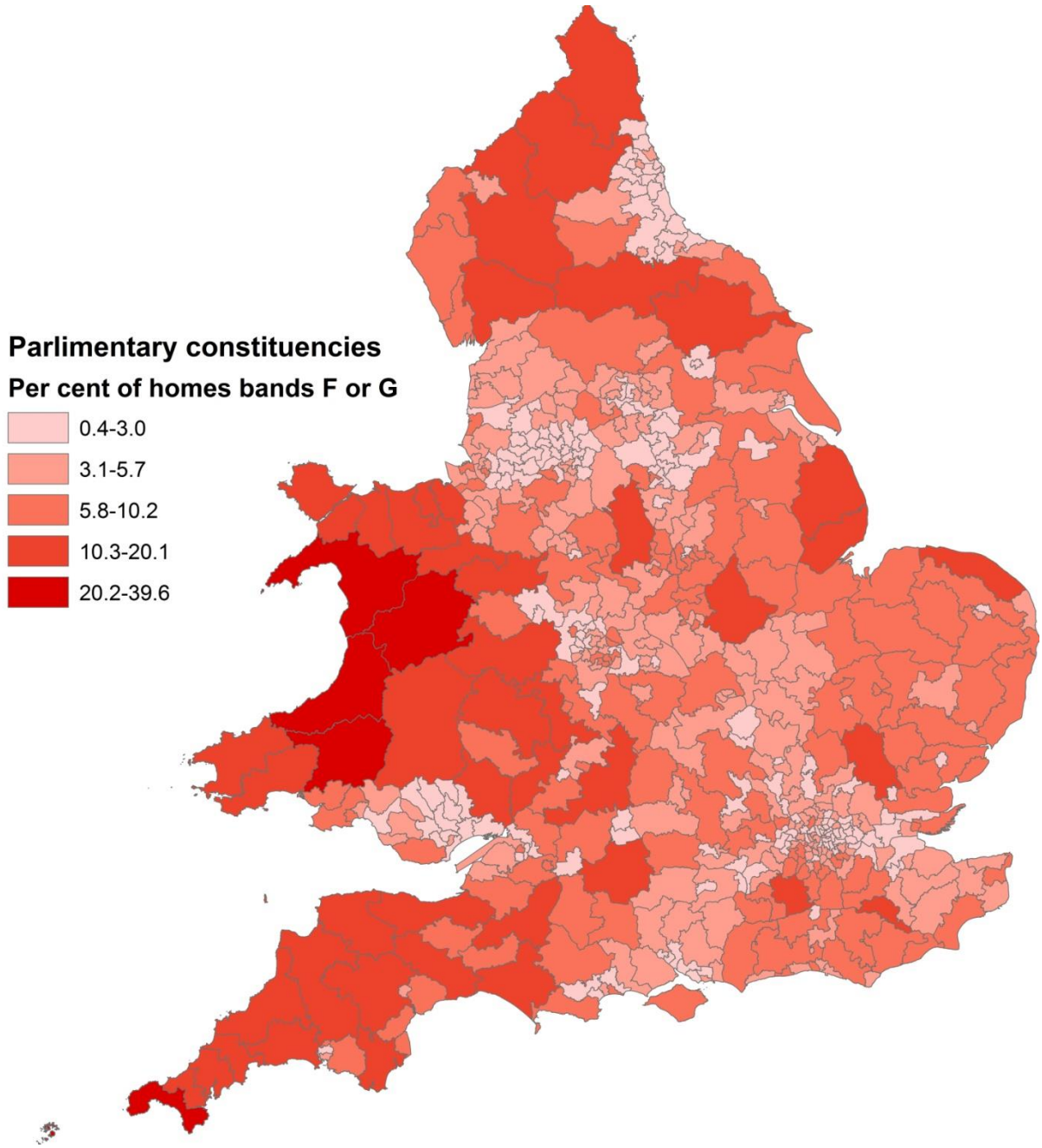


Figure 4.14: Proportion of dwellings with EPC ratings of F or G by parliamentary constituency from data provided by DECC using NEED

4.7 Request to the DCLG for EPC statistics from the domestic EPC register

EPC data on domestic properties, aggregated to a high enough level to ensure it is non-disclosive, would represent a valuable asset with respect to the identification, targeting and improvement of areas with high numbers of inefficient housing, as well as help reduce fuel poverty. Aggregated data on EPC data is already published at local authority level by the DCLG on the government website. This is very much welcomed, even though it is at a level which is not geographically specific enough in terms of targeting of energy efficiency schemes. However, the implication is that aggregated data of EPC records can be placed in the public sphere without concerns about revealing personal information or conflicting with the DPA.

In addition, 'constituency' (presumably Westminster parliamentary constituency[29]) is specified as one of the variables that would be included when purchasing a small data pack of EPC data, and therefore must be a key piece of information held in the register for each record. Some statistics are also available broken down by tenure and type of property, which suggests that inclusion of breakdowns by these characteristics would also be possible. Therefore, it seems a reasonable assumption that it is within DCLG's capability to produce aggregated data showing the number of domestic EPCs by each band (A to G) for each constituency, with additional breakdowns by tenure and property type.

So an additional alternative approach to access energy efficiency data for housing in the UK would be to directly request a free version of the EPC data from DCLG under a researcher licence or similar agreement. If permission were granted then all the considerations discussed in Section 4.2 concerning the direct use of EPC data would apply. The EPC data would be used to produce some aggregated statistics on EPC data at different geographical areas, and include the addition of socio-demographic data where available. Under such licence conditions, the original EPC data could then be deleted and the processed, and aggregated data mapped. The processing would ensure that any released data and maps produced would not present any data disclosure issues and follow all guidelines published by the ICO.

At the time of writing, requests to the DCLG have not been successful in accessing any other data other than the live tables available on the DCLG website.

Should the DCLG fail to comply with this request then a final possible approach could be to make a more formal FOI request for EPC data at parliamentary constituency level. This would not be a request for address level records from the EPC register but a request for aggregated data at constituency level. (It was revealed in a separate FOI request that the DCLG sought legal advice and clarified with the ICO, that EPC data should be categorised as personal data; it provides a significant amount of information about specific addresses. This can be further linked to and have an impact on an individual. Further information is available in the 'Data Protection Technical Guidance – Determining what is personal data', available from the ICO[30].)

There are not many examples of FOI requests concerning EPC data made to the DCLG. The most pertinent one was made in June 2011 and successfully requested the number of EPCs lodged for

[29] For some records, this could be to include pre-2010 parliamentary boundaries where lodgements were made between 2008 and 6 May 2010.

[30] ico.org.uk/media/for-organisations/documents/1554/determining-what-is-personal-data.pdf

each energy rating (that is, 1–100) in each year.[31] Although the data required for number of EPCs by constituency is more detailed and would require more processing and staff time to produce, it suggests that FOI requests of this nature are feasible and have a chance of being successful. Furthermore, the request cited above was completed within 20 days of the request being made.

Box 4.2 highlights some of the considerations and details about making an FOI request.

Box 4.2: A summary of the main considerations and details involved in FOI requests[32]

Considerations:

- Any information that is believed to be held by a public authority can be requested. The right only covers recorded information which includes information held on computers, in emails and in printed or handwritten documents as well as images, video and audio recordings.
- Some information may not be given to you because it is exempt, for example because it would unfairly reveal personal details about somebody else *(this would cover address level EPC data but not aggregated data)*.
- It is not necessary to know whether the information required is covered by the Environmental Information Regulations or the Freedom of Information Act. When making a request, it is for the public authority to decide which law they need to follow.

Legal requirements:

- Contact the relevant authority directly;
- Make the request in writing, for example in a letter or an email.
- Give your real name; and
- Give an address to which the authority can reply (postal or email address).

Outcomes:

- The authority contacted **must reply within 20 working days**. If additional time is required to consider the request this may be extended to 40 working days. A reason must be provided for extending the time limit, for example under the Environmental Information Regulations, more time is allowed if the information requested is particularly complex or of a significant size.
- The **authority is permitted to request a fee** for fulfilling the request, although there are rules regulating these charges.
- If the authority refuses to provide the information it must give reasoning behind the decision.

[31] https://www.whatdotheyknow.com/request/residential_epc_data#outgoing-137348
[32] ico.org.uk/for-the-public/official-information/

4.8 Other data sets and tools

4.8.1 The ONSPD

The ONSPD contains details of all postcodes in the UK, Isle of Man and the Channel Islands. The data set assigns all postcodes to different geographical areas including the following:

- Westminster parliamentary constituency.
- Electoral ward/division.
- Local authority district (or equivalent).
- 2011 Census Output Area (COA) (GB).
- 2011 Census Lower Level Super Output Area (LSOA).
- 2011 Census Middle Level Super Output Area (MSOA).

With the exception of 89 postcodes in the August 2014 release of the directory, all postcodes have a parliamentary constituency code associated with them. This would enable all address or postcode level data obtained to be matched to parliamentary constituency, as well as the other geographic areas mentioned above. Box 4.3 below provides a description of the constituencies available in the ONSPD. The ONSPD data set is freely available by request, with details available on the ONS website.

Box 4.3: Westminster parliamentary constituency areas in the ONSPD[33]

Westminster Parliamentary constituency areas in the ONSPD

England and Wales

Parliamentary Constituencies relate to those defined by the Parliamentary Constituencies (England) Order 2007, the Parliamentary Constituencies (England) (Amendment) Order 2008 and the Parliamentary Constituencies (Wales) (Amendment) Order 2008, and which came into effect at the May 2010 General Election. No further changes are envisaged until 2015.

Scotland

The Scottish Parliament (Constituencies) Act 2004, passed by the UK Parliament in July 2004, came into effect upon the dissolution of the last UK Parliament and breaks the linkage between UK Parliament constituencies and Scottish Parliament constituencies. This has enabled the Westminster Government to reduce the number of Scottish constituencies in the UK parliament in line with recommendations made by the Boundary Commission for Scotland following its Fifth Periodical Review of Constituencies (December 2004).

The Parliamentary Constituencies (Scotland) Order 2005 details the reduction of the number of constituencies for the UK Parliament from 72 to 59 Scottish seats. The changes bring the size of Scottish constituencies more into line with the size of English seats. English constituencies currently contain nearly 70,000 voters on average, compared to the Scottish average of 53,500.

[33] ONS Postcode Directory User Guide, 2014 Edition, ONS

4.8.2 Home Analytics tool (Scotland)

The EST Scotland's Home Analytics tool aims to help the Scotland Government and local authorities design, develop and target energy efficiency schemes and policies.[34]

EST Scotland has updated the latest version of the tool to include EPC data, enhancing the information held in the database and substantially increasing the number of variables available. The core data sets that now underpin the Home Analytics tool are:

- Energy Performance Certificate (EPC) data
- Home Energy Efficiency Database (HEED) data
- Energy Saving Trust Home Energy Check (HEC) data
- Ordnance Survey AddressBase, MasterMap Topography layer and StreetMap data
- Scotland Gas Networks data
- Scottish Census data
- Scottish Index of Multiple Deprivation (SIMD) data.

The key data sets in terms of describing properties are EPC, HEED and HEC data, and it is likely that this data set could provide all the data required to map energy efficiency in Scotland. Discussions with staff involved in the production of Home Analytics revealed that EPC data has been modelled for addresses without certificates, using property descriptions such as built form and property age to predict likely EPC rating and provide a band for all addresses. (This is a similar approach to the option suggested in Section 4.4.)

However, data access issues are again likely to be a problem for a non-governmental organisation to obtain, and unlikely to be a realistic option for this project.

4.8.3 Supplementary address level property data

Supplementary data could be used to enhance the depth of information about individual properties when constructing an address level database to use in order to estimate EPCs through use of energy models (for example if using Experian data or data from NEED). Figure 4.15 below shows a possible process of creating a database using Experian and additional property data to model EPC rating and produce an EPC database.

This data could be obtained from a variety of sources including local authorities, large housing associations, local and national scheme and policy information, local authority owned housing surveys, installer and national institution's data (for example Fenestration Self-Assessment Scheme (FENSA), Cavity Insulation Guarantee Agency (CIGA), CORGI/Gas Safe etc.).

While the majority of this data is freely available, not all organisations will have access or be eligible to obtain this data. In addition, address matching this information to the original address level database constitutes a lengthy and time consuming process. For some of the options described here that do not use EPC data directly, this would add significant information to the data, in particular an

[34] Details of the Home Analytics tool can be found at www.energysavingtrust.org.uk/scotland/organisations/national-and-local-government/home-analytics.

Experian data approach where modelled data on housing characteristics could be replaced by primary data.

However, many of the data sets are unlikely to be national data sets and will be owned or only obtainable by local authorities. Furthermore, obtaining this data would be a lengthy process; performing further data cleaning and then address matching of each of these files to a national address level data set would represent a very considerable amount of work and one that would involve significant timescales and substantial staff costs. Therefore, it is not recommended for a Great Britain-wide data set option. However, it is a more realistic prospect for smaller geographies, for example ranging from constituency areas to English region level.

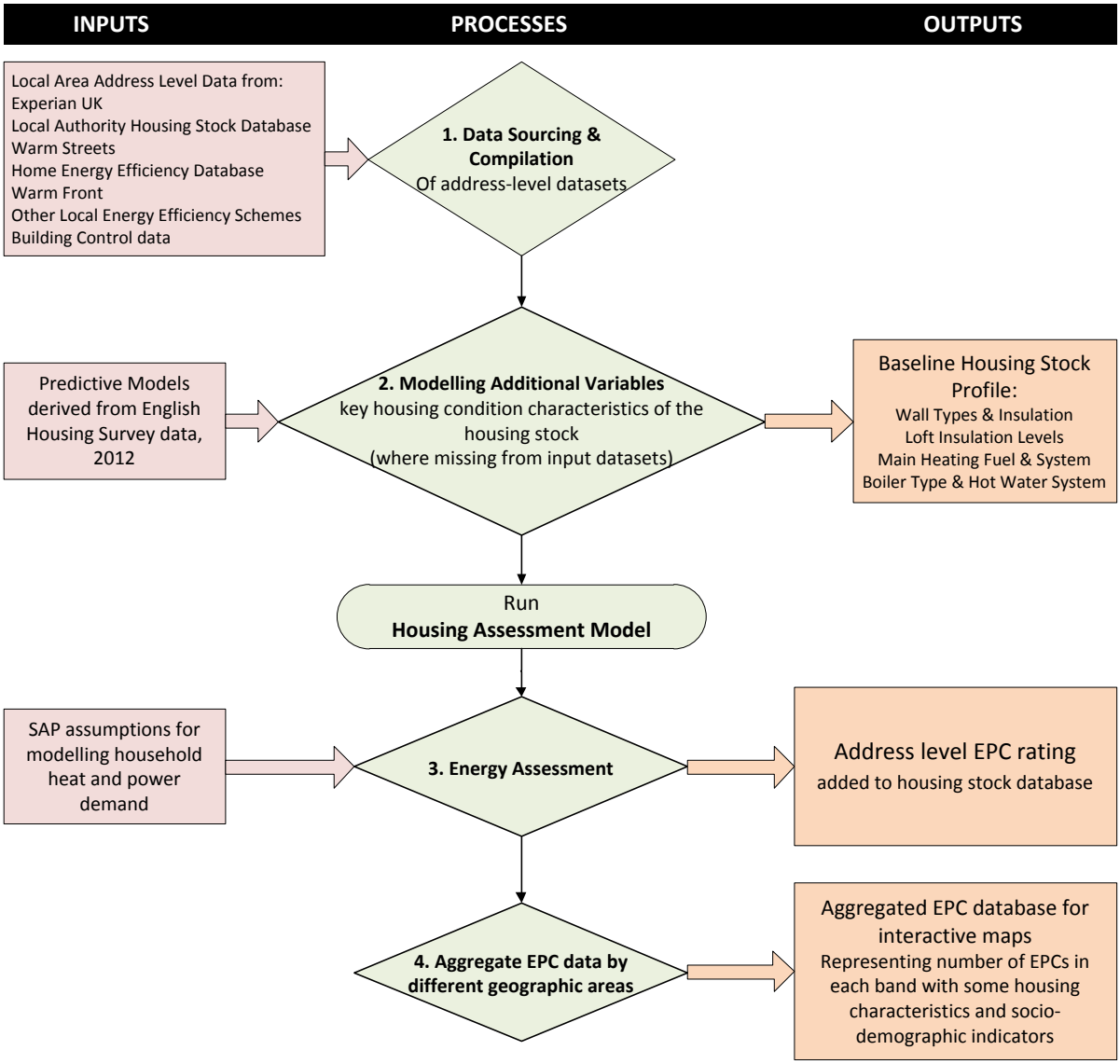


Figure 4.15: Suggested process of modelling EPC at address level using Experian data and other address level information (not recommended at national level)

5 Summary analysis and recommendations

5.1 Great Britain option

Section 4 evaluated a range of options for producing a comprehensive database of energy efficiency ratings. These include the direct use of energy efficiency data, modelling of energy efficiency ratings and approaches that combine the two. From exploratory analysis and modelling of statistics and predictor variables, the research then assessed the likely accuracy of the different approaches.

In order to produce a data set that would underpin maps of home energy efficiency standards, two of the options assessed were excluded due to concerns about accuracy of the modelled results.

The first rejected option – modelling energy efficiency from Experian data (**Option 2**) – can provide address level energy efficiency values from which to aggregate energy efficiency to different small area levels. However, the results would all be based on modelled data, including the prediction of intermediate variables. Comparing these with EPC data at different geographical levels (address level to regional), reveals a significant variety in the accuracy of these modelled values. This variation in accuracy is sufficiency large and unpredictable to prevent confidence in a resulting map produced from the data.

The second rejected option – predicting energy efficiency from Census 2011 data (**Option 4**) – similarly yields unsatisfactory prediction accuracy of energy efficiency ratings. Simplifying the prediction model to determine whether dwellings have EPC ratings above or below band D increases the prediction accuracy to a certain extent but still has significant limitations. Our exploratory model could only predict purpose-built flats as having an EPC band of C or above. Therefore, this option was also ruled out.

Two of the options – **Option 1** and **Option 3** – are likely to have high costs but are worth further consideration. First and foremost, would be the direct use of EPC data (**Option 1**). This would use actual EPC records from the EBR. The two limitations of this approach are:

1. the coverage of data, which is approximately 50 per cent of all dwellings in Great Britain, and the resulting uncertainties regarding the ability for these records to represent all dwellings
2. the costs of purchasing the data, which would be in excess of £120,000 for all of Great Britain. However, if the government was to make this data available free of charge to bona fide research bodies and local authorities, as is already the case in Scotland, this would make this option much more viable.

Nevertheless, this is an attractive option as it uses EPC data collected at address level. Furthermore, geographically representing the coverage of EPCs at small area level would represent a useful subset of information on a map.

A workable option to increase the coverage of EPC to all addresses would involve the purchase of address level household data from Experian and then matching addresses without EPC records to those with property characteristics similar to those with EPC records (**Option 3**). A version of this method is used by the Home Analytics tool in Scotland and is also likely to be used in the NEED

database. While this is likely to increase the accuracy of the data, it comes at increased cost and lack of certainty over the increased level of accuracy provided.

Finally, the DCLG may agree to a simple request to make aggregated data from the EPC register available to research bodies, or should it fail to respond to such a request, via an FOI request (**Option 6**). This would substantially reduce the purchasing costs as well as bypassing some of the data processing required to aggregate the data to the desired geography. However, there is no certainty that this attempt would be successful.

Supplementary analysis of the NEED data (performed by DECC) appears to be a useful source of information to produce maps (**Option 5**). While exploration of the publicly available NEED EUL file revealed that there was insufficient geographical information in the data set, a separate request to DECC as part of this research resulted in the publication of energy efficiency ratings at LSOA level.[35] This data set contains the number of properties in each EPC band for each LSOA in England and Wales. This is produced from the weighted EUL file comprising 4 million records representing all housing in England and Wales.

There are four main caveats with this data set:

1. The energy efficiency data is based on 4 million EPC records which have been weighted to represent all EPC data in England.
2. The weighting used in the EUL files was calculated at regional rather than LSOA level, meaning that LSOA level data should be treated with caution.
3. Some LSOAs have data missing – in total there are 800,000 records omitted due to disclosure concerns and statistical limitations.
4. The data only contains information from EPC records collected up to the end of October 2012.

Nevertheless, the data can be aggregated to higher levels of geography easily, which reduces some of the limitations of using LSOA level data and has a good match with national statistics for EPC data produced from the EUL file. Furthermore, **it is has been shown that this data set can be used to produce maps of energy efficiency data at parliamentary constituency level** in England and Wales at modest cost. This data should be used to show the proportion rather than the total number of dwellings in each band.

If cost were not a barrier then the analysis suggests that a method that combined EPC data and an Experian household data set is the preferred, but most expensive option (£165,000–£180,000, see Table 4.18). This would include Scotland data and enable the production of statistics and maps for the whole of Great Britain. However, even the less costly option of using only EPC data is likely to be prohibitively expensive for many organisations. One option for substantially reducing these costs would involve the DCLG providing free access to the EPC data under a researcher licence. It is not clear whether this is yet possible as requests made during this course of this research were not successful, but it may be an option in the future and should continue to be investigated.

If free access to EPC data is not possible under a researcher licence, a possible compromise could involve the purchase of EPC data at a smaller geographical area, for example, an English region. This option and its cost implications are discussed in Appendix D.

[35] www.gov.uk/government/statistics/national-energy-efficiency-data-framework-need-ad-hoc-requests-2014

In summary, the research suggests that it is technically possible to produce a reasonably accurate EPC database at parliamentary constituency level at modest cost through use of DECC's new NEED database, from which maps could be produced. There are some limitations to this option, such as the data being slightly out of date. A more accurate database could be produced through purchase of the EPC register and the use of Experian data to model ratings for homes without EPC values. However, this option entails considerable expense. This cost would be reduced considerably should the DCLG agree to make the data available free under a researchers licence. This is already effectively the case in Scotland. Should DCLG not agree to provide the data free, consideration should be given to buying EPC and Experian data for London only (see Appendix D for more information on this option). This would demonstrate the value of the exercise and help make the case for extending it to the whole of Great Britain. Meanwhile, continued attempts should be made to allow researchers access to EPC data under a specific license.

Appendix A: EPC data supplementary information

Appendix Table A: List of data available in the EPC data packs

EPC DATA PACK VARIABLES
SMALL DATA PACK
address of the property (including the postcode);
energy performance indicator;
potential energy performance indicator;
energy efficiency;
property type;
inspection date;
local authority area (individually or in groups);
constituency;
county; and,
date on which the data were entered onto the register.
MEDIUM DATA PACK ALSO INCLUDES
transaction type;
current environmental impact;
potential environmental impact;
current energy consumption;
potential energy consumption;
current CO2 emissions;
current CO2 emissions per floor area;
potential CO2 emissions;
current lighting cost;
potential lighting cost;
current heating cost;
potential heating cost;
current hot water cost;
potential hot water cost;
total floor area;
energy tariff;
whether the property is connected to the gas network;
floor level;
whether the property is on the top floor (for flats);
if the property is a flat not on the top floor the storey count for the property;
overall dimensions and size of the building;
type of heating controls;
proportion of windows which are multi-glazed;
types of glazing and area glazed;
number of extensions to the property;
number of habitable rooms;
number of heated rooms;
proportion of low energy light bulbs; and
number of open fireplaces

A **large data pack** means a set of data containing everything in the small and medium data packs together with additional data entered in order to register the certificates.

Organisations eligible to handle EPC data:

- either House of Parliament (or a member or official thereof);
- a government department;
- a person or body to which section 6 of the National Audit Act 1983 applies;
- the Welsh Ministers;
- the National Assembly for Wales
- the Scottish Parliament, the Scottish Executive, an office-holder in the Scottish Administration or a Scottish Public Authority;
- the Northern Ireland Assembly or a Northern Ireland Department;
- a county council, district council or parish council in England;
- a London borough council;
- the Common Council of the City of London;
- the Greater London Authority;
- a local authority within the meaning of section 61 of the Local Government in Scotland Act 2003;
- a local authority within the meaning of section 175 of the Local Government (Wales) Measure 2011;
- a council of a district or borough within the meaning of the Local Government Act (Northern Ireland) 1972;
- a university, including a university college (or any institution in the nature of a college in a university) or a school or hall of a university;
- a provider of higher education as defined in section 579(1) of the Education Act 1996;
- a provider of further education as defined in section 2(3) to (5) of the Education Act 1996;
- a charity (as defined by the Charities Act 2011) whose purposes include—
    - the conduct of research into the earth's climate or environment or research into the construction, design or use of buildings, or
    - the promotion of energy efficiency in buildings;
- a registered provider of social housing;
- a person who operates an accreditation scheme approved by the Secretary of State under regulation 22;
- a Green Deal relevant person;
- the holder of a licence under section 6(1)(c) or (d) of the Electricity Act 1989 as modified under section 41(1) of the Energy Act 2008;
- a person (other than a natural person) certificated under the Microgeneration Certification Scheme.

Prescribed purposes for using EPC (a detailed intended use must be supplied when applying for data)

- The prescribed purposes are those set out at Part 1 of Schedule 1 of the Energy Performance of Buildings (England and Wales) Regulations 2012, namely:
- Promoting energy efficiency improvements (as defined in section 2 of the Energy Act 2011) ("energy efficiency improvements") in relation to buildings (in a Green Deal plan);
- Conducting research into, or developing or analysing policy (or policy proposals) in relation to, the energy efficiency of buildings;
- Conducting research into the effectiveness or impact of energy efficiency improvements;
- Identifying geographic areas where the energy efficiency of buildings is low relative to other areas, or conducting research into the extent, causes or consequences of such lower levels of efficiency;
- Promoting and marketing energy efficiency improvements that may be made pursuant to a Green Deal plan;
- Identifying and analysing the impact of carbon emissions on the environment resulting from buildings with low levels of energy efficiency;
- Determining whether energy efficiency improvements that may be made pursuant to a Green Deal plan have or have not been made in respect of a particular building or buildings.

Appendix B: NEED variables

Appendix Table B: List of variables available in the NEED EUL file

Variable
Household ID
Former Government Office Regions
Index of multiple deprivation IMD 2010 for England, quantiles
Welsh Index of multiple deprivation 2011
Fuel Poverty Indicator
Date of EPC inspection
Annual gas consumption for 2005-2012 in kWh (by year)
Annual electricity consumption for 2005-2012 in kWh (by year)
Whether households with Economy 7 electricity meters
Main heating fuel
Age of construction of the property
Type of property
Floor area band
Energy efficiency band
Depth of loft insulation
Wall construction
Whether cavity wall insulation installed through a Government scheme
Year cavity wall insulation installed
Whether loft insulation installed through a Government scheme
Year of loft insulation installed
Whether boiler installed in property
Year of boiler installation
Weighting variable based on region, property age, property type and floor area

Appendix C: Home Analytics v3.0 Data

The table below lists the data variables available in Home Analytics v3

Appendix Table C: Data variables available in Home Analytics v3

Variable
UPRN, Address and Location
Property tenure
Main fuel type
Property age of construction
Boiler efficiency band (A- G)
Wall construction
Loft insulation
Energy Efficiency Rating band (EPC)
Modelled RdSAP fuel bill (£ per annum)
HTTC - Exposure to wind driven rain zone
Listed Building grade
Distance to nearest biofuels supplier
Solar PV potential output
Distance to nearest dwelling
ECO CSCO eligibility
Likelihood of income deprivation
High fuel cost, low income indicator
Property type
Number of habitable rooms
Total floor area (m2)
Number of dwellings within building (blocks of flats)
Wall insulation
Glazing type
Environmental Impact Rating band (EPC)
Modelled RdSAP CO2emissions (tCO2 per annum)
Hard to Treat Cavity (HTTC) wall risk
Distance to nearest gas meter
Roof orientation (E, SE, S, SW, W)
Wind speed at 10m above ground level
Garden size
ECO HHCRO eligibility
ECO CSCO Rural eligibility
Fuel bill (percentage of median)
High fuel cost, low income, vulnerable circumstances indicator

Appendix D: Regional mapping option: London

The cost of producing a useful and accurate map for the entire country is likely to be prohibitively expensive to produce, in one attempt at least. These costs are almost entirely associated with the construction of a statistically reliable and accurate data set to sit behind the mapping tool, and are predominantly data purchasing costs (although additional costs of note are likely to also include data cleaning, address matching and energy efficiency modelling).

Therefore, an alternative suggestion could be to consider a smaller region in the UK and purchase corresponding EPC data. If budgetary constraints allow, this could be further supplemented by Experian data to ensure full address coverage of housing in the area.

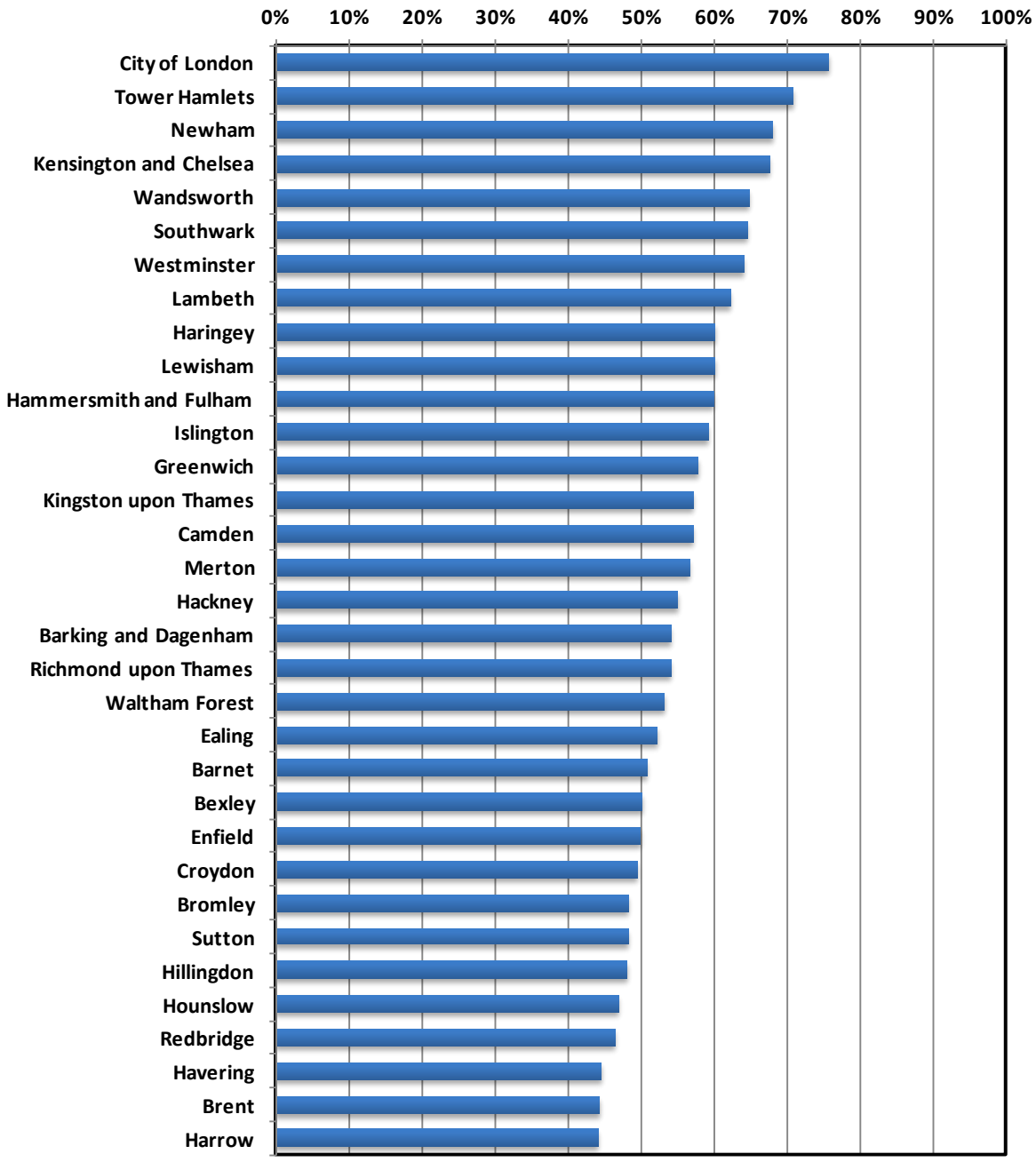
For a number of reasons, it appears to make sense that this smaller area is London. According to the stats on domestic EPC data published by DCLG, and as Appendix Table D shows, London has the highest coverage of all the English regions and is the only region to have EPCs for more than half the estimated housing stock (assuming a low number of duplicate EPC records).

Appendix Table D: EPC lodgements, estimated dwelling stock and EPC coverage by English region (Source: DCLG, ONS)

Region	Number of Lodgements	Estimated Total Dwelling Stock	percentage EPCs coverage
London	1,859,535	3,358,180	55%
South East	1,859,300	3,694,420	50%
East Midlands	989,878	1,971,520	50%
Yorkshire and The Humber	1,140,401	2,319,920	49%
North West	1,514,946	3,143,930	48%
East of England	1,217,701	2,531,910	48%
West Midlands	1,112,940	2,376,730	47%
South West	1,090,102	2,401,300	45%
North East	506,872	1,178,280	43%

Mapping London is therefore, likely to represent the most accurate representation of the housing stock. In addition, the relatively more uniform nature of the housing stock in London combined with its urban location mean that modelling missing EPC data for additional addresses using Experian data could achieve higher levels of accuracy than in other regions or devolved nations of Great Britain.

The coverage of EPCs by London Borough is shown in Figure 4.15. In general, the data suggests that the more central the Borough the higher the proportion of houses that have received energy assessments and resulting EPC certificates. (A map of London Boroughs is provided below in Appendix Figure B). The coverage in each borough ranges from 44% in Harrow to 76% in the City of London. All but seven London boroughs have more than 50% of dwellings with EPCs (assuming low levels of duplicate records).



Appendix Figure A: Percentage EPC coverage of the dwelling stock by London Borough (Source: DCLG, ONS)

5.1.1 Cost of building a London energy efficiency data set

The options for this data set are to use EPC data only, or to base the data on an address level data set using Experian data, matching EPC records to this data set and then predicting the EPC ratings of the properties without EPC records. The costs of this therefore vary to a certain extent and typical costs are provided in Appendix Table E. Estimated costs for EPC data were based on the supply of all available EPC lodgements in London (as specified in the DCLG live tables in December 2014).

Experian gave a direct quote for the provision of its data. Overall the cheapest option would cost between £18,000 and £19,000 (for EPC data only), while use of the two data sets could incur costs of over £42,000.

Appendix Figure B: Map of London Boroughs (Source: No Second Night Out[36])

Appendix Table E: Data purchasing costs for EPC data and Experian data for London

Data	Number of records	Min Costs	Max costs	Comments
EPC data	1,859,535	£18,595	£18,595	These could reduce slightly depending on the number of duplicated EPC records.
Experian data	3,358,180	£18,000	£23,500	Dependent on whether mosaic is purchased for low income household ID
Total data costs		**£36,595**	**£42,095**	

[36] http://www.nosecondnightout.org.uk/wp-content/uploads/2013/02/Outreach-map-for-NSNO-Feb-2013.jpg

Further costs relating to the main tasks required to process, model and complete an address level energy efficiency data set are listed (and costs estimated) in Appendix Table F. The main tasks are:

- Experian data cleaning
- EPC Data cleaning
- Address matching EPC records to Experian data
- Modelling EPCs for remaining Experian addresses without EPC lodgements
- Experian/EPC data matching to ONSPD (by postcode)

If only EPC data is used, then only EPC data processing and subsequent matching to ONSPD data is required. The costs provided here for staff resource should be considered purely indicative and likely to vary considerably based on the organisation and staffing expertise utilised. However, some indication of the scale of the data processing and matching was considered important in terms of comparing the costs of data processing to the costs of data purchasing[37].

Appendix Table F: Additional data tasks and staff resource required

Task	Number of records	Estimated number of staff days required[38]	Estimated staff costs based on a rate of £500/day	Task descriptions
Experian Data cleaning	3,358,180	2-3	£1,000-£1,500	e.g. Removing duplicates
EPC data cleaning[39]	1,859,535	2-3	£1,000-£1,500	e.g. Removing duplicates
Experian to EPC address matching	3,358,180	4-6	£2,000-£3,000	Joining both data sets together
Modelling EPCs for remaining Experian addresses	1,498,645	3-5	£1,500-£2,500	Assigning a modelled EPC band to Experian records without any corresponding EPC data or where address matching has failed.
Experian/EPC[39] data to ONSPD matching (postcode)	3,358,180	2-3	£1,000-£1,500	Joining additional area level information by postcode from ONS Postcode Directory
Total		**11-17 days**	**£5,500 – £8,500**	

[37] Should CSE be request to complete a task similar to that specified here, the project specification would be subject to a more thorough project plan and budget and costs are likely to differ to those stated.
[38] Depending on the expertise level of the staff used and the exact nature of the data processing requirements
[39] These are the minimum tasks that would be required if only EPC data were to be used.