NEED Annex D: Determinants of household gas use

27 June 2019

This report describes how annual household gas use per m^2 of floor area in England and Wales varies according to property characteristics, household characteristics and local area characteristics. This was done using linear regression models, where separate regression models were built for houses and for flats.

Key findings

A typical house in England and Wales has an annual gas use of 130 kWh per m^2, while a typical flat uses 165 kWh per m^2. These figures should not be compared, as floor area is not defined in the same way for flats as it is for houses (see page 3). The figures below present the estimated effects of various variables on annual household gas use per m^2 of floor area after controlling for other factors[1]:

- Newer properties tend to use less gas per m^2 than older properties. The newest houses (those built after 2012) use 58 kWh per m^2 less gas than the oldest properties (those built before 1919). There is a 63 kWh per m^2 reduction between the oldest and the newest flats.
- Mid-terrace houses use an average of 13-26 kWh of gas per m^2 less than all other types of house[2]. Single floor houses use 21 kWh per m^2 more than 2-floor houses. Considering types of flats, on average purpose-built flats use 23 kWh per m^2 less than converted flats.
- Properties with a larger number of rooms were found to use less gas per m^2. Among houses, those which had 5 bedrooms or more, use 33 kWh per m^2 less gas than 1-bedroom houses. Flats with 3 bedrooms or more use 24 kWh per m^2 less than 1-bedroom flats.
- Regarding energy efficiency measures taken under government schemes, the regression modelling indicated that solid wall insulation was the most effective measure, implying a saving of 19 kWh of gas per m^2 for houses[3], and a saving of 25 kWh per m^2 for flats.
- Regarding the area the property is in, the LSOA the property is in appears to have little effect on the gas use per m^2.

What you need to know about this report:

The modelling of annual household gas use in this report is based on data for 11.3 million houses and 1.3 million flats in England and Wales. The information on how much gas was used by these properties was based on temperature adjusted meter readings covering a year-long period from July 2016 to July 2017.

---

[1] The full list of factors included can be found in Appendix B (for houses) and Appendix C (for flats).

[2] The house types being referred to here are ‘mid-terrace’, ‘end-terrace’, ‘semi-detached’ and ‘detached’. Bungalows are included within each of these categories rather than being treated as a separate category.

[3] The Impact of measures analysis included in the tables in the same publication as this Annex provides a more robust assessment of the saving from energy efficiency measures installed.

Contents

Introduction

The National Energy Efficiency Data-framework (NEED) report[4] presents various breakdowns for the median annual household gas consumption in England and Wales including by property type, floor area and the number of adults. Note that the annual gas use figures for each property are temperature adjusted, so in principle the effect of variation in weather conditions has been removed from the gas use figures. In this Annex to the 2019 NEED report, regression modelling is used to look at how various factors affect household gas use, attempting to compare like for like by controlling for other variables.

The modelling approach chapter in this report gives an overview of the regression techniques used.

Separate models for houses and flats

In this analysis separate models were built for houses and for flats. The main reason for this was that, while for houses the floor area includes the entire footprint for the house, for flats it only includes the Effective Floor Area. The Effective Floor Area is defined as the usable area of the rooms within the property excluding bathrooms, WCs, showers and lobby areas.

In addition to this there are two explanatory variables included in the modelling which are only applicable to either houses or flats:

- the floor the property is on (only applicable to flats)
- the number of floors the property has (only applicable to houses)

Controlling for the size of properties

Annexes to previous editions of the NEED report have used regression models to understand the effects of various factors on household gas use. The regression analysis carried out by Katalysis[5] found floor area to be the most important factor in explaining variation in household gas use. This remains the case. Fitting a linear regression models for gas use in terms of floor area for both houses and flats:

- For houses, floor area explains 26.2% of the variability in gas use. Assuming that gas use is directly proportional to floor area and fitting a no-intercept model, only 22.3% of the variability in gas use is explained by the floor area. The results of these models can be found in Appendix A1.
- For flats, floor area (as recorded by the VOA) explains 8.4% of the variability in gas use. Assuming that gas use is directly proportional to floor area and fitting a no-intercept model,

[4] The 2019 NEED report: www.gov.uk/government/statistics/national-energy-efficiency-data-framework-need-report-summary-of-analysis-2019

[5] Summary of Katalysis Regression (2012): www.gov.uk/government/uploads/system/uploads/attachment_data/file/65975/6870-need-report-annex-f.pdf

only 1.4% of the variability in gas use is explained by the floor area. These models can be found in Appendix A2.

The Katalysis report modelled total household gas use and included floor area as an explanatory variable in the model. Other annexes to past NEED reports[6] have taken the same approach.

Households use gas primarily for heating, and so the larger the property (as measured by floor area), the greater the amount of energy required to heat the available space. In this report the size of the property is accounted for explicitly by modelling gas use per m^2 of floor area. This approach is not perfect as the gas used for cooking and heating water is not expected to be to be directly linked to floor area.

The advantage of this approach is that it allows the inclusion of variables such as the number of bedrooms in the model. The Katalysis report for example, does not include this variable owing to it being correlated with floor area. However, this problem does not arise for the approach taken in this report as floor area is not included in the model as an explanatory variable.

When a regression model is built for gas use including both floor area and the number of bedrooms, this gives the result that gas use increases with the number of bedrooms (even after controlling for floor area). This would suggest that houses with more bedrooms are less efficient which is a misleading result. If instead the average gas use per m^2 is considered against the number of bedrooms in the property, there is a clear trend of gas use per m^2 falling as the number of rooms increases. This pattern is preserved when this is broken down by other variables such as the number of floors. It also persists when a large number of variables are controlled for by building a regression model for gas use per m^2. This is the approach taken in this report.

Properties excluded from the analysis

Households were excluded from the analysis if there were any of the following issues with the data for the household:

- The gas use recorded for the household for 2016 was unchanged from that recorded for the previous year, suggesting that the gas use figure was imputed as a result of the true use figure not being available.
- The recorded gas use was not in the range 2,500 kWh – 50,000 kWh, this is consistent with the “Impact of Measures” analysis included in the NEED publication.
- Where data was unavailable for floor space, number of rooms, the age of the property, the number of floors (for houses) or the floor level (for flats).
- The floor space per floor was less than $10m^2$

[6] Summary of Katalysis Regression (2012): www.gov.uk/government/uploads/system/uploads/attachment_data/file/65975/6870-need-report-annex-f.pdf
Summary of NERA Work (2012): www.gov.uk/government/uploads/system/uploads/attachment_data/file/65974/6869-need-report-annex-e.pdf
Predicting Gas Consumption report (2016): www.gov.uk/government/uploads/system/uploads/attachment_data/file/532538/Annex_C_Predicting_gas_consumption.pdf

Results: Houses

This section sets out the determinants of gas use per m^2 of floor area for houses. A regression model was built for gas use per m^2 of floor area for houses. This model explains 15.6% of the variability in gas use per m^2. The age of the property and the number of adults living within it (per 100 m^2) were two of the main contributors to the proportion of the variability explained. Older houses and those with the most adults per (100 m^2) tend to use the most gas per m^2. The full results of this regression model can be found in Appendix B and this chapter presents the main findings.

The distribution of gas use per m^2 of floor area for houses is shown in Figure 1. A typical house uses 130 kWh of gas per m^2. The vast majority (95%) of houses use between 47 and 277 kWh per m^2 (see Table 1).

Figure 1: The distribution of annual gas use per m^2 of floor area, for houses

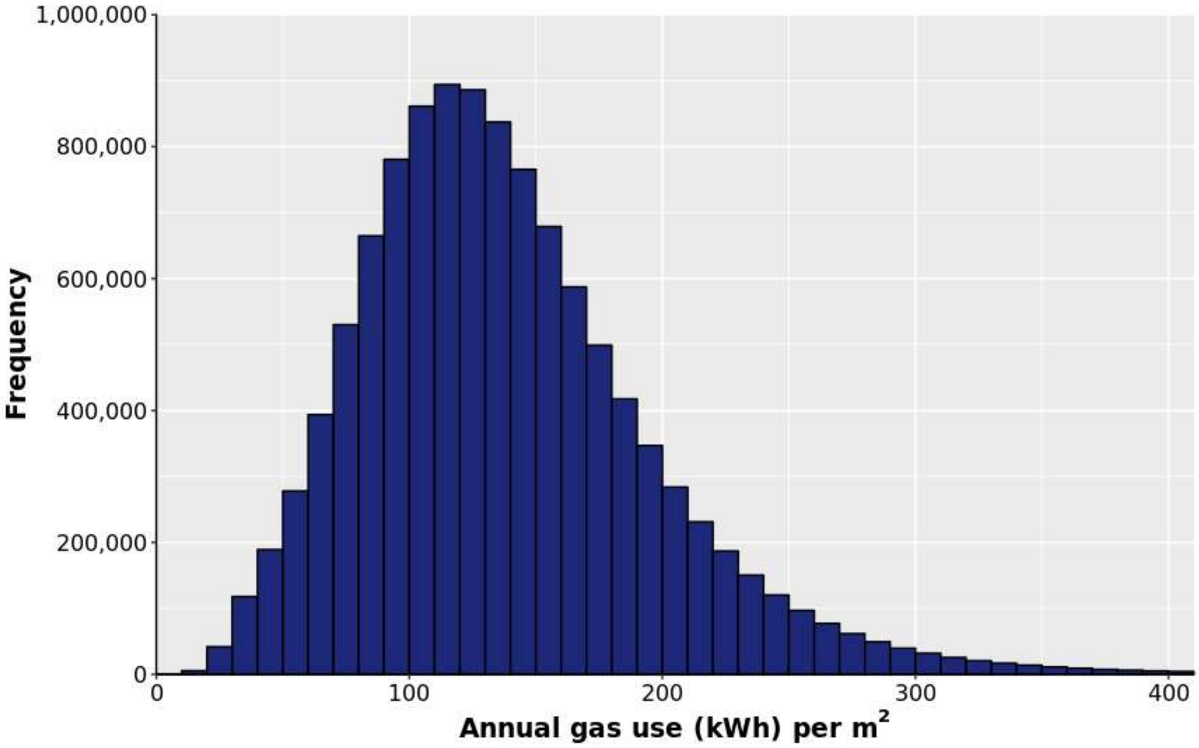


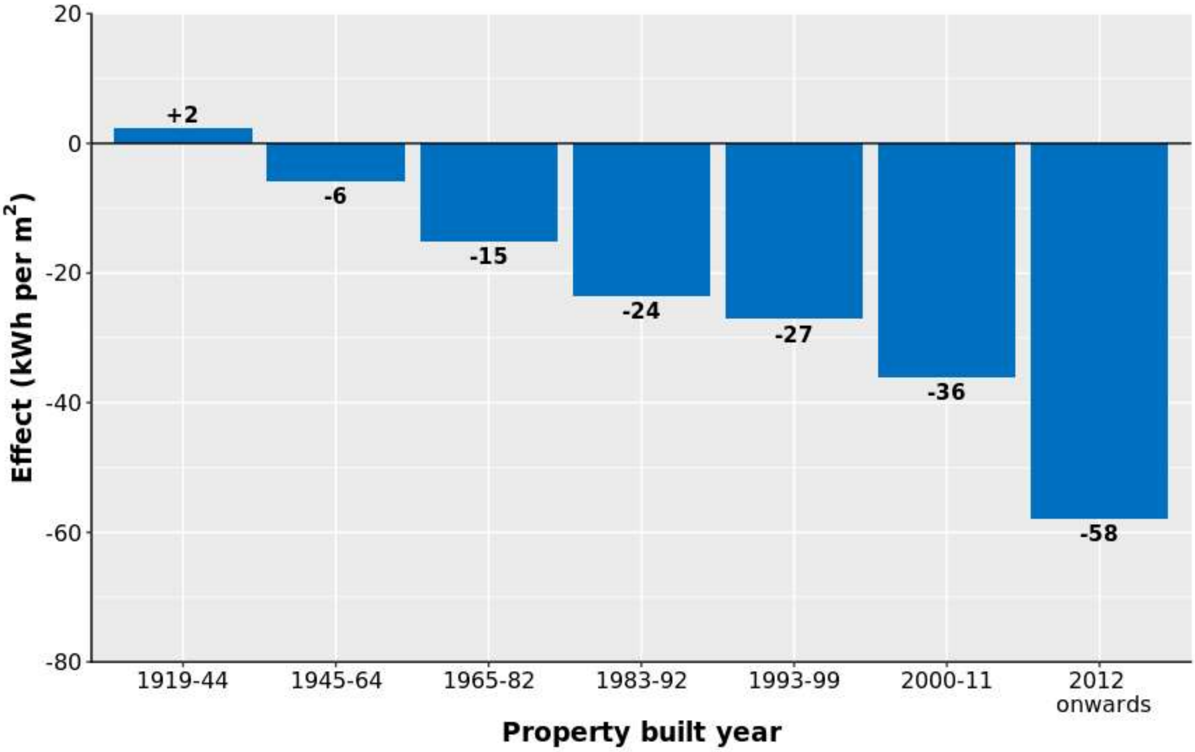
Table 1: Summary statistics for houses: annual gas use, floor area, and gas use per m^2 of floor area

	Quantiles						
	2.50%	Lower quartile	Median	Upper quartile	9.75%	Mean	Standard deviation
Gas use per m^2 (kWh / m^2)	47	98	130	169	277	138	59
Gas use (kWh)	4,237	9,384	13,049	17,776	33,028	14,379	7,212
Floor area (m^2)	59	85	98	118	203	106	38

Property characteristics

Out of the possible determinants of gas use per m^2 of floor area examined, the age of the house, appeared to be the most important factor. After controlling for other variables[7], newer houses tend to use less gas per m^2 than older houses (see Figure 2). On a like for like basis the newest houses (those built 2012 onwards) use 58 kWh per m^2 less than the oldest houses (those built prior to 1919).

Figure 2: The estimated effect of the age of the house on annual gas use per m^2 of floor area, relative to houses built before 1919



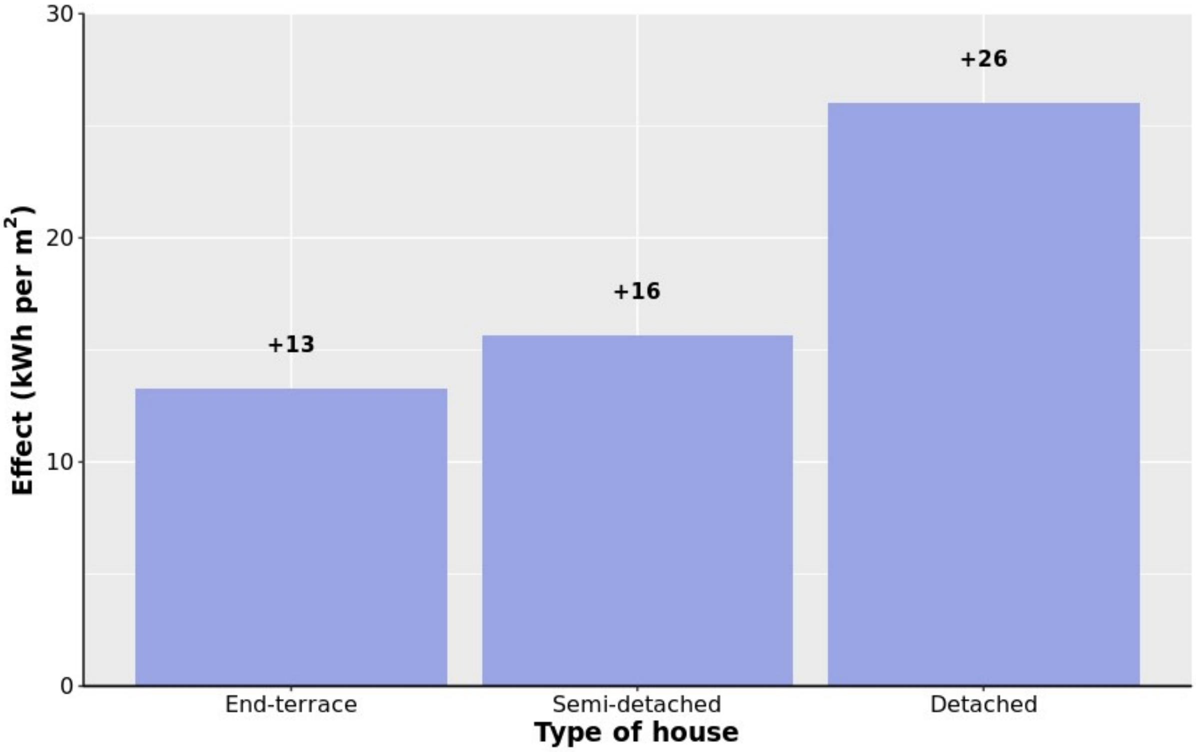
Regarding property type, mid-terrace houses stand out as having the lowest levels of gas use per m^2 (see Figure 5). This is to be expected owing to the extra insulation provided by the neighbouring properties on both sides. In contrast houses which are completely detached from neighbouring properties tend to use the most gas, with 26 kWh per m^2 more used on average than for mid-terrace houses.

The vast majority of houses have two floors. Houses with only one floor use an estimated 21 kWh of gas per m^2 more than those with two floors. Note that this is after controlling for the other explanatory variables[9] including the number of bedrooms and the number of bathrooms.

Houses with more bedrooms tend to use less gas per m^2. On average houses with 5 bedrooms or more use 33 kWh per m^2 less than 1-bedroom houses. Note that this is after controlling for the other factors in the regression model[9] including the number of floors and the number of adults per $100m^2$ of floor area.

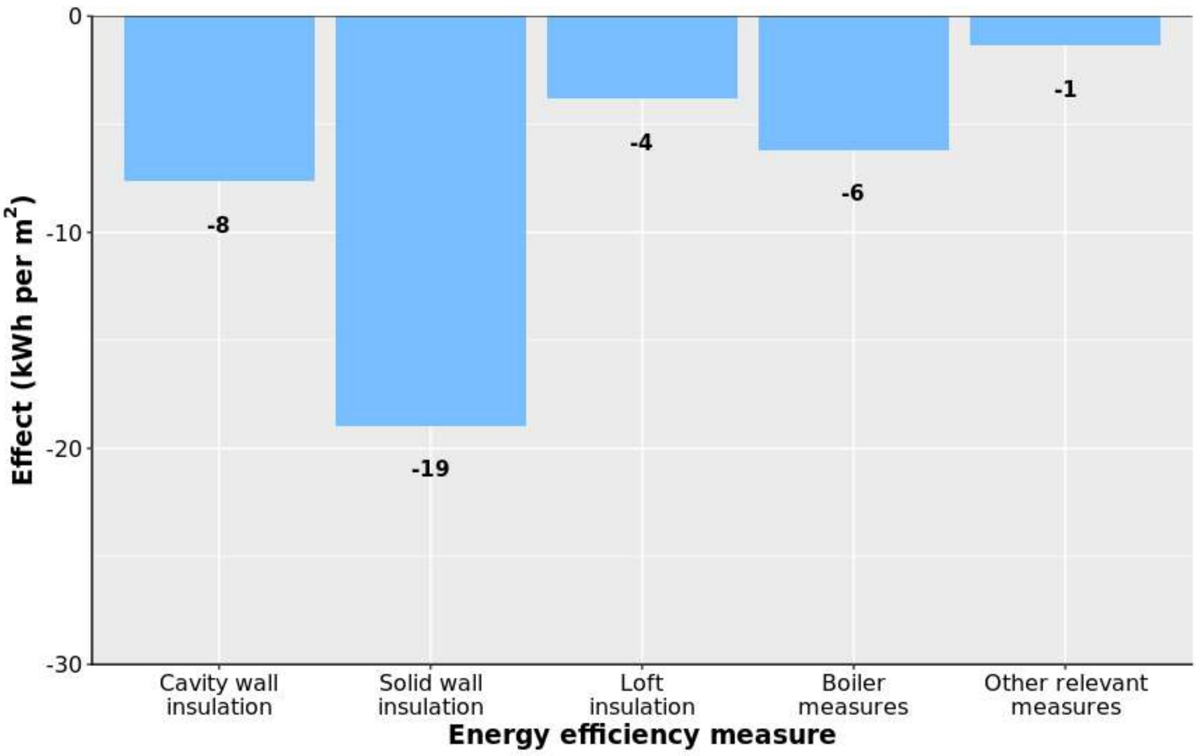
[7] A full list of the variables being controlled for using the regression model for can be found in Appendix B.

Figure 3: The estimated effect of the type of house on annual gas use per m^2 of floor area, relative to mid-terrace houses



Regarding energy efficiency measures taken under government schemes (e.g. ECO and Green Deal), the regression model corroborates the findings in the main National Energy Efficiency Data-framework (NEED) report, that the most effective household energy efficiency measure being taken under government schemes is solid wall insulation (see Figure 4).

Figure 4: The estimated effects of energy efficiency measures taken under government schemes on annual gas use per m^2 of floor area, in houses



The model implies a saving of 19 kWh per m^2 from solid wall insulation, by comparing houses which have had this measure installed (under a government scheme) to those which have not. The implied saving from each of the other measures installed was less than 10 kWh per m^2.

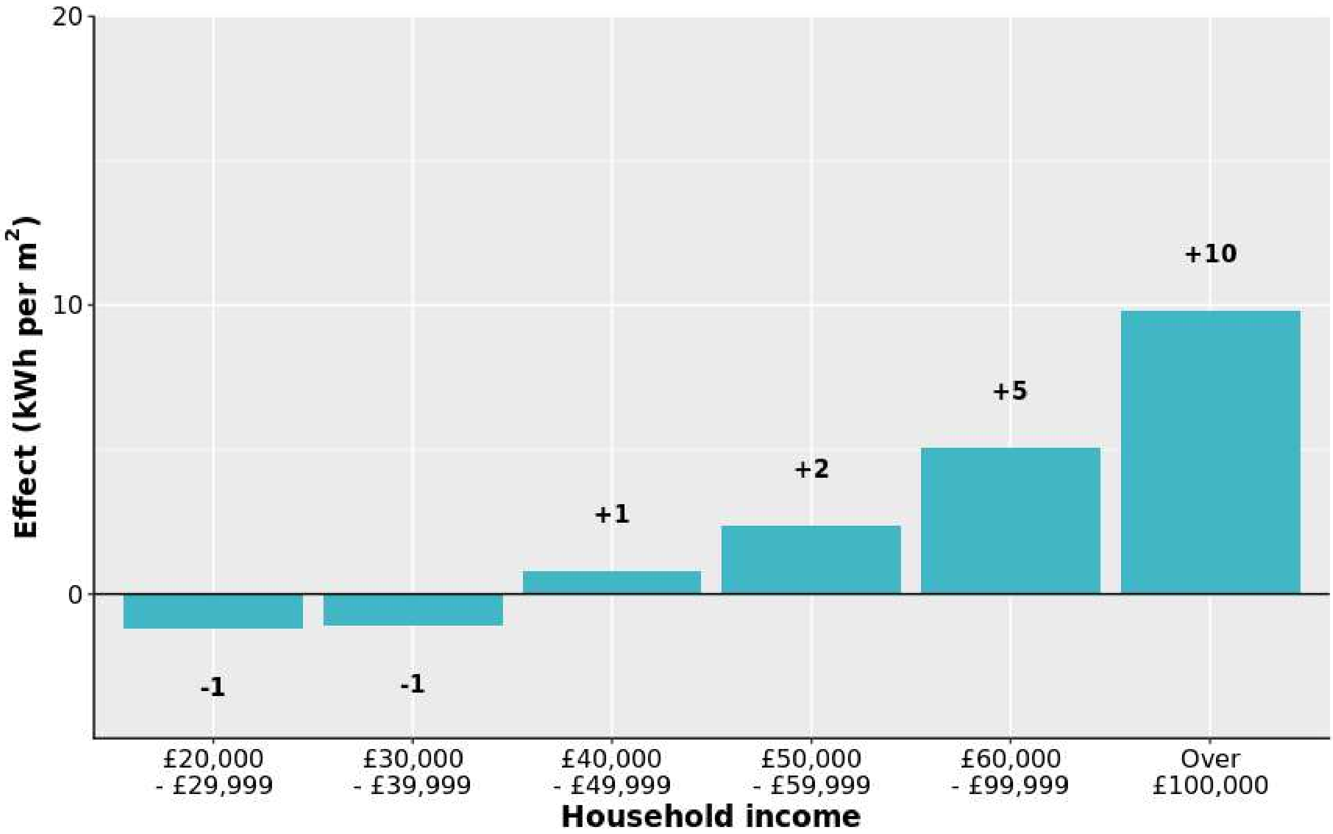
Note that the "Impact of Measures" analysis presented in the main NEED report contains a much more robust assessment of the effectiveness of energy efficiency measures. This is because it uses a control group to compare the change in gas use for properties which have had an energy efficiency measure installed to those which have not, whereas this regression analysis does not examine changes over time.

Household characteristics

The regression model indicates that, after controlling for other factors such as the number of bedrooms, annual gas use increases with the number of adults. On average an additional adult per $100m^2$ of floor area corresponds to an additional 9 kWh of gas used per m^2.

Figure 5 shows that gas use among households on incomes from £20,000 to £50,000 are similar to the households on the lowest incomes (less than £20,000). Gas use per m^2 rises more sharply at the other end of the income scale with those households on more than £100,000 using 10 kWh per m^2 more on average, than the households on the lower incomes (less than £20,000).

Figure 5: The estimated effect of household income on gas use per m^2 of floor area, in houses, relative to households earning less than £20,000



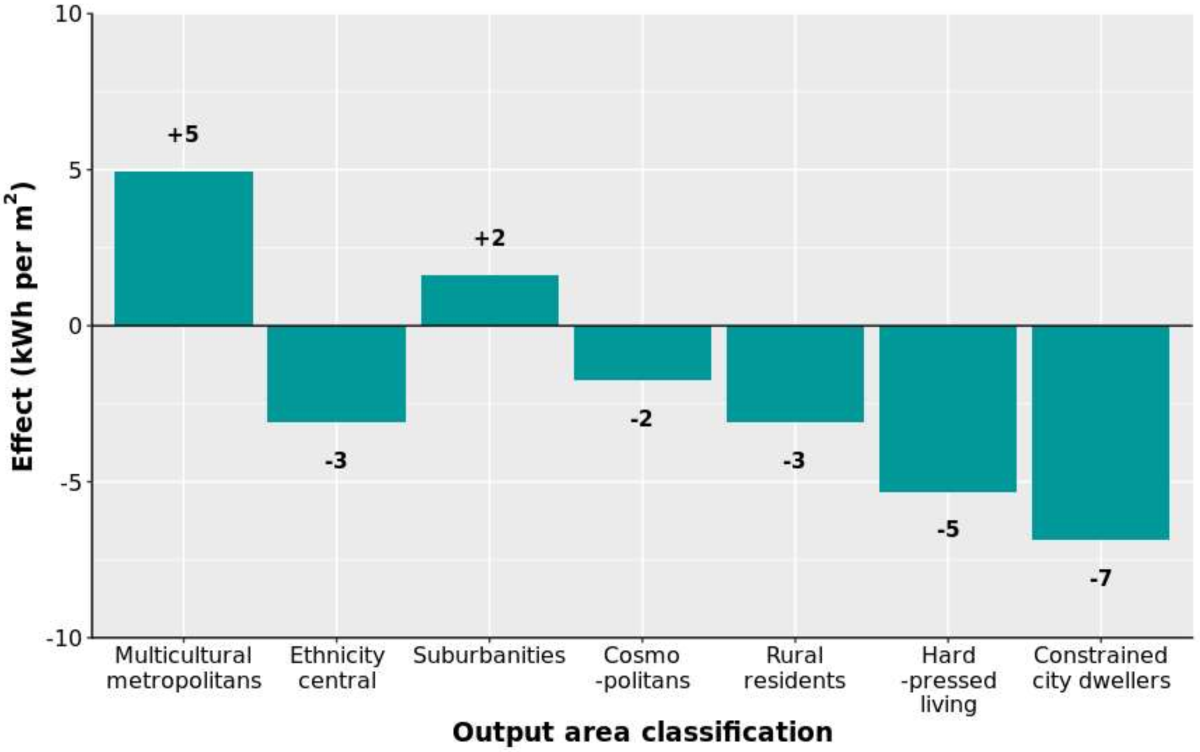
Households in properties owned by the council or housing associations use less gas per m^2 than owner occupiers. Gas use for these houses is 9 kWh per m^2 lower than owner occupied houses and 8 kWh per m^2 lower than for privately rented houses. Note that this is after controlling for the other variables in the regression model[8] including household income. It should also be noted that while the installation of energy efficiency measures was included in the regression model, the Energy Performance Certification (EPC) rating was not. If the EPC rating of properties was included in the model it is possible that this would reduce the apparent impact of housing ownership, as properties owned by council or housing associations are known to be more energy efficient in general.

Local area characteristics

Houses in conurbations (the most urban parts of the country) tend to use the most gas per m^2 of floor area. These houses use 7 kWh per m^2 more on average, than houses in smaller cities and towns.

Based on the 2011 Census Output Area Classifications[9], houses in areas categorised as “Constrained City dweller” or “Hard-pressed living” tend to have the lowest gas use per m^2 (see Figure 6).

Figure 6: The estimated effect of the output area classification on gas use per m^2 of floor area, in houses, relative to output areas classified as Urbanites



[8] A full list of the variables being controlled for using the regression model for can be found in Appendix B.

[9] An overview of the 2011 Census Output Area Classifications can be found at: www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/ns-2011-area-classifications/pen-portraits-and-radial-plots/pen-portraits-oa.pdf

The regions with highest gas use per m^2 for houses tend to be in the North (e.g. 'North West' and 'Yorkshire and The Humber'), while those with the lowest tend to be further south (e.g. South West). Houses in the most southern region of England (the South West) use 26 kWh per m^2 less gas on average than those in the most northern region of England (the North East). Note that the annual gas use figures are temperature adjusted, so in principle the effects of variation in temperatures across the country have been removed from the data.

The variability in gas use per m^2 of floor area

The section presents the results of the random effects and random intercepts models. These models are explained in the *Random and mixed effects modelling* section on pages 21 to 22.

Examining the gas use per m^2 for individual houses grouped by the LSOA they are in, gives an Interclass Correlation Coefficient[10] (ICC) of 0.090 (see Table 2). This indicates a low degree of homogeneity (clustering) in the gas use per m^2 between houses within the same LSOA.

This remains the case after controlling for differences between houses in terms of the characteristics of the property and of the household[11] (an ICC of 0.067).

After the inclusion of property, household and local area characteristics (all the explanatory variables included in the linear regression model for houses which are listed in Appendix B) the percentage of the unexplained variance attributable to differences between LSOAs was only 3.8%.

Table 2: Unexplained variance in annual gas use per m^2 of floor area, in houses, before and after the inclusion of the explanatory variables

	Standard deviation		Variance			ICC (proportion of total variance which is between LSOAs)
	Between LSOAs	Within LSOAs	Between LSOAs	Within LSOAs	Total	
Before inclusion of explanatory variables	17.9	56.9	321.5	3,234.4	3,555.9	0.090
After the addition of property and household characteristics	14.4	53.7	207.7	2,885.8	3,093.5	0.067
After the addition of property, household and local area characteristics	10.7	53.7	115.3	2,885.1	3,000.4	0.038

[10] The Interclass Correlation Coefficient (ICC), provides a measure of the degree to which households within the same LSOA are alike in terms of their gas use per m^2 relative to households in general. It takes values from 0-1, where a value close to one indicates a high degree of similarity in gas use per m^2 with LSOAs.

[11] For the full list of the property and household characteristics being controlled for here see Appendix B.

Results: Flats

This section sets out the determinants of gas use per m^2 of floor area for flats. Readers should be aware that at the time of publication, linking datasets together in NEED (based on address matching) is a less reliable process for flats. This means that while the results of this analysis remain valid, there is a higher level of uncertainty. In addition to this, the recorded floor area for flats is not consistent with the floor area recorded for houses and it may not give as reliable an indication of the amount of space being heated (see page 3).

A regression model was built for gas use per m^2 of floor area for flats. This model explains 11.4% of the variability in gas use per m^2. The age of the property and the number of adults living within it (per 100 m^2) were the two main contributors to the proportion of the variability explained. Older flats and those with the most adults (per 100 m^2) tend to use the most gas per m^2. The full results of this regression model can be found in Appendix C and this chapter presents the main findings.

The distribution of gas use per m^2 for flats is shown in Figure 7. A typical flat uses 165 kWh of gas per m^2 of floor area. The vast majority (95%) of flats use between 59 and 455 kWh of gas per m^2 (see Table 3).

Figure 7: The distribution of annual household gas use per m^2 of floor area, for flats

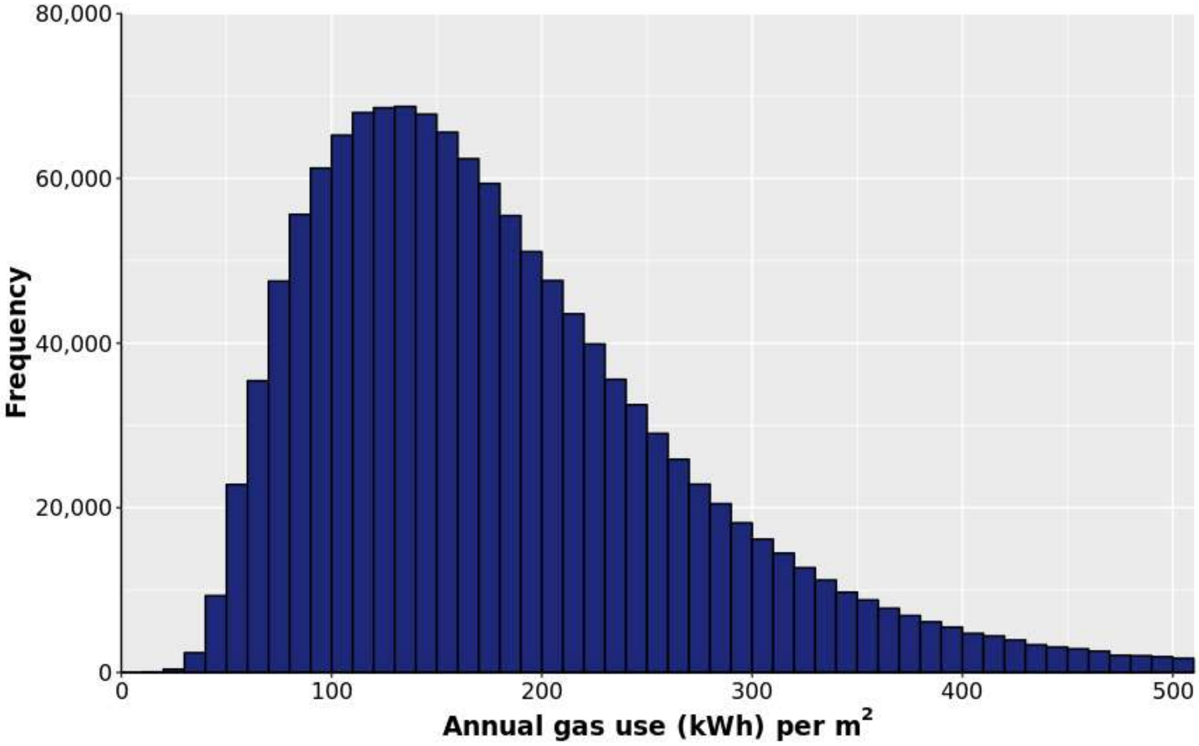


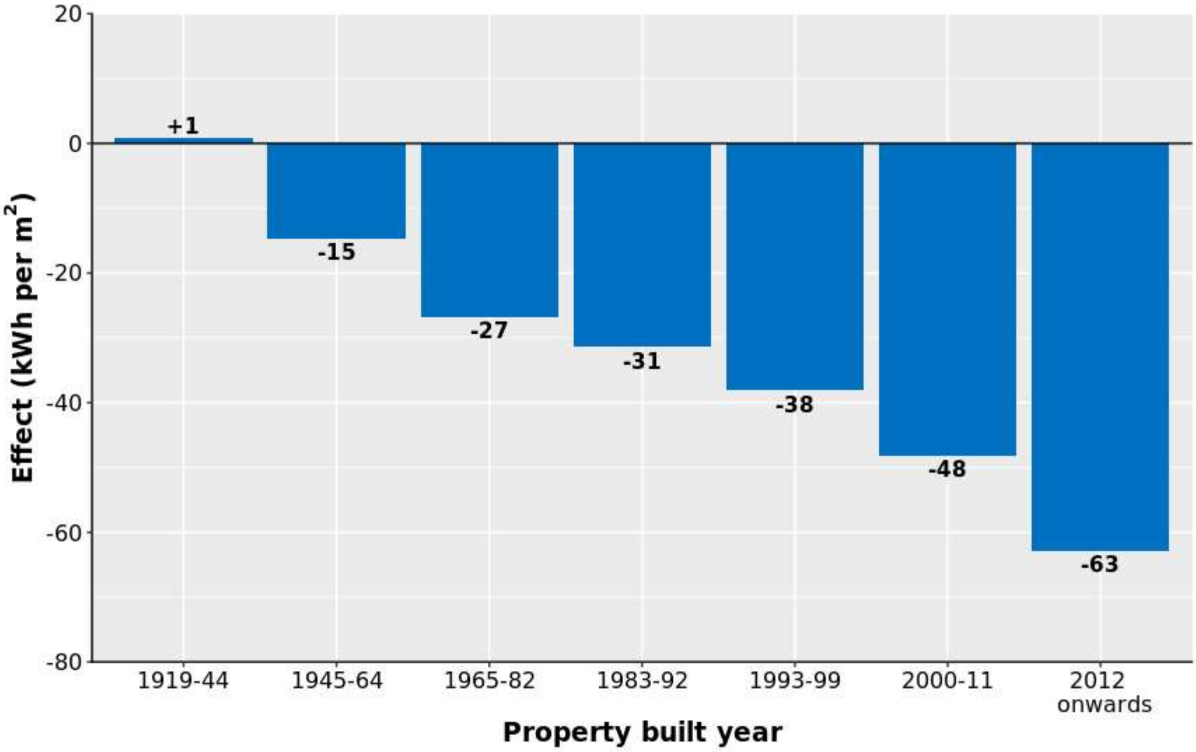
Table 3: Summary statistics for flats: annual gas use, floor area, and gas use per m^2 of floor area

	Quantiles						
	2.50%	Lower quartile	Median	Upper quartile	9.75%	Mean	Standard deviation
Gas use per m^2 (kWh / m^2)	59	115	165	231	455	188	111
Gas use (kWh)	2,849	5,239	7,627	10,901	22,098	8,821	5,234
Floor area (m^2)	28	39	47	55	84	49	16

Property characteristics

As with houses, one of the key factors associated with gas use was the age of the property, with newer properties using less gas per m^2 of floor area than older properties (see Figure 8). On a like for like basis (after controlling for the other variables included in the regression model[12]) the newest flats (those built 2012 onwards) use 63 kWh less per m^2 of floor space than the oldest flats (those built prior to 1919).

Figure 8: The estimated effect of the age of the flat on annual gas use per m^2 of floor area in flats, relative to flats built before 1919



On a like for like converted flats use 23 kWh of gas more per m^2 of floor area more than purpose-built flats.

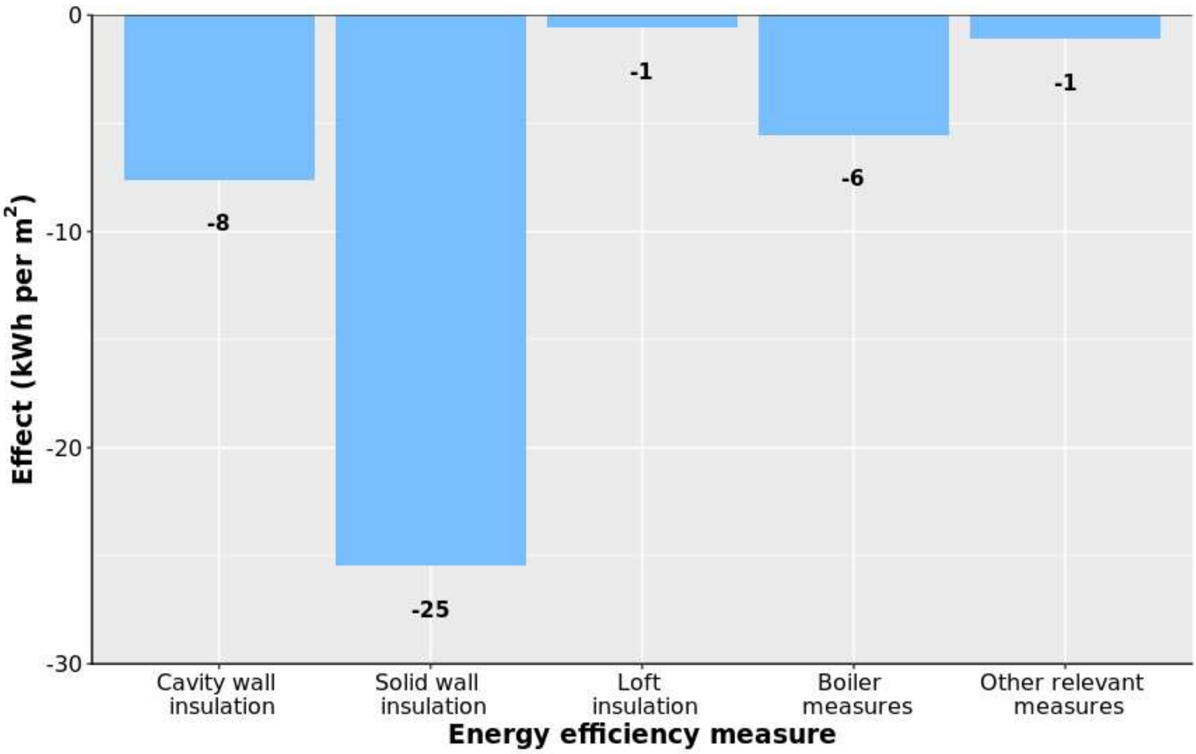
As was found to be the case with houses, flats with more bedrooms tend to use less gas per m^2. Flats with 3 or more bedrooms use 24 kWh per m^2 less than those with only 1 bedroom. Note that this is after controlling for the other factors in the regression model[14] including the number of adults per $100m^2$ of floor area.

Ground floor flats were found to use more gas per m^2 than flats on higher floors. Ground floor flats use 13 kWh per m^2 more than 1st floor flats and 19 kWh per m^2 more than 2nd floor flats.

[12] A full list of the variables being controlled using the regression model for can be found in Appendix C.

Regarding energy efficiency measures taken under government schemes (e.g. ECO and Green Deal), as was the case for houses, the model suggests that solid wall insulation is the most effective energy efficiency measure (see Figure 9). The regression model implies a saving from this measure of 25 kWh gas per m^2 of floor area. The implied saving from each of the other measures installed was less than 10 kWh per m^2.

Figure 9: The estimated effects of energy efficiency measures taken under government schemes on annual gas use per square metre of floor area in flats

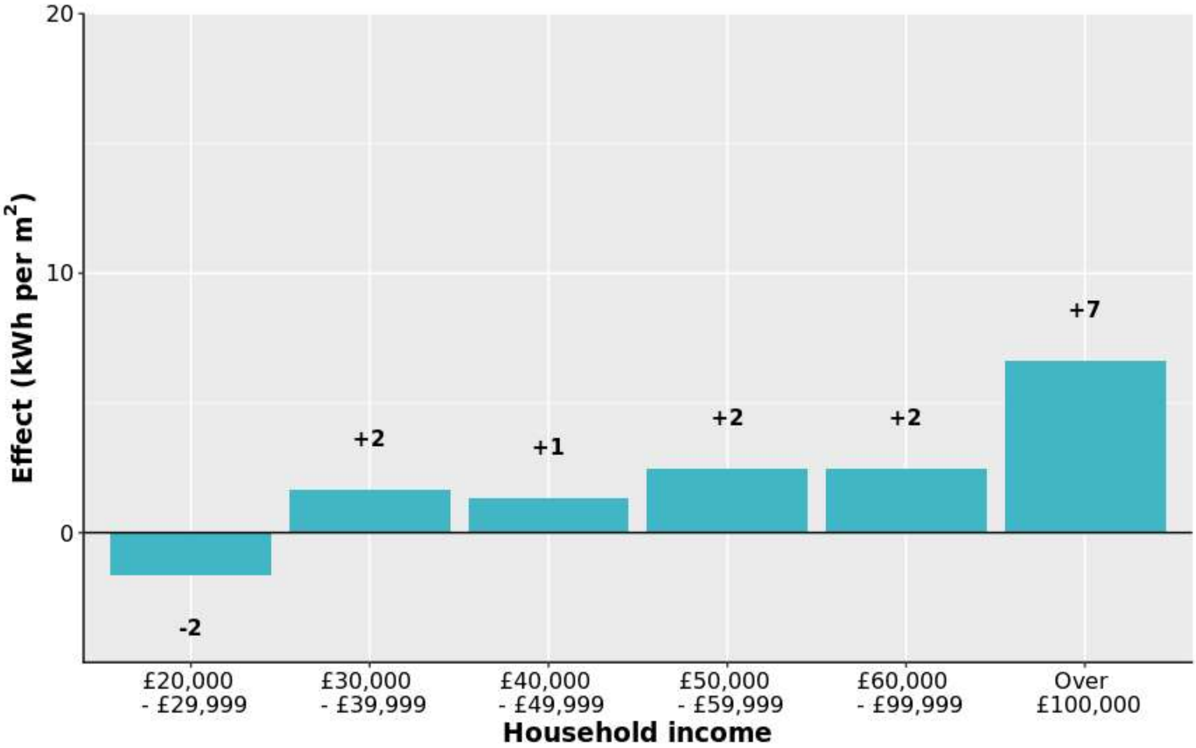


Household characteristics

As was the case for houses, for flats the number of bedrooms, annual gas use increases with the number of adults. On average an additional adult per 100m^2 of floor area corresponds to an additional 9 kWh of gas used per m^2 of floor area.

Among households living in flats, those on the very highest incomes (at least £100,000) stand out as using the most gas, 7 kWh per m^2 more than those on the lowest incomes (less than £20,000), see Figure 10.

Figure 10: The estimated effect of household income on gas use per m^2 of floor area, in flats, relative to households earning less than £20,000



As was the case with houses, council or housing association flats use less gas per m^2 than privately owned flats. Gas use for these flats is 8 kWh per m^2 lower than owner occupied flats and 6 kWh per m^2 lower than for privately rented flats. Note that this is after controlling for the other factors in the regression model[13] including household income.

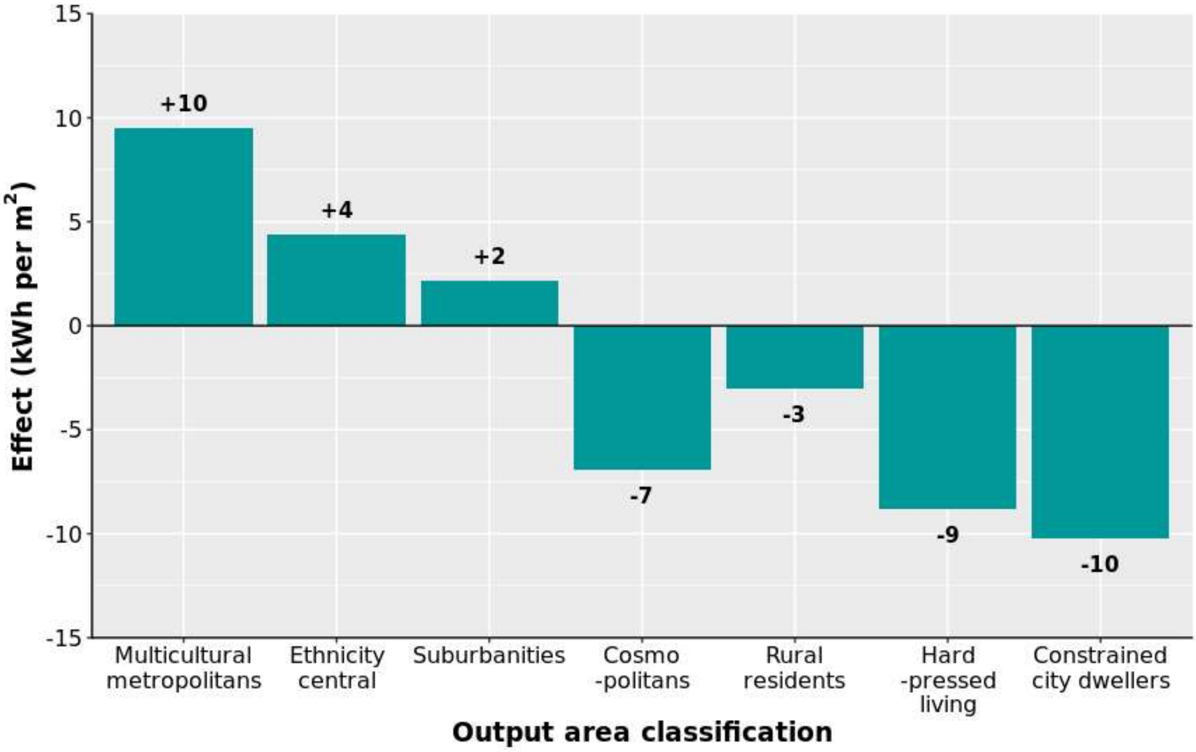
[13] A full list of the variables being controlled using the regression model for can be found in Appendix C.

Local area characteristics

As was the case for houses, flats in conurbations (the most urban parts of the country) tend to use the most gas per m^2 of floor area. These flats are estimated to use 13 kWh per m^2 more than flats in smaller cities and towns.

Based on the 2011 Census Output Area Classifications[14], as was the case for houses, flats in areas categorised as “Constrained City dweller” or “Hard-pressed living” tend to have the lowest gas use per m^2 (see Figure 11).

Figure 11: The estimated effect of the output area classification on gas use per m^2 of floor area in flats, relative to output area classified as *Urbanites*



[14] An overview of the 2011 Census Output Area Classifications can be found at: www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/ns-2011-area-classifications/pen-portraits-and-radial-plots/pen-portraits-oa.pdf

Conclusion

For both flats and houses, the variable which is most closely associated with gas use is floor area. Linear regression models were built for annual gas use per m^2 of floor area. As the recorded floor area for flats is not consistent with houses, separate models were built for flats and houses. For both houses and flats the two main contributors to the variability in gas use per m^2 explained by the models were:

- the age of the property (with lower gas use per m^2 among newer properties)
- the number of adults (with gas use per m^2 increasing linearly with the number of adults per $100m^2$)

The proportion of the variability in gas use per m^2 explained by the models were small (15.4% for houses and 11.6% for flats), suggesting that there may be considerable room for improvement in the fit the models. For houses almost all (96.2%) of the unexplained variability was within LSOAs, with unexplained differences between LSOAs making up only a small proportion (3.8%).

Additional variables could to be considered to try to explain more of the variability is gas use per m^2. The key variable that could be added to the models is the Energy Performance Certification (EPC) rating.

The modelling approach

Separate linear regression models were built for houses and for flats to estimate how various factors affect household gas use per m^2 of floor area. Linear regression models allow more like for like comparisons by controlling for all other explanatory variables added to the model.

For houses, as well as a linear regression model, random and mixed effects models were also used to provide further insights about the variance in gas use per m^2 between and within LSOAs[15]. This analysis was not also carried out for flats. This was because a large number of LSOAs had very few flats remaining after the exclusion of properties from the analysis.

Building the linear regression models

The following variables were considered for inclusion in the model:

Variable	Source
Property characteristics	
Type of property	Valuation Office Agency
Floor area	Valuation Office Agency
Numbers of rooms (bedrooms, bathrooms, other rooms, total rooms)	Valuation Office Agency
Number of floors (for houses)	Valuation Office Agency
Floor number (for flats)	Valuation Office Agency
Property age	Valuation Office Agency
Energy efficiency measures taken as a part of government schemes	BEIS[16] Energy Efficiency Measures Data
Household characteristics	
Number of adults	Experian
Household income	Experian
Tenure (Owner occupied, privately rented, council/ housing association)	Experian
Length of residency (based on longest lasting resident)	Experian
Local area characteristics	
Region (the Government of Regions of England, Wales)	Office for National Statistics (ONS)
Urban/Rural classification of Output Areas[17,17]	Office for National Statistics (ONS)
Output Area classifications[17,18]	Office for National Statistics (ONS)

[15] Output areas were built from adjacent postcode units and were designed to be as socially homogenous as possible. Lower super output areas (LSOAs) are in turn built from adjacent output areas, and in England and Wales these contain 400-1,200 households. See: www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography#super-output-area-soa

[16] Department of Business, Energy and Industrial Strategy

[17] See www.gov.uk/government/statistics/2011-rural-urban-classification

[18] An overview of the 2011 Census Output Area Classifications can be found at: www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/ns-2011-area-classifications/pen-portraits-and-radial-plots/pen-portraits-oa.pdf

The overall approach taken

The models were based on 11.3 million houses and 1.3 million flats. With such a large number of data points, even very small insignificant effects, applicable to only a small proportion of all properties, would result in a statistically significant improvement in the fit of the model. So, a threshold was adopted for deciding which of the explanatory variables should in included in the model. For a variable to be included it had to explain at least 0.1% of the variability in gas use per m^2 in the absence of any other variables. This was based on looking at the R^2 value for the model which only included the variable in question.

The variables were added to the model in 4 stages with a linear regression model fitted at each stage:

1. Property characteristics
2. Energy efficiency measures installed
3. Household characteristics
4. Local area characteristics

When the explanatory variables included in the model are too correlated with each other, this can result in misleading parameter estimates. At the end of each stage the variance inflation factors (VIFs) for each explanatory variable in the model were checked. If the highest VIF was 5 or above, then the variable with the highest VIF was excluded, the model rerun and the VIFs of the remaining variables checked again, until all remaining variables had VIFs below 5.

Variance Inflation Factor (VIF)

If an explanatory variable has a VIF of 2, this means that the variance in the corresponding parameter estimate is twice as large as it would be if the explanatory variables were not at all correlated with each other. An explanatory variable not at all affected by multicollinearity (correlations between the explanatory variables), would have a VIF of 1.

If an explanatory variable has a high VIF this means that a model for the explanatory variable in terms of the other explanatory variables, has a high R^2 value. 5 and 10 are commonly used thresholds for VIFs to reach before there is cause for concern.

The variables excluded

There were 4 possible variables to include in the model referring to the numbers of rooms in the property:

- the number of bedrooms
- the number of bathrooms
- the number of other rooms
- the total number of rooms

After considering which variables were the least correlated with each other, and the most closely associated with gas use per m^2 of floor area, the variables included in the models were “the number of bedrooms” and “the number of bathrooms”.

For flats, “region” was excluded as it had a VIF of 5.55. Once it was excluded the VIFs of several other variables in the model decreased.

No other variables were excluded from either the houses model or the flats model.

One of the criteria set for variables to be included in the model was that they must explain at least 0.1% of the variation in gas use per m^2. Some of the variables referring to energy efficiency measures installed as a part of government schemes did not meet this 0.1% threshold. However, combined, all the “energy efficiency measures installed” variables did meet the threshold. All the energy efficiency measures variables were included for completeness. While readers should be aware that the “Impact of Measures” analysis in the NEED 2019 publication gives more reliable estimates of savings from energy efficiency measures, the analysis in this report provides an alternate way of assessing these savings.

The number of adults

The larger the number of residents living at a property the greater the amount of gas used for heating water. As might be expected, gas use appears to increase linearly with the number of adults. Gas use per m^2 of floor area appears to increase linearly with “adults per m^2 of floor area”. This is case for both houses and flats. Moreover, gas use per m^2 has a higher correlation with “number of adults per m^2” than with “number of adults”. This is why, “adults per m^2 of floor area” was included as an explanatory variable rather than “number of adults”.

Model fit

The model for houses explains 15.6% of the variability in gas use per m^2 of floor area, while the model for flats explains 11.4% of the variability between flats. These figures are referred to as R^2 values.

R^2 (Coefficient of determination) of a linear regression model

This is a measure of the proportion of the variability in the y-variable that can be explained by the model. A high R^2 value means that the model fits the data well. A low R^2 value implies that there may be considerable room for improvement in the fit of the model, possibly through the addition of extra explanatory variables.

The R^2 values in this report cannot be compared with those in previous reports, for two reasons:

- The models in previous reports were for total household gas use and included floor area as an explanatory variable. The models in this report are for gas use per m^2 of floor area and do not include floor area as an explanatory variable (for the reasons explained in the “Controlling for the size of properties” section on page 4). Therefore, the R^2 values in this report do not take into account the additional variability explained as a result of dividing by floor area.
- In previous reports there was a single model for both houses and flats combined, whilst in this report there are separate models for houses and flats (for reasons explained in the “Separate models for houses and flats” section on page 3). The R^2 value for the combined model may be artificially higher. This is because the difference between houses and flats is included in the data, and this additional variability is explained by the property type variable in the model.

However, if models are built for total gas use, using all the same explanatory variables as were included in the models in this report, along with floor area, this gives the following R^2 values:

- an R^2 value of 38.0% for houses (similar to the R^2 values given in previous reports[19])
- an R^2 value of 16.6% for flats

The lower R^2 value for flats, reflects the fact that that gas use is less correlated with floor area in flats, than it is in houses. For flats the recorded floor area may not give as reliable an indication of the amount of gas required to heat the available space. This is because, unlike for houses, for flats the recorded floor area does not include the entire footprint of the house (see the *Separate models for houses and flats* section on page 4).

Comparing old and new approaches

Old approach: Previous reports have presented models for household gas use which have included floor area as an explanatory variable and have excluded variables correlated with floor area from the model.

New approach: In this report, models were instead built for gas use per m^2 of floor area. Floor area was not included as an explanatory variable in the models and so variables correlated with floor area could be included in the models.

As explained on page 7, it is not possible to compare R^2 values from the models for gas use per m^2 with the R^2 values from the models in previous reports. However, by using the same explanatory variables included in the models in this report, the old approach can be replicated. This is to compare how well the two approaches fit the observed values for total gas use. For the new approach the fitted values given by the regression model are for gas use per m^2. These can be multiplied by the floor area to give the fitted values for total gas use.
The root mean squared error (RMSE) for the fitted values for gas use from the models is a measure of how much on average they differ from the actual gas use values in the data (lower is better):

- For houses the new approach gives an RMSE of 5,799 kWh, slightly higher (2% higher) than the RMSE for the old approach (5,684 kWh).
- For flats the new approach gives an RMSE of 4,910 kWh, similar to the RMSE for the old approach (4,899 kWh).

[19] The Summary of Katalysis Regression:
www.gov.uk/government/uploads/system/uploads/attachment_data/file/65975/6870-need-report-annex-f.pdf
The Summary of NERA work:
www.gov.uk/government/uploads/system/uploads/attachment_data/file/65974/6869-need-report-annex-e.pdf
Predicting Gas Consumption report:
www.gov.uk/government/uploads/system/uploads/attachment_data/file/532538/Annex_C_Predicting_gas_consumption.pdf

An alternative measure is the mean percentage difference between the fitted and actual values for household gas use:

- For houses, the fitted values for gas use differed from the actual values by 37.7% on average. This compares to the fitted values differing by 37.8% on average, under the old approach.
- For flats, the fitted values for gas use differed from the actual values by 48.6% on average. This compares to the fitted values differing by 49.8% on average, under the old approach.

So to summarise, the old and new approaches give comparable results in terms of how well they fit the data. However, the new approach has advantages in terms of the interpretability of the results (these are discussed in the “Controlling for the size of properties” section on page 3 and 4).

Random and mixed effects models for houses

For houses, as well as a linear regression model, random and mixed effects models were also used to provide further insights about the variance in gas use per m^2 between and within LSOAs[20].

The interclass correlation coefficient

What is of interest here is whether the LSOA a household is in has a bearing on its gas use per m^2. This can be assessed by the Interclass Correlation Coefficient (ICC).

Interclass Correlation Coefficient (ICC)

Suppose there is a population of units (e.g. households) for which some characteristic y is being considered (e.g. gas use per m^2). Also suppose that the population can be split into a number of groups (e.g. LSOAs). Then the ICC is defined as:

$$ICC = \frac{\sigma^2_{between\ groups}}{\sigma^2_{between\ groups} + \sigma^2_{within\ groups}}$$

where:

$\sigma^2_{between\ groups}$ = the variance of y between households within the same group

$\sigma^2_{withi\quad groups}$ = the variance in the mean of y between groups

Suppose the LSOA a household was in had a large bearing on how much gas it used per m^2. This would mean that there was a high degree of homogeneity in gas use per m^2 between the households within each LSOA (relative to households in general). A large proportion of the variability in gas use per m^2 would then be attributable to systematic differences between LSOAs. The ICC would then be close to 1.

[20] Output areas were built from adjacent postcode units and were designed to be as socially homogenous as possible. Lower super output areas (LSOAs) are in turn built from adjacent output areas, and in England and Wales these contain 400-1,200 households. See: www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography#super-output-area-soa

On the other hand, the LSOA the household is in might have virtually no bearing on its gas use per m^2. In such a case, the variance between houses within the same LSOA would be indistinguishable from the total variance between households in general. In this situation the estimated variance between LSOAs would then be negligible in comparison, resulting in an ICC close to 0.

The ICC therefore simultaneously provides both a measure of:

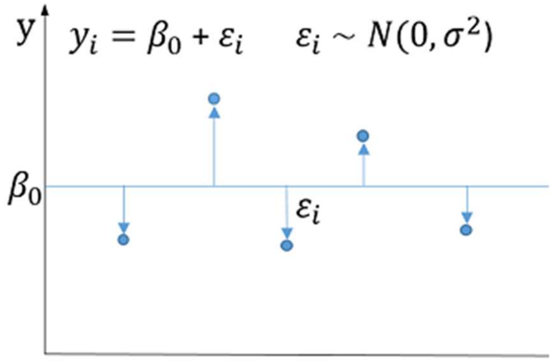
a. the degree of similarity in gas use per m2 between houses within the same LSOA relative to houses in general
b. the proportion of variance in gas use per m2 between houses that is attributable to systematic differences between LSOAs

Fixed and random effects

In order to set the background for explaining what random and mixed effects models are, the components which make up linear regression models are first considered. In a linear regression model, as well as the explanatory variables there is also an error term to reflect the differences between the actual values of the variable being modelled (the y variable, gas use per m^2 in this case) and the fitted values given by the model (see Figure 1b for an example with one explanatory variable).

Figure 12: Linear regression models

12a: Linear regression – No explanatory variables



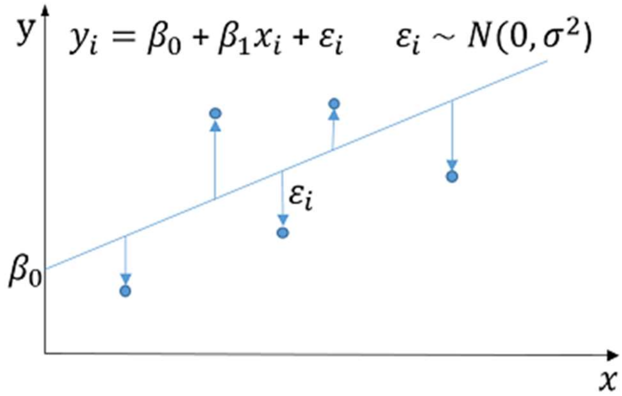
Fixed: β_0 (the overall mean)

Random:

ε_i = The difference between the y-value for the ith household (y_i) and the overall mean (β_0) . These are assumed to be independent and normally distributed with mean 0 and variance σ^2.

Parameters estimated by the model: β_0, σ^2

12b: Linear regression – One explanatory variable



Fixed: β_0 (the intercept), β_1 (the gradient)

Random:

ε_i = The difference between the actual y-value for the ith household (y_i) and the fitted value y-value for the ith household ($\beta_0 + \beta_1 x_i$). These are assumed to be normally distributed with mean 0 and variance σ^2.

Parameters estimated by the model: β_0, β_1, σ^2

When no explanatory variables are included (see Figure 1a) the error term reflects the variability in the y-variable about the mean.

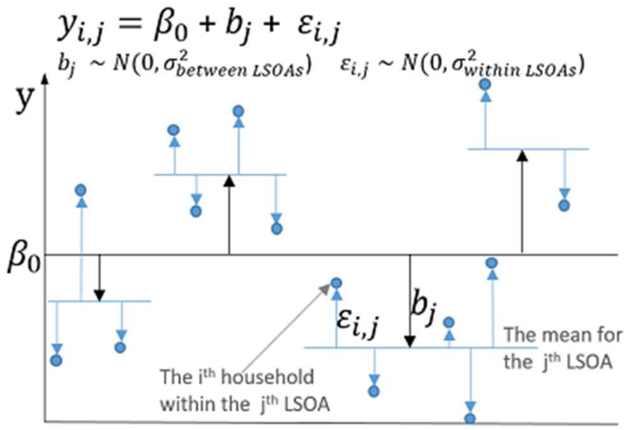
The explanatory variables may be referred to as fixed effects and the error term may be referred to as a random effect.

The random effects model

Suppose a linear regression model was fitted for gas use per m^2, with “LSOA” as an explanatory variable. The model would use the data to produce an estimated effect for each of the more than 30,000 LSOAs in England and Wales (apart from one, which it would use as a baseline to compare the rest). These estimates would be unreliable for the LSOAs with relatively few houses remaining in the analysis after exclusions.

Figure 13: Random and mixed effects models

13a: Random effects model



Fixed:

β_0 (the overall mean)

Random:

b_j = The difference between the mean for the jth LSOA ($\beta_0 + b_j$) and the overall mean (β_0). These are assumed to be independent and normally distributed with mean 0 and variance $\sigma^2_{between\ LSOAs}$.

$\varepsilon_{i,j}$ = The difference between the y-value for the ith household, within the jth LSOA ($y_{i,j}$) and the mean for the jth LSOA ($\beta_0 + b_j$). These are assumed to be independent and normally distributed with mean 0 and variance $\sigma^2_{within\ LSOAs}$.

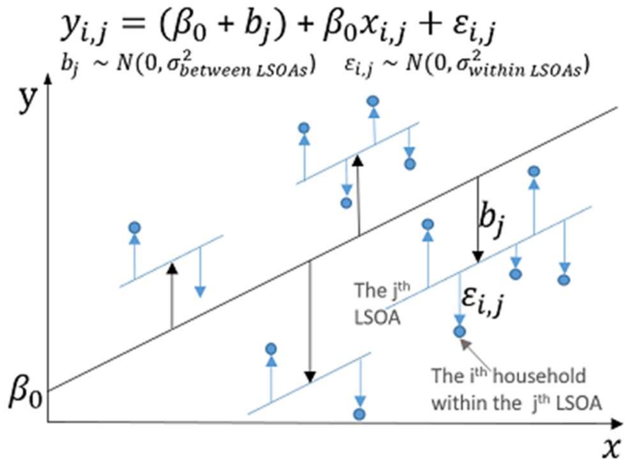
Parameters estimatd by the model:

β_0, $\sigma^2_{between\ LSOAs}$, $\sigma^2_{within\ LSOAs}$

Interclass correlation coefficient

ICC = $\frac{\sigma^2_{between\ LSOAs}}{\sigma^2_{between\ LSOAs} + \sigma^2_{within\ LSOAs}}$

13b: Random intercepts model



Fixed:

β_0 (the overall intercept),

β_1 (the gradient)

Random:

b_j = The difference between and the intercept for the jth LSOA ($\beta_0 + b_j$) and the overall intercept (β_0). These are assumed to be independent and normally distributed with mean 0 and variance $\sigma^2_{between\ LSOAs}$.

$\varepsilon_{i,j}$ = The difference between the y-value for the ith household, within the jth LSOA ($y_{i,j}$) and the expected y-value given the LSOA it is in (($\beta_0 + b_j$) + $b_j x_{i,j}$). These are assumed to indpendent and normally distributed with mean 0 and variance $\sigma^2_{within\ LSOAs}$.

Parameters estimated by the model:

β_0, β_1, $\sigma^2_{between\ LSOAs}$, $\sigma^2_{within\ LSOAs}$

Residual interclass correlation coefficient

ICC = $\frac{\sigma^2_{between\ LSOAs}}{\sigma^2_{between\ LSOAs} + \sigma^2_{within\ LSOAs}}$

Instead, “LSOA” could be included in the model as a random effect. There would then be two random effects in the model to represent:

a. the variability in gas use per m^2, between LSOAs
b. the variability in gas use per m^2, between houses within the same LSOA

Such a model, only having random effects and no explanatory variables (illustrated in Figure 13a), is an example of a random effects model. This model would not estimate the effects of individual LSOAs but instead would estimate the between and within-LSOA variances. This would allow the ICC to be estimated.

Mixed effects model

To the random effects model for gas use per m^2 in houses, the explanatory variables from the linear regression model were added to gives a basic mixed effects model (a model containing both fixed explanatory variables and random effect). This mixed effects[21] model is known as a random intercepts model and it is illustrated in Figure 13b (shown with only 1 explanatory variable for simplicity).

The residual ICC (as defined in Figure 2.2) provided a measure of:

a. the degree of similarity in gas use per m2 between houses within the same LSOA relative to houses in general, after controlling for (removing the effects of) the explanatory variables added to the model
b. the proportion of the unexplained variance in gas use per m2 that is attributable to systematic differences between LSOAs

If there were relatively large systematic differences between the LSOAs, then controlling for this between-LSOA variability, could have resulted in substantial changes to the parameter estimates for the explanatory variables. Fitting the mixed effects model showed that this was not the case. The parameter estimates for the explanatory variables, given by the mixed effects model were broadly the same as those given by the linear regression model.

[21] Mixed effects models are also referred to as hierarchical or multilevel models.

Appendix A: Linear regression models for annual household gas use against floor area

A1: Houses

	Estimated effects on annual gas use (kWh) per m^2		
	Intercept	Effect of every addition m^2 of floor area	The R^2 value of the model
No intercept model (assuming gas use is directly proportional to floor area)	0.0	+130.8	22.3%
Best fitting model (including an intercept)	4190.0	+96.0	26.2%

A2: Flats

	Estimated effects on annual gas use (kWh) per m^2		
	Intercept	Effect of every addition m^2 of floor area	The R^2 value of the model
No intercept model (assuming gas use is directly proportional to floor area)	0.0	+172.4	1.4%
Best fitting model (including an intercept)	4322.4	+92.5	8.4%

Appendix B: The linear regression model for annual household gas use per m^2 of floor area, in houses

	Effect on annual gas use (kWh) per m^2		Number of households included in the analysis
	Estimate	Standard error	
Intercept	159.0	0.2	11,251,453
Property characteristics			
Type of house			
Mid-terrace (baseline)	0.0	0.0	2,780,497
End-terrace	+13.3	0.1	1,343,417
Semi-detached	+15.7	0.0	4,228,125
Detached	+26.0	0.1	2,899,414
Number of floors			
1	+21.4	0.1	1,235,647
2 (baseline)	0.0	0.0	9,388,558
3 or more	+2.1	0.1	627,248
Number of bedrooms			
1 (baseline)	0.0	0.0	210,606
2	-16.8	0.1	2,548,834
3	-21.9	0.1	6,456,331
4	-27.6	0.1	1,692,851
5 or more	-33.5	0.2	342,831
Number of bathrooms			
1 (baseline)	0.0	0.0	9,654,255
2	-3.7	0.1	1,387,200
3 or more	-6.7	0.1	209,998

	Effect on annual gas use (kWh) per m^2		Number of households included in the analysis
	Estimate	Standard error	
Age of house			
Pre 1919 (baseline)	0.0	0.0	2,041,991
1919-44	+2.3	0.1	2,329,304
1945-64	-5.8	0.1	2,291,143
1965-82	-15.2	0.1	2,357,614
1983-92	-23.6	0.1	833,385
1993-99	-27.1	0.1	641,955
2000-2011	-36.0	0.1	663,765
2012 onwards	-57.8	0.2	92,296
Energy efficiency measures taken as a part of government schemes Baseline: In each case all houses not recorded as having had the given measure installed			
Cavity wall insulation	-7.6	0.0	3,037,343
Solid wall insulation	-18.9	0.2	94,558
Loft insulation	-3.8	0.0	2,987,336
Boiler	-6.2	0.0	4,589,936
Other relevant measure	-1.3	0.0	2,163,996

	Effect on annual gas use (kWh) per m²		Number of households included in the analysis
	Estimate	Standard error	
Household characteristics			
Number of adults in the house per 100 m²			
Every extra adult per 100 m²	+8.6	0.0	
Household income			
Less than £20,000 (baseline)	0.0	0.0	2,298,707
£20,000 - £29,999	-1.2	0.1	2,249,275
£30,000 - £39,999	-1.1	0.1	2,028,552
£40,000 - £49,999	+0.8	0.1	1,650,810
£50,000 - £59,999	+2.3	0.1	994,352
£60,000 - £99,999	+5.0	0.1	1,479,600
Over £100,000	+9.8	0.1	550,157
Length of residency (of the resident who has lived in the house the longest)			
Less than 1 year	+3.7	0.2	54,923
1 years	-3.0	0.1	406,798
2 years	-4.9	0.1	623,772
3 years	-6.5	0.1	580,213
4 years	-7.2	0.1	486,670
5 years	-7.3	0.1	494,837
6 years	-7.0	0.1	431,128
7 years	-6.0	0.1	500,876
8 years	-4.4	0.1	399,567
9 years	-3.4	0.1	396,001
10 years	-2.1	0.1	330,917
11 years or more (baseline)	0.0	0.0	6,545,751
Tenure			
Owner occupied (baseline)	0.0	0.0	8,444,395
Privately rented	-1.0	0.1	1,239,353
Council/Housing association	-8.9	0.1	1,567,705

	Effect on annual gas use (kWh) per m²		Number of households included in the analysis
	Estimate	Standard error	
Local area characteristics			
Region			
North East (baseline)	0.0	0.0	603,210
North West	-11.4	0.1	1,579,625
Yorkshire and The Humber	-2.8	0.1	1,173,869
East Midlands	-7.5	0.1	1,011,029
West Midlands	-6.5	0.1	1,159,367
East of England	-9.7	0.1	1,134,934
Wales	-17.5	0.1	602,244
London	-5.1	0.1	1,079,538
South East	-13.3	0.1	1,926,524
South West	-26.5	0.1	981,113
Urban/ rural			
Conurbation (baseline)	0.0	0.0	3,804,916
City or town	-7.0	0.0	5,768,728
Rural	-5.4	0.1	1,677,809
Output area classification			
Multicultural metropolitans	+5.0	0.1	1,537,376
Ethnicity central	-3.1	0.1	194,397
Suburbanites	+1.6	0.1	3,202,381
Urbanities (baseline)	0.0	0.0	2,339,638
Cosmopolitans	-1.8	0.1	182,405
Rural residents	-3.1	0.1	759,027
Hard-pressed living	-5.3	0.1	2,490,710
Constrained city dwellers	-6.9	0.1	545,519
R^2 = 15.6%			

Appendix C: The linear regression model for annual household gas use per m^2 of floor area, in flats

	Effect on annual gas use (kWh) per m^2		Number of households included in the analysis
	Estimate	Standard error	
Intercept	+211.6	0.5	1,336,741
Property characteristics			
Type of flat			
Purpose built flat (baseline)	0.0	0.0	1,141,326
Converted flat	+22.7	0.4	195,415
Number of bedrooms			
1 (baseline)	0.0	0.0	462,030
2	-17.7	0.2	724,208
3 or more	-23.6	0.3	150,503
Number of bathrooms			
At most 1 (baseline)	0.0	0.0	1,252,204
2 or more	-4.7	0.4	84,537
The floor of the flat			
Below ground floor	-8.7	1.0	12,822
Ground floor (baseline)	0.0	0.0	518,262
Floor 1	-12.6	0.2	517,447
Floor 2	-19.4	0.3	184,815
Floor 3	-20.8	0.5	49,202
Floor 4 or higher	-16.1	0.5	54,193

	Effect on annual gas use (kWh) per m^2		Number of households included in the analysis
	Estimate	Standard error	
Age of flat			
Pre 1919 (baseline)	0.0	0.0	217,133
1919-44	+0.7	0.4	119,400
1945-64	-14.6	0.4	330,520
1965-82	-26.7	0.4	362,771
1983-92	-31.4	0.5	95,283
1993-99	-38.1	0.6	45,309
2000-2011	-48.2	0.5	152,044
2012 onwards	-62.9	1.0	14,281
Energy efficiency measures taken as a part of government schemes Baseline: In each case all houses not recorded as having had the given measure installed			
Cavity wall insulation	-7.6	0.3	203,759
Solid wall insulation	-25.5	0.9	15,058
Loft insulation	-0.6	0.4	91,067
Boiler	-5.5	0.2	570,718
Other relevant measure	-1.1	0.3	158,358

	Effect on annual gas use (kWh) per m^2		Number of households included in the analysis
	Estimate	**Standard error**	
Household characteristics			
Number of adults in the flat per 100 m^2			
Every extra adult per 100 m^2	+8.6	0.0	
Household income			
Less than £20,000 (baseline)	0.0	0.0	429,373
£20,000 - £29,999	-1.6	0.3	276,546
£30,000 - £39,999	+1.7	0.3	231,532
£40,000 - £49,999	+1.3	0.3	149,874
£50,000 - £59,999	+2.5	0.4	84,816
£60,000 - £99,999	+2.5	0.4	135,177
Over £100,000	+6.6	0.7	29,423
Length of residency (of the resident who has lived in the flat the longest)			
Less than 1 year	+6.0	0.7	24,250
1 years	-2.1	0.4	116,214
2 years	-5.6	0.3	144,697
3 years	-7.1	0.4	118,591
4 years	-8.5	0.4	98,131
5 years	-7.2	0.4	93,576
6 years	-7.8	0.4	86,434
7 years	-6.0	0.4	87,644
8 years	-5.1	0.5	56,856
9 years	-4.2	0.5	49,128
10 years	-2.1	0.5	44,550
11 years or more (baseline)	0.0	0.0	416,670
Tenure			
Owner occupied (baseline)	0.0	0.0	407,843
Privately rented	-1.7	0.3	361,725
Council/Housing association	-7.5	0.3	567,173

	Effect on annual gas use (kWh) per m^2		Number of households included in the analysis
	Estimate	**Standard error**	
Local area characteristics			
Urban/ Rural			
Conurbation (baseline)	0.0	0.0	741,524
City or town	-12.9	0.2	524,633
Rural	-9.2	0.5	70,584
Output area classification			
Multicultural metropolitans	+9.5	0.3	197,990
Ethnicity central	+4.4	0.3	326,253
Suburbanites	+2.2	0.7	27,745
Urbanites (baseline)	0.0	0.0	256,186
Cosmopolitans	-6.9	0.4	138,173
Rural residents	-3.0	0.9	14,255
Hard-pressed living	-8.8	0.4	109,833
Constrained city dwellers	-10.2	0.3	266,306
R^2 = 11.4%			

This publication is available from: https://www.gov.uk/government/collections/national-energy-efficiency-data-need-framework

If you need a version of this document in a more accessible format, please email energyefficiency.stats@beis.gov.uk. Please tell us what format you need. It will help us if you say what assistive technology you use.