Department
of Energy &
Climate Change

National Energy Efficiency Data-Framework

Anonymised dataset accompanying documentation

July 2014

URN 14D/267

Any enquiries regarding this publication should be sent to us at
EnergyEfficiency.Stats@decc.gsi.gov.uk.

Contents

1. Executive summary

DECC has published two datasets containing property level data from the National Energy Efficiency Data-Framework (NEED):

1. **Public Use File (PUF):** A sample of 49,815 records selected to be representative of the housing stock and available to all via the DECC website and data.gov.uk: https://www.gov.uk/government/publications/national-energy-efficiency-data-framework-need-anonymised-data-2014.

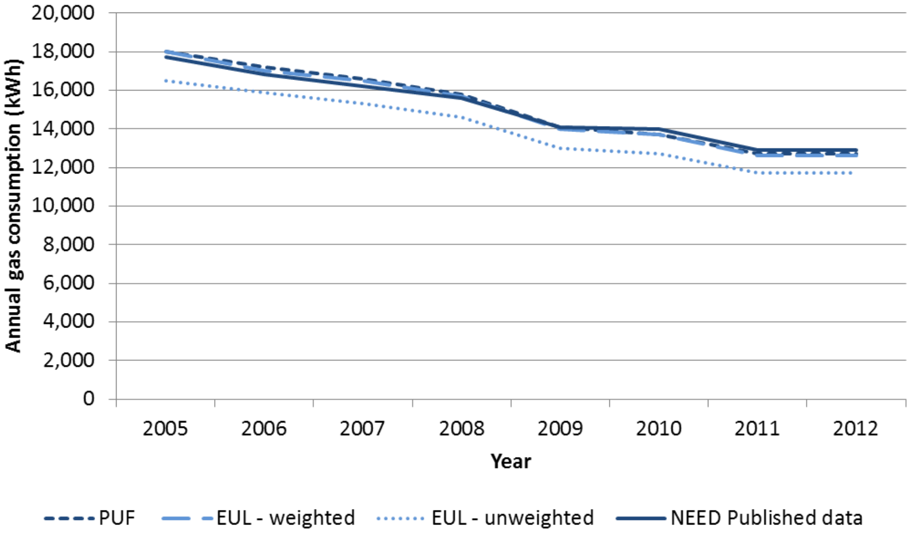
2. **End User Licence File (EUL):** The end user licence file is a sample of 4 million records (4,086,448). This dataset is available from the UK Data Archive under an end user licence: http://discover.ukdataservice.ac.uk/catalogue/?sn=7518.

The two datasets are based on samples of properties which have had an energy performance certificate. Both datasets contain general variables as well as information on gas and electricity consumption and energy efficiency of the property.

The content and format of the datasets has been determined following engagement with users including a consultation. The datasets have been anonymised to prevent identification of a specific property. This process included: removal of household identifiers and detail geographic information; banding of variables; sample selection; and a small amount of record swapping.

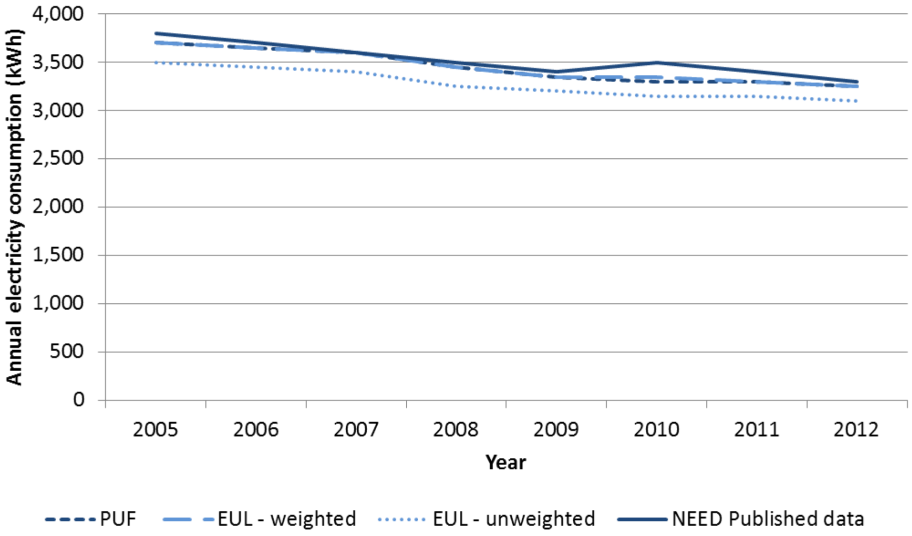
The anonymised datasets use data from Energy Performance Certificates (EPCs) as the source of property attribute data, rather than Valuation Office Agency property attribute data which is used in NEED analysis undertaken by DECC. This has advantages, as it allows for additional variables such as energy efficiency band to be included in the datasets. However, as EPCs are required for very specific groups of properties, the data are not representative of the housing stock in England and Wales. The PUF has been selected as a stratified random sample in order to address this bias. It was not possible to select a fully stratified sample for the EUL dataset, so a weighting variable has been included with the dataset to attempt to address the remaining bias. Figures 1.1 and 1.2 show DECC's published estimates from NEED[1] alongside estimates from the EUL and PUF.

Figure 1.1: Median annual gas consumption, 2005 to 2012



[1] https://www.gov.uk/government/publications/national-energy-efficiency-data-framework-need-report-summary-of-analysis-2014

Figure 1.2: Median annual electricity consumption, 2005 to 2012



The figures show that the typical consumption based on the PUF and EUL weighted data are consistent with the NEED headline results, for both gas and electricity in all years from 2005 to 2012. The lower results seen for the EUL unweighted data reflect the bias in the EPC dataset, which contains a higher proportion of more energy efficiency properties and therefore has a lower typical consumption[2].

It is hoped that the publication of these datasets will increase the value of NEED. DECC would welcome information on how the datasets are being used.

For any queries or feedback on the published datasets please contact DECC by email: EnergyEfficiency.Stats@decc.gsi.gov.uk.

[2] The EPC data contains a higher proportion of newer properties, and more flats. See section 4 for more details.

2. Introduction

The information held by the Department of Energy and Climate Change (DECC) as part of the National Energy Efficiency Data-Framework (NEED) is a valuable resource for researchers looking at energy efficiency and energy consumption in households.

The UK helped secure the G8's Open Data Charter[3], which establishes the presumption that the data held by governments will be publicly available, unless there is good reason to withhold it. As part of this commitment to Open Data, DECC has published anonymised data[4] from NEED.

Two dataset containing household level data have been published:

3. **Public Use File (PUF):** A sample of 49,815 records selected to be representative of the housing stock (based on region, property age, property type and floor area band). This dataset is available to all via the NEED pages of the Government website and data.gov.uk: https://www.gov.uk/government/publications/national-energy-efficiency-data-framework-need-anonymised-data-2014

4. **End User Licence File (EUL):** The end user licence file is a sample of 4 million records (4,086,448). This dataset is available from the UK Data Archive under an end user licence: http://discover.ukdataservice.ac.uk/catalogue/?sn=7518.

This approach to publication of two datasets with different content and different access requirements is in line with ICO guidance, and supported by the ICO as it "allows the measures taken to protect individuals' privacy to be tailored to each dataset, bearing in mind the purpose for which each dataset is released, who is likely to use them and the different levels of risk to individuals' privacy"[5].

Both datasets are made up of samples of domestic properties in England and Wales and contain:

- General variables (e.g. region);
- Gas and electricity consumption variables; and
- Property variables including energy efficiency measures installed.

The content and format of these two datasets has been determined following a consultation. DECC used feedback from NEED users to inform the proposals set out in the consultation, including feedback received from a seminar with energy suppliers and an event held for NEED users. The consultation sought views on the proposals, including the content of the dataset and approach to anonymisation and publication. The consultation document and Government's response are available here: https://www.gov.uk/government/consultations/national-energy-efficiency-data-framework-making-data-available.

[3] https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex

[4] Anonymised data are data relating to a specific individual or property where the identifiers have been removed to prevent identification of that individual or property (directly or indirectly).

[5] http://ico.org.uk/~/media/documents/consultation_responses/ICO-response-to-DECC-National-Energy-Efficiency-Data-Framework-consultation-on-anonymised-data.pdf

Background

NEED was set up in order to assist DECC in its business plan priority to “save energy with the Green Deal and support vulnerable consumers”.

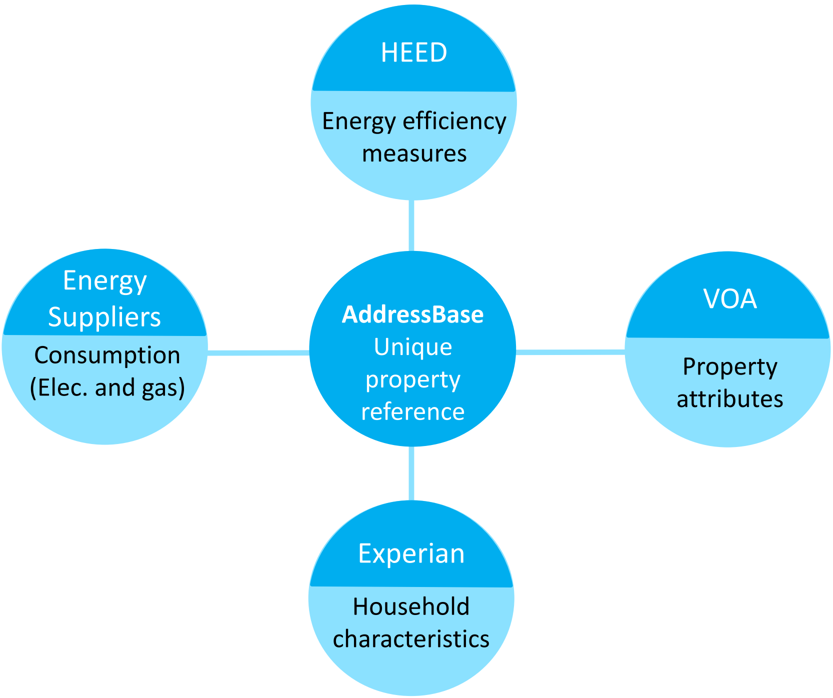
It is a key element of DECC’s evidence base supporting DECC to:

- develop, monitor and evaluate key policies (including the Green Deal);
- identify energy efficiency potential which sits outside the current policy framework;
- develop a greater understanding of the drivers of energy consumption; and
- gain a deeper understanding of the impacts of energy efficiency measures for households and businesses.

NEED enables this analysis by combining data from existing sources (administrative and commercial) at property level. The address information in each dataset is used to assign a unique property reference number (UPRN) to each record. Data from different sources can then be matched to each other via the UPRN (figure 2.1).

Four key data sources have been used in DECC’s analysis of domestic energy consumption and the impact of installing energy efficiency measures: meter point electricity and gas consumption data, Valuation Office Agency (VOA) property attribute data, the Homes Energy Efficiency Database (HEED) containing data on energy efficiency measures installed, and data modelled by Experian on household characteristics.

Figure 2.1: Structure of domestic NEED


HEED
Energy efficiency measures
Energy Suppliers
Consumption (Elec. and gas)
AddressBase
Unique property reference
VOA
Property attributes
Experian
Household characteristics

While the proposed anonymised datasets are primarily based on data currently used in NEED, it has not been possible to use all the data sources which form NEED. Property attribute data collected by the VOA and data from Experian (covering household characteristics) cannot be used due to legal and contractual restrictions.

DECC has instead used information from Energy Performance Certificates (EPCs) for information on property attributes. Use of EPC data has a number of advantages compared to

VOA data, including the potential to use relevant information which is not available from VOA, such as Energy Efficiency Band. However, EPC data coverage is not as comprehensive as VOA data and there is more uncertainty about the quality of the data.

No information is available on household characteristics, but indicators assigned to a property based on the geographic location of the property have been included. For example the index of multiple deprivation is assigned based on the Lower Layer Super Output Area (LSOA) of the property. Fuel poverty indicator is assigned in the same way.

All DECC outputs from NEED, including latest results and further information on creation of the dataset and methodology can be found on the Government website at: https://www.gov.uk/government/collections/national-energy-efficiency-data-need-framework.

Further information on how the anonymised dataset was created, including the content and approach to anonymisation is set out in section 3 of this report. Section 4 shows how results from the anonymised datasets compare with data from other sources, including DECC published official statistics from NEED.

For any queries or feedback on this publication email: EnergyEfficiency.Stats@decc.gsi.gov.uk.

3. Creation, content and anonymisation

This section outlines how the anonymised datasets were created, the variables included in the datasets and the approach taken to anonymisation of the datasets.

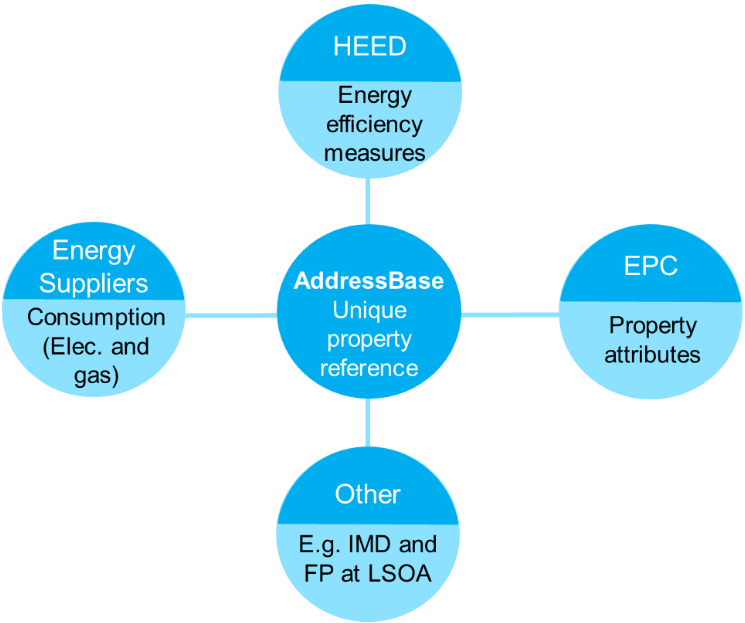
Creation

The two datasets published are both samples taken from a larger dataset of properties for which an EPC record could be matched to consumption data ("matched EPC dataset"). This dataset was created in the same way as the NEED sample used for DECC analysis, except EPC data have been used instead of VOA property attribute data, and no household characteristics data are included.

EPCs were first introduced in England and Wales in 2007. Certificates are needed whenever a property is built, sold or rented. EPCs are now also required prior to a property having a measure through the Green Deal[6], Renewable Heat Incentive (RHI) or Feed in Tariffs (FiTs)[7]. The EPC data included in the datasets are based on an extract from the EPC register for EPCs lodged up to October 2012. This has been used to be consistent with the most recent consumption data in NEED; annual data for 2012. By the end of the third quarter of 2012 there had been 8.1 million EPCs lodged[8].

In order to assign a UPRN to each property, the EPC dataset was matched to AddressBase based on the address information provided in the dataset. These data were then linked with information on electricity and gas consumption and energy efficiency measures installed in each property using the UPRN. In addition, information on index of multiple deprivation and fuel poverty indicator were assigned to each property based on the LSOA the property is located in (figure 3.1).

Figure 3.1: Structure of NEED anonymised dataset



[6] https://www.gov.uk/government/policies/helping-households-to-cut-their-energy-bills

[7] https://www.gov.uk/government/policies/increasing-the-use-of-low-carbon-technologies

[8] Official Statistics on Energy Performance Certificates are published by DCLG at: https://www.gov.uk/government/collections/energy-performance-of-buildings-certificates.

As EPCs are required for very specific groups of properties the matched EPC dataset is not representative of the housing stock in England and Wales. For example, properties built since 2008 are more likely to be included in the dataset, and there are a higher proportion of more energy efficiency properties in the EPC data than in the England and Wales dwelling stock. More information on the impact of this bias and how it has been addressed in the creation of the datasets is provided in the relevant sections below.

Prior to selecting the samples for the PUF and EUL, all records with missing information in any of the key property attributes[9] were excluded from the dataset. Any record without valid electricity[10] consumption in 2012 was also excluded.

Public use file (PUF)

The public use file is a stratified random sample of 49,815 properties, drawn from the matched EPC dataset. The dataset was selected to be representative of the England and Wales housing stock, stratified by region, property type, property age[11] and floor area band. Population totals for England and Wales from VOA were used to determine the size of each stratum.

End user licence file (EUL)

The creation of the EUL file is very similar to the PUF. The matched dataset which the sample is selected from is the same. However, due to the bias in the matched EPC dataset and the desire to create a large file for detailed analysis, it was not possible to select a stratified random sample which would accurately reflect the dwelling stock in England and Wales.

Rather than selecting an entirely random sample from the matched EPC dataset, properties were selected based on their frequency in dataset relative to the total dwelling stock. For example, there are a much higher proportion of smaller modern flats in the matched EPC dataset then there are in the population as a whole, so only a relatively small proportion of those that appeared in the matched EPC dataset were selected for the EUL file. By contrast, older detached properties are less likely to be in the matched EPC dataset, so a higher proportion of those that do appear have been selected. This does not entirely remove the bias from the EPC dataset, but does ensure that properties with less common combinations of property attributes are not under represented more than necessary and can therefore be used for analysis.

In order to correct for the bias further, a weighting variable has been included in the published EUL dataset. The weighting variable was determined using the same variables used in the stratification of the PUF[9]. It allows for estimates for the England and Wales dwelling stock to be produced. The impact of this variable on results is shown in more detail in section 4.

Content

Table 3.1, sets out the variables included in the datasets and the source of each of these variables. Three variables are only included in the EUL file. Look-up tables and formats of variables are published separately on the government website[12].

[9] Property Age, Property Type, Floor Area Band and Energy Efficiency Band.

[10] Electricity consumption is assumed to be valid in NEED if it is between 100kWh and 25,000 kWh (inclusive). In DECC's NEED analysis suspected estimated readings are also excluded, these have not been excluded in the anonymised datasets, but made up approximately 2 per cent of all electricity meters in 2012.

[11] Only three property age bands were used in the stratification so that bands on the EPC and VOA datasets matched; pre-1930, 1930-1982 and 1983 or later.

[12] https://www.gov.uk/government/publications/national-energy-efficiency-data-framework-need-anonymised-data-2014

Table 3.1

Variable	Description	Source
HH_ID	Household identifier. This identifier has been created specifically for these datasets and has no relation to any identification number from the original data.	DECC
REGION	Former Government Office Regions (GORs) in England, and Wales.	National Statistics Postcode Lookup http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/administrative/england/government-office-regions/index.html
IMD_ENG	Index of multiple deprivation (IMD) 2010 for England. Households are allocated to five groups (quintiles) based on the deprivation rank of the 2001 Lower Layer Super Output Area (LSOA) they are located in. Households in the 20 per cent most deprived LSOAs are in the bottom quintile (1) and least deprived are in the top quintile (5).	Department for Communities and Local Government (DCLG) https://www.gov.uk/government/collections/english-indices-of-deprivation
IMD_WALES	Welsh Index of multiple deprivation 2011. Households are allocated to one of five bands based on the deprivation rank of the LSOA (2001) they are located in. One is most deprived and five is least deprived.	Welsh Assembly Government http://wales.gov.uk/statistics-and-research/welsh-index-multiple-deprivation/?lang=en
FP_ENG	**EUL only.** Fuel Poverty Indicator. Households are allocated to one of five bands based on the estimate of the proportion of households in fuel poverty in the LSOA they are located in. Uses the 2011 low income high cost definition of fuel poverty.	DECC https://www.gov.uk/government/publications/2011-sub-regional-fuel-poverty-data-low-income-high-costs-indicator
EPC_INS_DATE	**EUL Only.** Provides information on the date of the EPC inspection (grouped by pre-2010 and 2010 or later).	Landmark/DCLG, EPC https://www.gov.uk/buy-sell-your-home/energy-performance-certificates
GconsYEAR	Annual gas consumption in kWh.	Xoserve and Independent Gas Transporters https://www.gov.uk/government/publications/regional-energy-data-guidance-note
GconsYEARValid	Flag indicating households with valid gas consumption (V), households off the gas network (O) and invalid consumption.	DECC – using combination of gas consumption data and EPC off gas data.
EconsYEAR	Annual electricity consumption in kWh.	Industry data aggregators

Variable	Description	Source
		https://www.gov.uk/government/publications/regional-energy-data-guidance-note
EconsYEARValid	Flag indicating households with valid electricity consumption.	DECC – using electricity consumption data
E7Flag2012	Flag showing households with Economy 7 (profile 2) electricity meters in 2012.	Industry data aggregators
MAIN_HEAT_FUEL	Description of main heating fuel (gas or other).	EPC
PROP_AGE	Age of construction of property (six bands).	EPC
PROP_TYPE	Type of property (e.g. detached, semi-detached).	EPC
FLOOR_AREA_BAND	Floor area band.	EPC
EE_BAND	Energy Efficiency Band (A and B grouped).	EPC
LOFT_DEPTH	Depth of loft insulation (150mm or more, or less than 150 mm).	EPC and HEED
WALL_CONS	Wall construction (cavity wall or other).	EPC
CWI	Cavity wall insulation installed through a Government scheme (includes EEC, CERT, CESP).	Homes Energy Efficiency Database (HEED) http://www.energysavingtrust.org.uk/Organisations/Government-and-local-programmes/Programmes-we-deliver/Homes-Energy-Efficiency-Database
CWI_YEAR	Year cavity wall insulation installed.	HEED
LI	Loft insulation installed through a Government scheme (includes EEC, CERT, CESP).	HEED
LI_YEAR	Year of loft insulation installed.	HEED
BOILER	Boiler installed in property (certified by Gas Safe or CORGI, or installed through a Government scheme).	CORGI/Gas Safe/HEED
BOILER_YEAR	Year of boiler installation.	CORGI/Gas Safe/HEED
WEIGHT	**EUL only.** Weighting based on region, property age, property type and floor area band.	DECC

Anonymisation

The Information Commissioner's Office (ICO) anonymisation code[13] states that "there is a clear legal authority that where an organisation converts personal data into an anonymised form and discloses it, this does not amount to a disclosure of personal data". Information on how the records have been anonymised is set out in this section and has been carried out with reference to relevant guidance from the ICO[13] and with input from members of the UK Anonymisation Network, including the Office for National Statistics.

The approach to anonymisation for the PUF and EUL datasets reflected the risk associated with the release of each of the datasets. The initial steps taken were the same in each case and were undertaken prior to the samples for each dataset being selected.

The main method used to anonymise the data was the removal of the household identifier and detailed geographic information.

The next stage in the anonymisation was applying banding or recoding to variables. The banding employed was determined based on the risk of disclosure. A number of "visible" variables were selected; these are variables which an intruder[14] would be most likely to use to identify a property in the data, such as floor area band and property type. If a property could be identified through these variables then non visible more sensitive information could be discovered. Analysis of these variables determined the required banding.

Potential disclosure problems were considered in tabulations, such as low counts or columns/rows with many zeros. Where potential disclosure problems arose, additional banding was applied to variables within the dataset to reduce the risk of disclosure. For example, a high proportion of occurrences of unique combinations of visible variables included properties with a floor area band of over 200 square metres. As a result, this floor area band was grouped with the band below, to give a top floor area band of over 150 square metres.

The analysis of visible variables, alongside knowledge of information available in the public domain and consultation responses on priorities, also informed the decision on which variables could safely be included in the dataset and the small number of variables which should be excluded. This led to the exclusion of two variables, environmental impact band and solid wall insulation installation.

The full matched EPC dataset was tested for disclosure by analysts from DECC and the Ministry of Justice. Individuals noted the approach taken when attempting to identify specific households in the dataset, which allowed an assessment of the most "dangerous" variables i.e. those that were most likely to allow an intruder to identify a household in the dataset. This allowed decisions to be made on any further banding required and identification of variables which should not be included in either one of the datasets.

Further details of the banding of variables and rationale are included in the consultation response[15].

The final stages of anonymisation and testing differed for each dataset.

[13] http://www.ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Practical_application/anonymisation_code.ashx

[14] Refers to a group or individual who wishes to identify people in the data or attributes relating to these people. Also known as an attacker they may or may not have malicious intent.

[15] https://www.gov.uk/government/consultations/national-energy-efficiency-data-framework-making-data-available

Public use file (PUF)

A lot of protection is offered for the PUF through the fact it is a sample of properties, made up of less than one per cent of all properties on the EPC register in October 2012.

This size of the sample alongside the banding outlined above leads to a dataset with a low risk of a property being identified. In addition to this, two variables included in the EUL file are not included in the PUF (EPC inspection date and fuel poverty indicator). These two variables could be used to support attempted identification of properties when combined with other variables in the dataset and publically available information. Although the risk of identification with these variables included is low, the exclusion of these two variables - which did not receive strong support to be included in the PUF in the consultation responses - avoids unnecessary risk.

Testing of the PUF was undertaken by Southampton University and did not lead to the identification of any households in the dataset.

End user licence file (EUL)

In addition to the approach to anonymisation outlined above, the EUL file has some protection through the end user licence required to be signed by users before gaining access to the dataset. However, it is still important to limit the risk of disclosure and two further methods were employed to support this.

Some record swapping was used to alleviate the risk of disclosure from a small number of properties with less common combinations of attributes. In a small number of cases, the region associated with two records which were similar in all other aspects was swapped. Efforts were made to limit any damage to the dataset and all other information relating to the property remained unchanged.

The sample for the EUL dataset was then selected from the full matched EPC dataset as described above. The EUL sample is made up of approximately half of all EPCs lodged by October 2012[16]. This offers further protection as an intruder will not know whether a property has been selected in the sample or not, or whether a property which is unique in the sample is unique in the population.

The two datasets created balance the desire to provide as much detailed information as possible to users with the need to minimise the risk of disclosure.

[16] Some records were dropped prior to sample selection, for example, records with missing or invalid information for some variables and records which could not be matched to valid electricity consumption in 2012.

4. Comparisons with other sources

This section outlines how the data from the published datasets compare with data from other sources, including NEED headline results.

As EPCs are required in a very specific group of properties, the dataset used to form the anonymised datasets is not fully representative of the population of dwellings in England and Wales. For example, properties built since 2008 are more likely to be included in the dataset, and there are a higher proportion of more energy efficiency properties in the EPC data.

The samples have been selected and weighted in order to address some of this bias. The results presented in this section show results for the samples and the impact of the weighting. It also shows the impact of the rounding applied to gas and electricity consumption in the published datasets.

Distribution of data

Energy Performance Certificate Data

Energy Performance Certificates (EPCs) were first introduced in England and Wales in 2007. Certificates are needed whenever a property is built, sold or rented. EPCs are now also required prior to a property having a measure through the Green Deal[17], Renewable Heat Incentive (RHI) or Feed in Tariffs (FiTs)[18].

An EPC contains:

- information about a property’s energy use and typical energy costs; and
- recommendations about how to reduce energy use and save money.

In order to produce this information, the Reduced Standard Assessment Procedure (RDSAP)[19] is used to assess and compare the energy and environmental performance of dwellings. This includes gathering information on physical characteristics of the property and the main heating fuel. RDSAP then assigns a score to a property based on how much energy a dwelling will consume given standard assumptions about occupancy and behaviour. It quantifies a dwelling’s performance in terms of an efficiency rating (the Energy Efficiency Rating). The energy efficiency ratings are grouped into bands from A (most efficient with lower running costs) to G (least efficient and higher running costs).

The data in the anonymised datasets are based on EPCs lodged in England and Wales up to October 2012, by which point 8.1 million EPCs had been lodged for dwellings[20]; at the end of March 2012 there were around 25 million dwellings in England and Wales[21]. Figure 4.1 shows the number of EPCs lodged in each year from 2008 to the end of 2013.

[17] https://www.gov.uk/government/policies/helping-households-to-cut-their-energy-bills

[18] https://www.gov.uk/government/policies/increasing-the-use-of-low-carbon-technologies

[19] https://www.gov.uk/standard-assessment-procedure.

[20] Some properties may have had more than one EPC during the period, therefore there will be fewer than 11.2 million properties with an EPC, by 27 April 2014 approximately 11.2 million EPCs had been lodged for dwellings.

[21] Live tables on dwelling stock published by the Department for Communities and Local Government: https://www.gov.uk/government/statistical-data-sets/live-tables-on-dwelling-stock-including-vacants.

Figure 4.1: Number of EPCs by year, for all dwellings

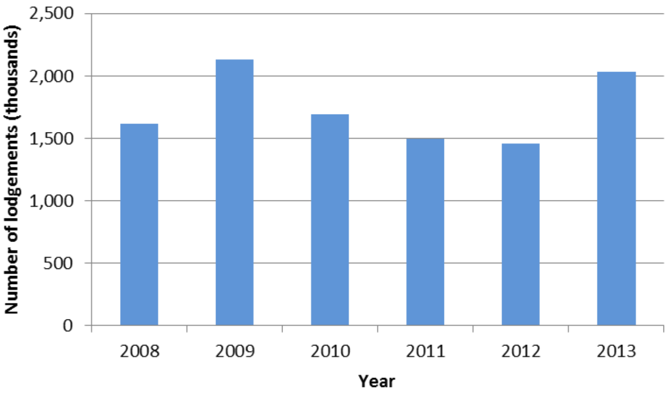
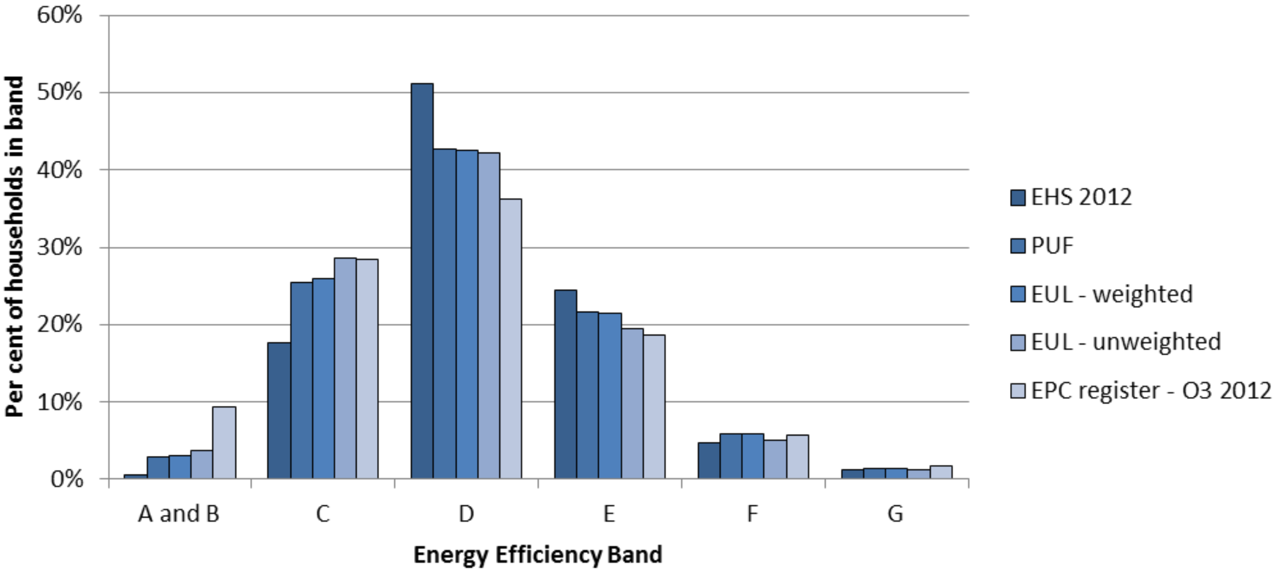


Figure 4.2 shows the energy efficiency band of properties in England, based on data from the English Housing Survey (EHS)[22], compared to the data for England and Wales in the anonymised datasets and for all EPCs lodged by October 2012. The most common energy efficiency band is D, with 11.6 million households or 51 per cent of all households in England in this band. The anonymised datasets slightly under represents properties in band D (42 and 43 per cent in England and Wales in the weighted EUL and PUF respectively). Fewer properties are in bands A and G, therefore results for these groups will be subject to greater uncertainty and should be treated with more caution. Bands A and B are over represented in the EPC register due to the properties which require an EPC[23] and this follows through into the EUL and PUF datasets.

Figure 4.2: Properties in England (and Wales) by energy efficiency band



The rest of this section shows how the distribution of records in the published datasets compare with the distribution in the NEED sample used for DECC's official estimates[24].

[22] https://www.gov.uk/government/publications/english-housing-survey-2012-to-2013-headline-report.

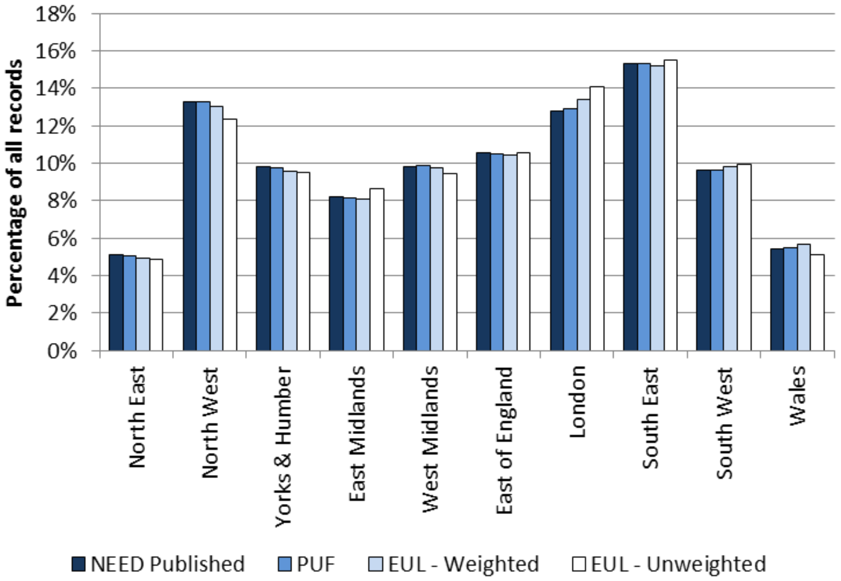
[23] For example, according to the EHS 2012, one per cent of properties in England are rated A or B. For properties which had an EPC (England and Wales) between 2007 and October 2012, nine per cent were in bands A or B.

[24] Distribution for NEED sample is based on records with valid electricity consumption in 2012 unless otherwise stated. Further information on the quality of the data used in NEED's headline estimates is available in Annex A: Quality Assurance to the June 2014 NEED publication https://www.gov.uk/government/publications/national-energy-efficiency-data-framework-need-report-summary-of-analysis-2014.

Region

The chart below shows the percentage of records in each dataset allocated to each region, including a comparison of the weighted and unweighted EUL.

Figure 4.3: Distribution of properties by region, NEED published data compared with anonymised datasets



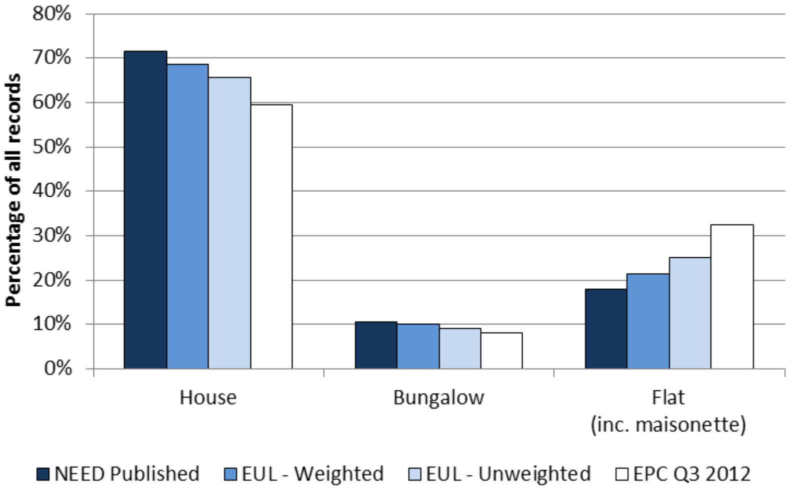
The distribution of records by region is similar for all datasets. The biggest differences are in London. The public use file has an additional 0.1 per cent of records assigned to London compared with the dataset used for DECC's NEED analysis. The equivalents for the EUL weighted and unweighted distributions are 0.6 and 1.3 per cent respectively.

The difference between the DECC NEED sample distribution and the anonymised datasets distributions occurs because of the stage the samples are selected. Due to access to data and the need to limit the transfer of data, the DECC NEED sample is selected prior to matching with consumption data, so some records are lost when the consumption data are matched. Flats are the most likely property type to be unmatched (due to the address information associated with flats) and as a result, the difference in approach to sample selection has the biggest impact on London, where flats are most common.

Property Type

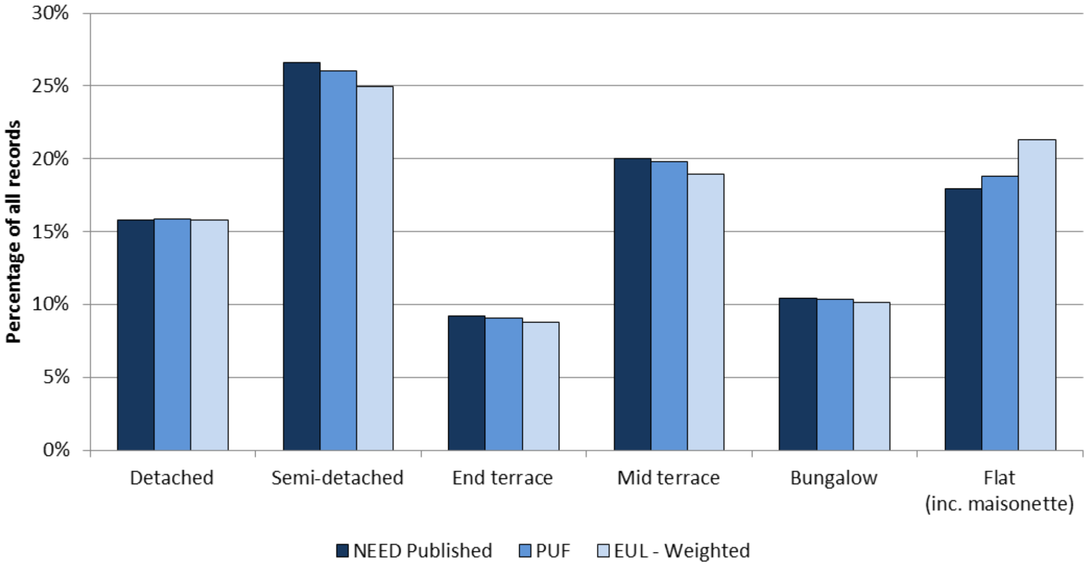
A closer look at the distribution by property type shows the bias towards flats in the EPC data more clearly, figure 4.4. It also shows the impact of the sample selection and weighting on the EUL dataset.

Figure 4.4: Distribution of properties by property type, NEED published data compared with EUL and EPC data



The figure shows that the dataset used for NEED publications is more likely to include houses and less likely to include flats compared with the anonymised dataset and the EPC quarter 3 2012 data. As outlined in section 3, the sample selection for the EUL attempts to help correct for some of the bias in the EPC data, this can be seen in figure 4.4. The weighting variable further reduces the difference relative to the data used for NEED official estimates. However, the issues with matching still have a small impact, meaning there is still a small difference between in distribution between the NEED published data and the EUL weighted results. Figure 4.5 shows the distribution for the NEED published data, the PUF and the weighted EUL.

Figure 4.5: Distribution of properties by property type, NEED published data compared with PUF and weighted EUL



It shows that the weighting and sample selection have largely removed the differences in the distribution, though some small differences remain, particularly for flats.

Consumption

This section shows how results from the PUF and EUL datasets compare with NEED published estimates of typical consumption. It also sets out the impact of weighting, rounding and sample selection in the anonymised datasets.

Figure 4.6 shows the mean and median annual electricity consumption for each of the three datasets for 2005 to 2012. Figure 4.7 shows the equivalent for gas.

Figure 4.6: Median and mean annual gas consumption, NEED published data compared with PUF and weighted EUL

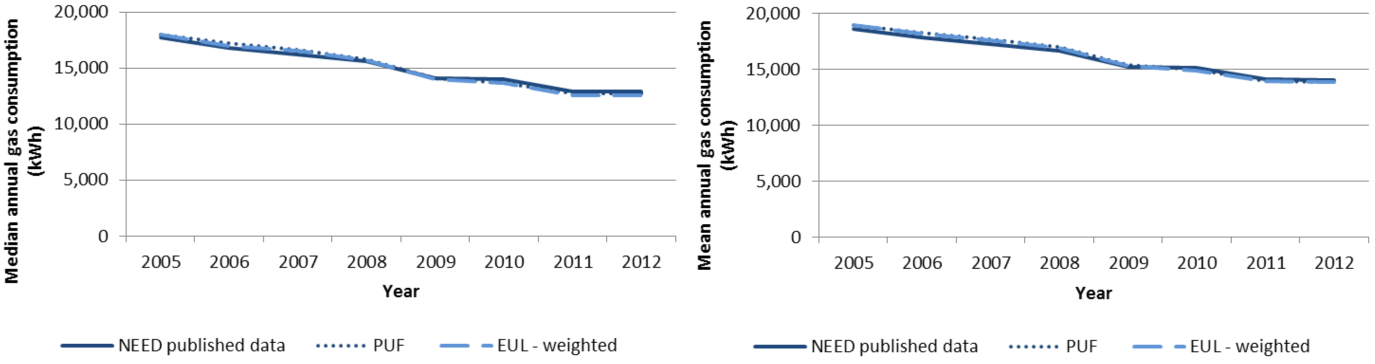
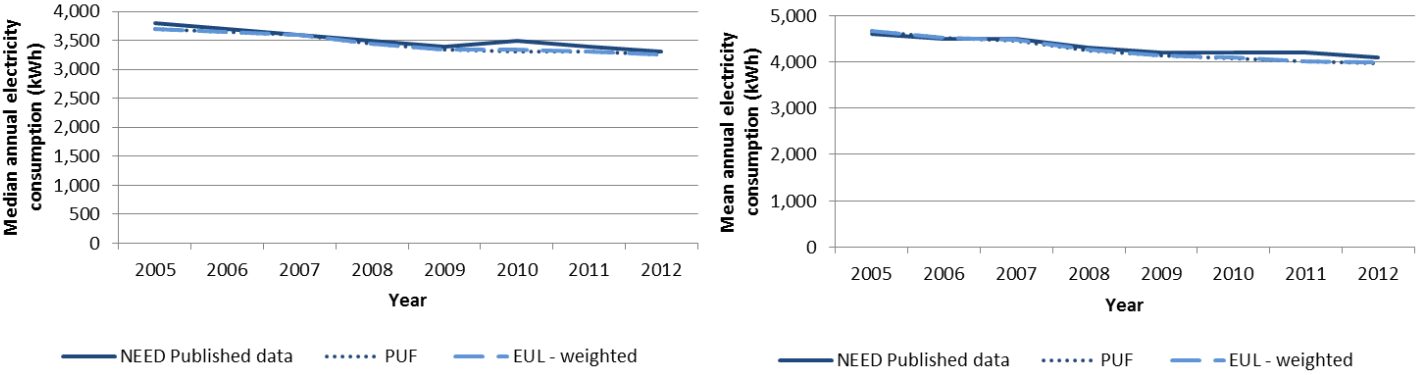


Figure 4.7: Median and mean annual electricity consumption, NEED published data compared with PUF and weighted EUL



In all cases, the mean and median consumption values are very similar. This indicates that the samples (when weighted in the case of the EUL) reflect the distribution of properties in the England and Wales housing stock well and that the rounding applied to the datasets is not distorting the estimates. Figures 4.8 and 4.9 show the impact of the rounding on the PUF in more detail.

Figure 4.8: Annual gas consumption, PUF rounded and unrounded data 2005 to 2012

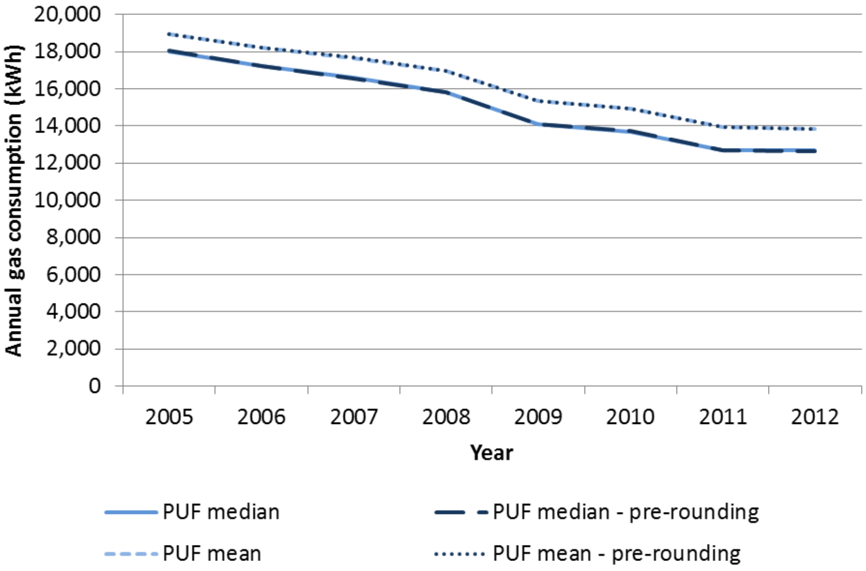
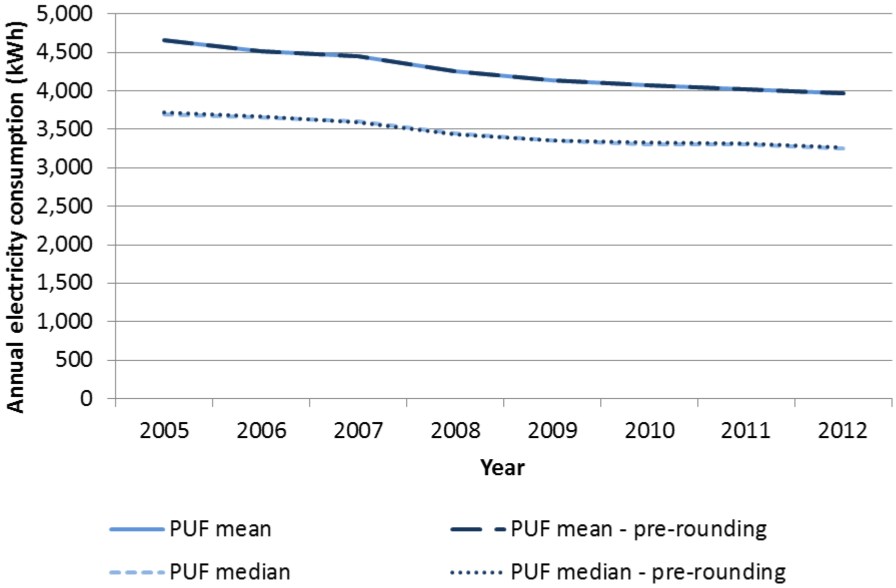


Figure 4.9: Annual electricity consumption, PUF rounded and unrounded data 2005 to 2012



The charts show that there is almost no difference between the results when calculating them using the rounded data published in the public use file compared to using the more detailed unrounded data. The impact on the end user licence dataset is similar, see figures 4.10 and 4.11. The impact of the weighting applied to the EUL dataset can also be seen from the charts.

Figure 4.10: Median annual gas consumption, EUL rounded and unrounded, 2005 to 2012

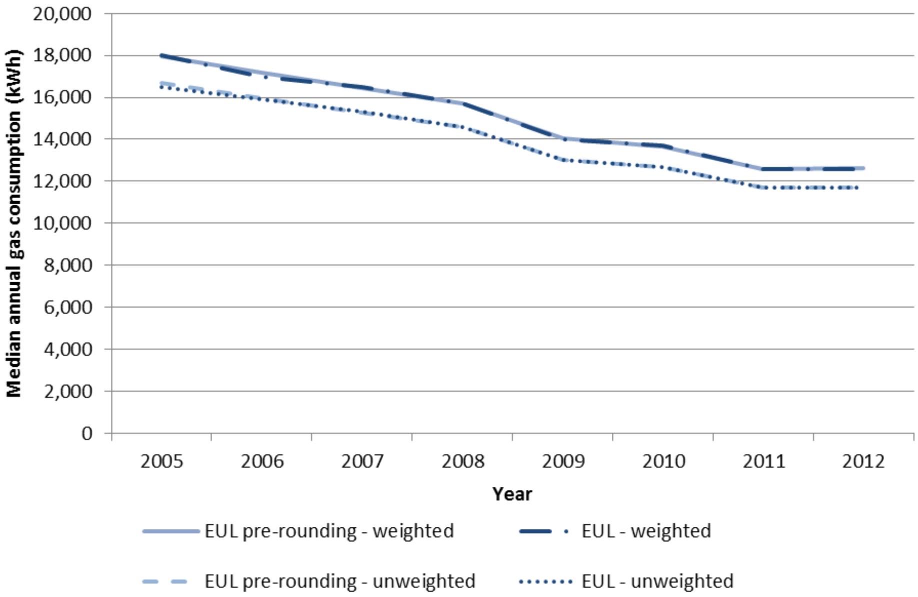
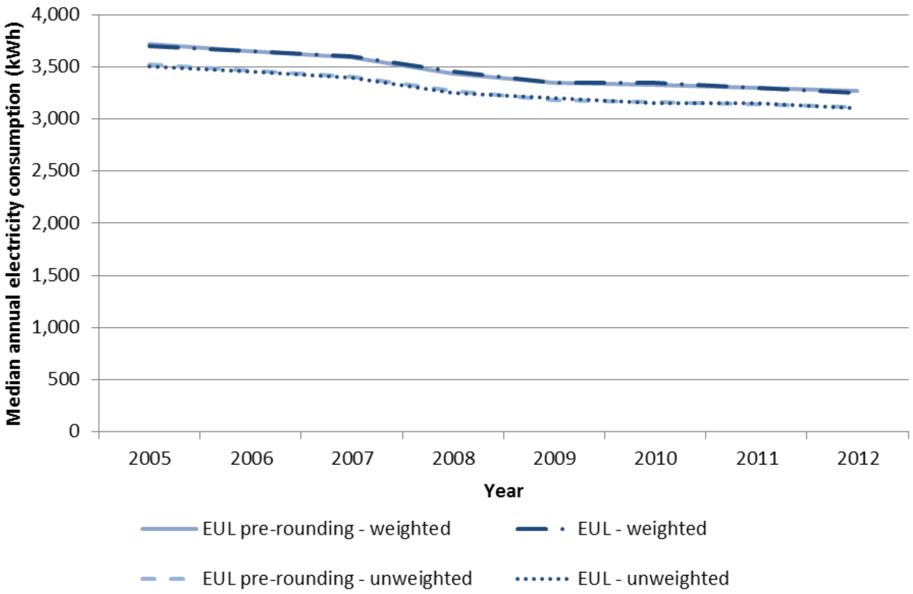


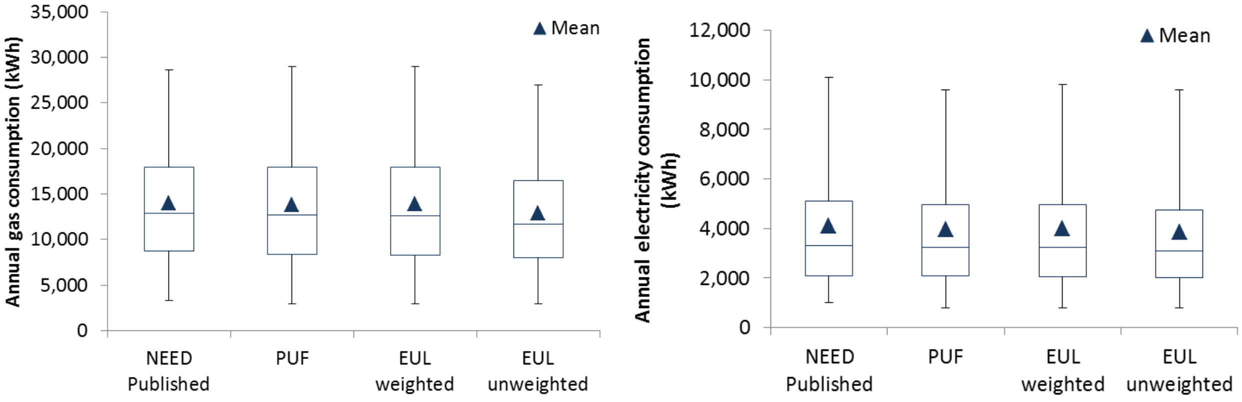
Figure 4.11: Median annual electricity consumption, EUL rounded and unrounded, 2005 to 2012



The charts show that again the rounding has very little impact on the estimates of typical consumption. The weighting has a more significant impact. Without the weighting, typical gas consumption would be approximately seven per cent lower in all years, and up to ten per cent below the NEED published data. The impact for electricity is similar, with the unweighted estimates five per cent below the weighted estimates in all years.

The mean and median consumption values disguise variation within the dataset. Figure 4.12 shows the distribution of annual consumption in each of the datasets, it shows mean and median consumption, as well as the upper and lower quartiles and the 5th and 95th percentiles.

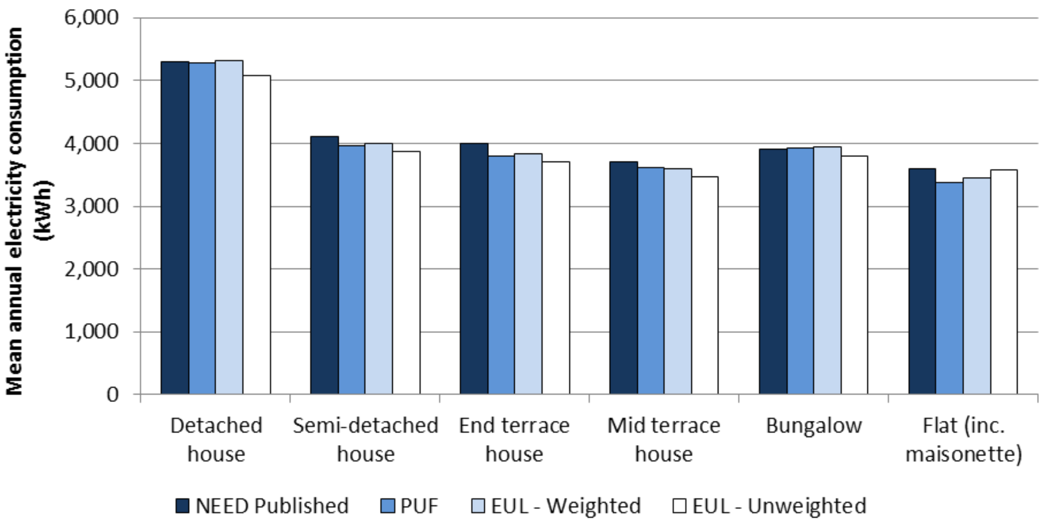
Figure 4.12: Comparison of distribution of annual consumption, 2012



The box plots show that for gas and electricity the distribution of the data is similar for all datasets, with the EUL unweighted dataset showing the greatest difference (the same or lower than the other estimates for all statistics shown). There is generally slightly more variation in distributions for electricity than for gas.

The variation in the data can also be understood further by considering typical consumption for different property attributes. Figure 4.13 shows the mean electricity consumption by property type for each of the datasets.

Figure 4.13: Mean electricity consumption by property type, 2012



This shows that there is more variation seen between the results from different datasets when considering breakdowns. However, estimates are still very close in each case. The biggest variation between datasets is for flats, suggesting that the weighting is not entirely addressing the differences in these properties.

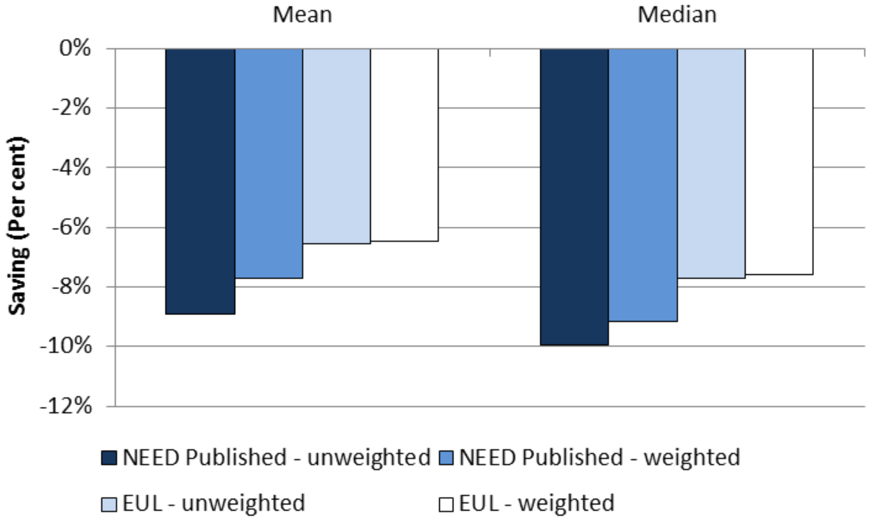
Impact of energy efficiency measures

This section shows the difference in estimates of the impact of installing energy efficiency measures produced using the anonymised dataset compared with NEED official statistics estimates. The results are produced based on the same methodology as that used for the headline NEED results[25], but using a different sample of properties; the NEED EUL dataset[26]. In almost all cases savings estimates based on the EUL dataset are lower than the headline savings from NEED. Further work is required to understand the reasons for this, but it is likely to be influenced by the bias in the EPC data sample. For example the fact it includes more properties with a higher turnover of occupants (properties that have been sold recently and properties which are rented). There is also more uncertainty surrounding the quality of the property attributes data in the EPC dataset; further work is on-going to understand this. Cavity wall insulation and loft insulation are considered in more detail below.

Cavity wall insulation

Figure 4.14 shows weighted and unweighted results for cavity wall insulation installed in 2011 using the two difference samples.

Figure 4.14: Percentage saving following installation of cavity wall insulation in 2011, comparison of NEED headline estimates and estimates from EUL dataset



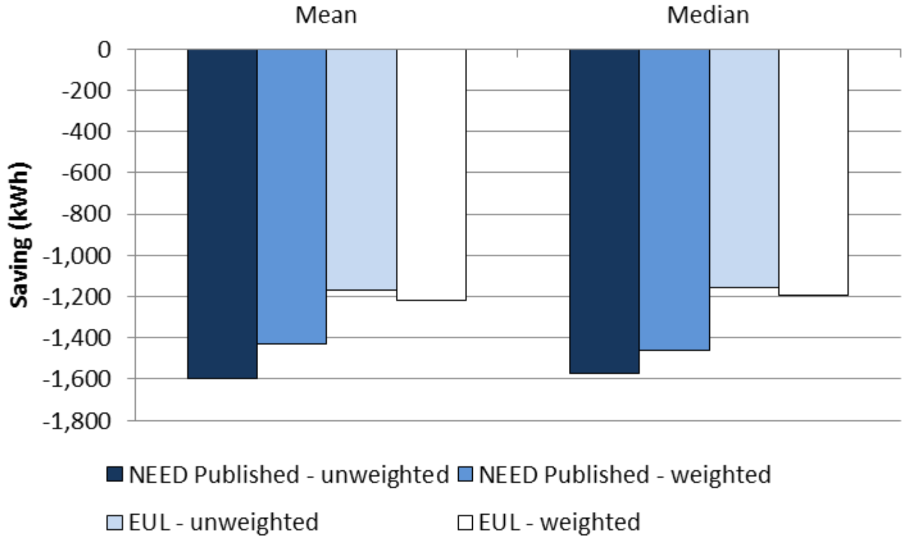
Savings estimates for the EUL dataset are lower than the headline savings from NEED. For the unweighted results, there is a 2.3 percentage point difference for mean and median between the NEED official estimates and estimates produced using the same methodology using the EUL dataset. For the weighted data the differences are 1.2 percentage points (mean) and 1.6 percentage points (median). Figure 4.14 also shows that the weighting has a much smaller impact on the results from the EUL dataset than on the NEED headline results.

The savings in kWh show similar differences between the two datasets. However the weighted saving is greater than the unweighted saving for the EUL dataset, this is not the case for the NEED headline savings, see figure 4.15.

[25] https://www.gov.uk/government/publications/national-energy-efficiency-data-framework-need-report-summary-of-analysis-2014 See: Headline tables: impact of measures 2011, cavity wall insulation 2011 and loft insulation 2011.

[26] Properties included in the dataset are the same, but more detailed data (rather than rounded data or additional banding) has been used for this analysis.

Figure 4.15: Saving following installation of cavity wall insulation in 2011, comparison of NEED headline estimates and estimates from EUL dataset

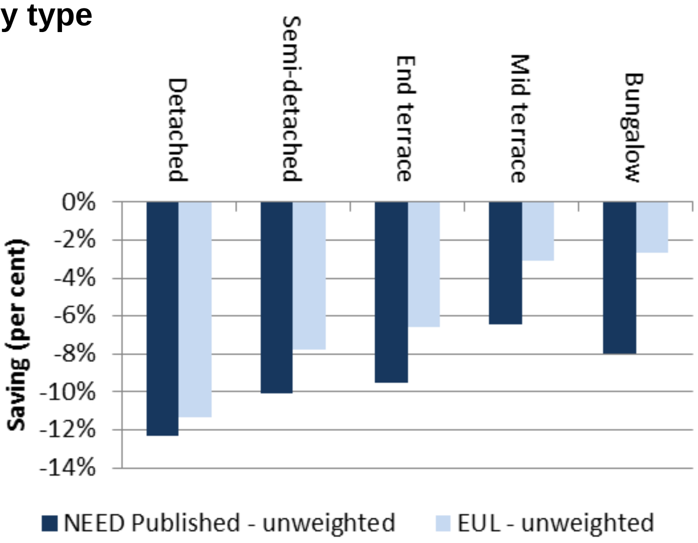


The difference in the results following weighting is partly due to the different approach taken to weighting. The weighting attempts to produce estimates to reflect the typical saving for all properties, not just those which have had a measure installed. In DECC's NEED outputs, the weighting factors are calculated specifically for the group of properties which had a measure installed. The EUL results are based on the weights published in the dataset and therefore are based on weights which are produced in order to ensure the dataset as a whole reflects the population of dwellings in England and Wales, not specifically for the group of properties which have had cavity wall insulation.

The rest of this section shows comparisons of unweighted results, in line with the headline results by property attributes published in the NEED report.

Figure 4.16 shows the typical savings following the installation of cavity wall insulation in 2011 by property type for NEED compared with the results produced using the same methodology on the EUL dataset.

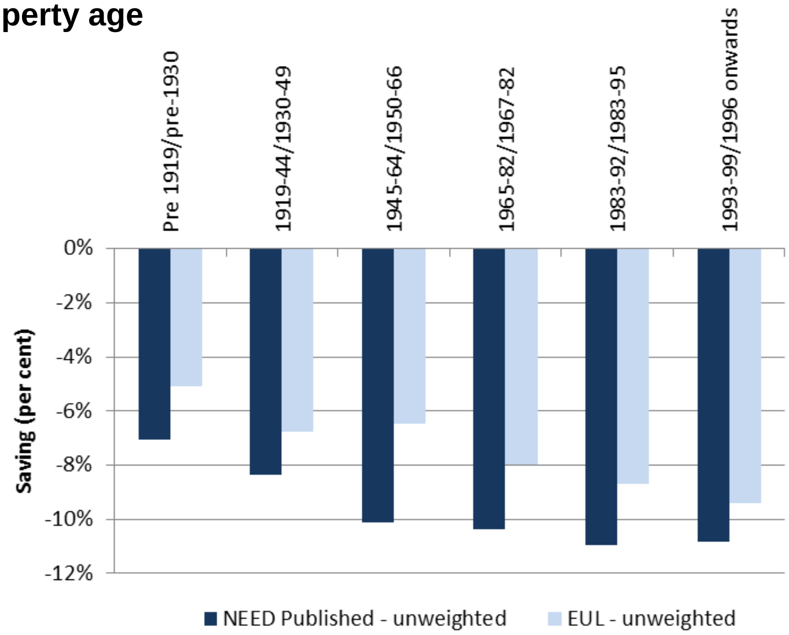
Figure 4.16: Typical percentage saving following installation of cavity wall insulation in 2011, by property type



In all cases the saving estimates based on data in the the EUL dataset are lower than the NEED headline estimates. There is also variation in the difference for different types of properties. Savings estimates for detached properties are the most similar (difference of 0.9 percentage points) and savings for bungalows show the greatest difference (5.3 percentage points).

A comparison by property age can be seen in Figure 4.17[27].

Figure 4.17: Typical percentage saving following installation of cavity wall insulation in 2011, by property age



The differences in results vary from EPC savings being 64 per cent of NEED headline savings (1945-64/1950-66) to 87 per cent (1993-99/1996 onwards). Part of this variation will be due to the differences in periods covered on the two datasets; however it will also result from the properties included in the two samples. In the EUL file 27 per cent of properties with cavity wall insulation in 2011 (and no other measure recorded as being installed) were built in 1983 or later, the equivalent in DECC's NEED sample is 20 per cent. Some variation will result from the additional measures excluded in DECC's NEED analysis. Properties with draft proofing or new glazing recorded as being installed are not included in the NEED analysis. There is no information on these properties in the EUL.

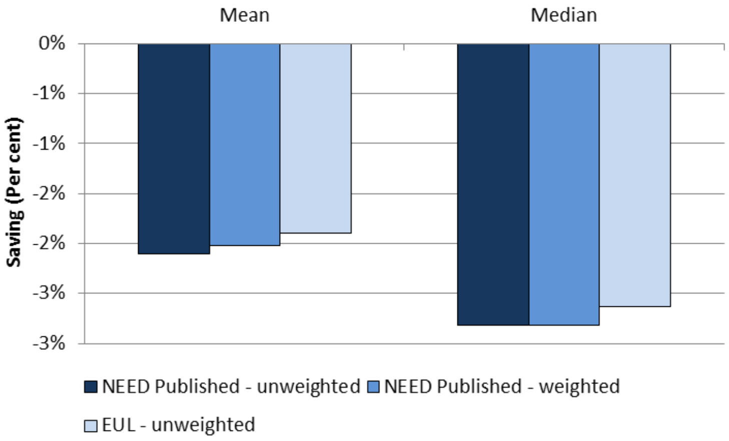
In addition, data quality and uncertainty outlined in the NEED report also applies to the estimates outlined here. There is significant variation in savings experienced by households much of which is down to behaviour of individuals and how they use their homes.

Loft insulation

Similar analysis was undertaken for loft insulation. The EUL estimates of savings following installation of loft insulation in 2011 are close to those published in NEED. The mean and median savings estimated using the EUL dataset are both 0.2 percentage points below the NEED published estimates, see figure 4.18.

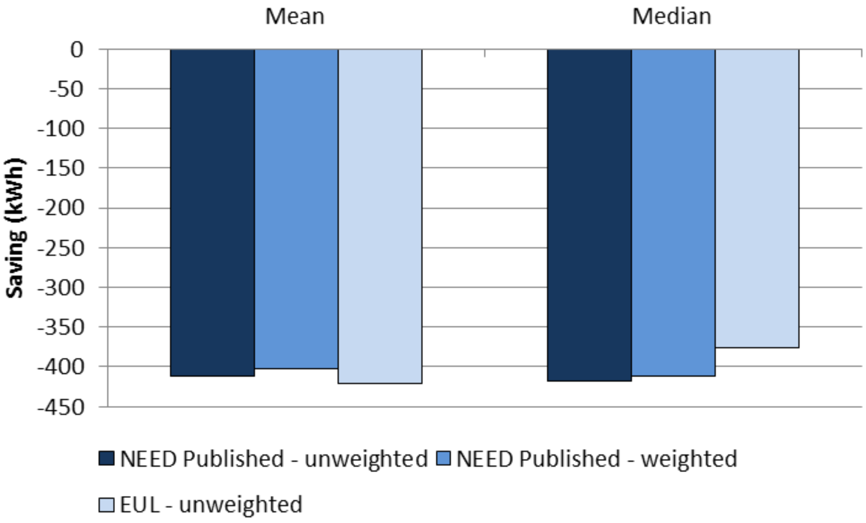
[27] First dates are NEED VOA years; second dates are EPC year groups. EPC groupings have been used to match VOA as closely as possible.

Figure 4.18: Percentage saving following installation of loft insulation in 2011, comparison of NEED headline estimates and estimates from EUL dataset



The actual savings (in kWh) are also similar, see figure 4.19.

Figure 4.19: Saving following installation of loft insulation in 2011, comparison of NEED headline estimates and estimates from EUL dataset



The results consistently give lower estimates from the EUL file than the NEED published estimates. These differences are more evident when considering the impact of installing energy efficiency measures than when considering typical consumption. This suggests that despite the sample structure and weighting some bias in the data remains.

Boiler data have been included in the dataset, but are not presented here. The data are based on data from Corgi and Gas safe. The boiler data were provided late in the process of putting together the anonymised dataset and DECC are looking at these data further to understand the quality of the data. Analysis using the boiler data should be treated with caution.

Annex B of DECC's latest NEED report provides further analysis of the EUL dataset, including estimates of consumption and savings following installation of measures by Energy Efficiency Band and Environmental Impact Band[28].

[28] Annex B: Energy Performance Certificate Data https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/323946/Annex_B_-_Energy_Performance_Certificate_data.pdf.

5. Future plans

These datasets represent the first release of record data from NEED. DECC is extremely grateful to all the parties who have worked with DECC to allow this project to progress, including a range of data providers and potential users of the data.

DECC will now work with organisations and individuals to review the publication of the datasets, including looking at:

- how the datasets have been used;
- the range of data included in the dataset, such as which variables have been most valuable, increasing coverage to include Scotland and more information on household characteristics; and
- the anonymisation of the data released and approach to release - including whether more data can be made available as Open Data and whether a more detailed dataset can be made available through a secure environment.

DECC welcomes input into this review from all users and potential users, including via a planned event for NEED users in September 2014. Alongside this, DECC will continue to work with academic and Open Data communities to understand the value of and priorities for future datasets. A significant part of this will be to understand how the data have been used and which variables are most valuable to each group of users. It is anticipated that the PUF will give users an opportunity to understand the data and its potential uses in order to consider priorities without the need to access the EUL dataset.

DECC plans to publish an updated dataset in 2015. This dataset or datasets will be informed by the review and include gas and electricity consumption data for 2013. DECC will continue to liaise with anonymisation experts and the ICO to ensure future publications include as much useful data as possible while remaining consistent with ICO guidance and Government best practice.

DECC will also continue to publish outputs from its own analysis from NEED, including analysis of consumption for different property types and household characteristics, and estimates of the typical reduction in annual gas consumption following the installation of energy efficiency measures[29].

Comments, queries and feedback on the datasets are welcomed and can be provided by email: EnergyEfficiency.Stats@decc.gsi.gov.uk.

[29] Available at: https://www.gov.uk/government/collections/national-energy-efficiency-data-need-framework.

Responsible statistician: Mary Gregory

Department of Energy & Climate Change
3 Whitehall Place
London SW1A 2AW

www.gov.uk/decc

URN 14D/267