Article

Developing admin-based property floor area statistics for England and Wales: 2021

Research exploring the feasibility of using Valuation Office Agency (VOA) and Energy Performance Certificate (EPC) data to produce a harmonised measure of property size (floor area) across property types for England and Wales

Contact:
Shannon Mahon, Andreea Butnaru, Emily Mason-Apps, Joseph Herson, Stephan Tietz
Admin.Based.
Characteristics@ons.gov.uk
+44 1329 444528

Notice

26 October 2022

This article outlines new methods for developing property floor area statistics, building on our previous methodology, Admin-based statistics for property floor space, feasibility research: England and Wales.

Table of contents

1 . Main points

- We used two administrative data sources to explore the feasibility of producing a harmonised measure of floor area for residential properties in England and Wales that could be used to develop an alternative measure of overcrowding that considers living space per person.

- Valuation Office Agency (VOA) data cover residential properties in England and Wales, but have two different methods for measuring floor area depending on the property type; and Energy Performance Certificate (EPC) data use one measure of floor area across all property types, but only 57.2% of residential properties have an EPC (and could be linked to VOA data).

- We used regression modelling to test if we could utilise the geographical completeness of VOA data to predict EPC floor area; this would provide a harmonised measure of floor area for residential addresses in England and Wales across all property types.

- The best-performing multiple linear regression model used VOA floor area (square meters), VOA floor area measure flag, and country to predict EPC floor area, and predicts 31% of properties within 5% of their actual floor area, 60% within 10%, and 84% within 20% of their actual floor area.

- We conclude that the model does not produce harmonised address-level floor area estimates of high enough statistical quality to provide an alternative measure of overcrowding that is based on available living space per person and is comparable across all property types.

- Initial findings suggest that the variance observed in the models is caused by quality of the EPC floor area variable when used for statistical purposes; and differences in property structure that cannot be fully accounted for in the data available.

- Without being able to harmonise the two different VOA floor area measures, caution needs to be taken when comparing available living space across different property types; however, floor area can still be compared for properties of the same type.

These research outputs are not official statistics on floor area in England and Wales and should not be used for policymaking or decision-making. They are published to allow users to compare the VOA data with the EPC data as an alternative approach to collecting information on floor area. We advise caution when using the data.

2 . About our transformation research

At the Office for National Statistics (ONS), we are [exploring the use of administrative data on housing](#). Until now, information about floor area has been collected through surveys such as the [English Housing Survey](#) and [Welsh Housing Conditions Survey](#). However, because of sample size, analysis of floor area for sub-regional geographies has been limited. We are exploring the feasibility of using administrative data to provide detailed information on floor area down to small geographies across England and Wales. This may help housing planners and policymakers to better understand the characteristics of the dwelling stock in their areas and therefore better meet the future housing needs of local residents ([see the Census 2021 topic consultation](#)).

This research is a progression of our [previous research into property floor area](#) and explores the feasibility of using administrative data to produce a harmonised measure of floor area for residential properties in England and Wales that could be used to develop an alternative measure of overcrowding that considers living space per person, instead of using occupancy ratings (see [Section 3, Measuring available living space across property types](#)).

This research forms part of our [population and social statistics transformation programme](#), which aims to provide the best insights on population, migration and society using a range of data sources. The findings will form part of the evidence base for the National Statistician's recommendation in 2023 on the future of population, migration and social statistics in England and Wales.

3 . Measuring available living space across property types

An important aspect of housing policy when assessing living conditions is the amount of living space available to a household. Accommodation that does not provide enough space for a household of a given size is considered overcrowded, as defined in the [Overcrowded housing research briefing from the House of Commons Library](#).

Overcrowding is often measured using occupancy ratings, usually [the room standard, as defined in the Housing Act 1985](#) or [the bedroom standard, as defined in 2012 by the Department for Communities and Local Government (DCLG)](#), now the Department for Levelling Up, Housing and Communities (DLUHC). These measures do not consider that rooms can vary in size, or that the actual use of a room may be different to the intended and recorded use (for example, a bedroom converted to home office would still be counted as a bedroom using the bedroom standard). Measuring available living space using floor area could be one way to better reflect the diversity of living conditions through an alternative measure of overcrowding that measures available living space per person. For more information on overcrowding measures, see our previous research [on deriving occupancy ratings from VOA number of rooms](#), and [on deriving the bedroom standard using VOA number of bedrooms](#).

Valuation Office Agency property characteristics data

Valuation Office Agency (VOA) data cover residential properties within England and Wales and includes a measurement of each property's floor area. However, the VOA have two distinct ways of measuring floor area depending on the type of property being measured: Reduced Cover Area (RCA) is used for houses and bungalows; and Effective Floor Area (EFA) is used for flats and maisonettes.

RCA includes external walls, and areas such as hallways, landings and passages in the measurements, so we would expect this method to typically overestimate the available living space compared with the EFA method, which measures the usable area of the rooms to the internal face of the walls of the property. A description of what is included in each measure can be found in [Section 7, Glossary](#).

Energy Performance Certificate data

An Energy Performance Certificate (EPC) provides a measure of the energy efficiency of properties within England and Wales and includes a measurement of each property's floor area. In March 2022 around 60% of properties in England and just less than 60% in Wales had an EPC, as noted in this [release from the DLUHC](#). In contrast to the VOA data, EPC data only use the Total Floor Area (TFA) method to measure the floor area for all property types. TFA is measured to the internal face of the external walls and only includes areas that are heated, habitable and internally accessible from the main dwelling, meaning the measure more closely represents available living space. It also enables comparison of floor area across all property types. A breakdown of what is included in this measure can be found in [Section 7, Glossary](#).

Producing a single measure of available living space

Because of these differences, we would expect the TFA from EPC data to be smaller than the RCA measure from VOA for houses and bungalows, and greater than the EFA measure for flats and maisonettes, with some variation.

This research explores the feasibility of producing a statistical model (see [Section 5, Method for harmonising floor area measures](#)) that uses the geographical completeness of VOA floor area measures (RCA and EFA) to predict the EPC floor area measure (TFA) with the aim to produce a single measure of available living space for residential properties in England and Wales.

4 . Data used to harmonise VOA floor area measures

We used Valuation Office Agency (VOA) data linked to Energy Performance Certificate (EPC) data to explore the feasibility of producing a single measure of available living space. [Section 8, Data sources and quality](#), details the data cleaning and linkage steps.

Ensuring representativeness of linked VOA and EPC addresses

VOA data should cover all residential properties in England and Wales. At the time of this research, only 57.2% of residential properties have an EPC and could be linked to VOA data. To understand the representativeness of EPC data, the Office for National Statistics (ONS) is collaborating with the Department for Levelling Up, Housing and Communities (DLUHC). DLUHC report initial exploratory work in their statistical release presenting Experimental Official Statistics based on Energy Performance Certificates (EPCs). The ONS is planning to include a section about the representativeness of EPC data in the next annual Energy Efficiency of Housing publication.

We looked at linkage rates of VOA addresses to EPC addresses by property types to ensure that the linked dataset was representative of all residential addresses across England and Wales. Table 1 shows that flats have the highest linkage rates for both England and Wales. The lowest rate of linkage for each country is for "other" properties (such as caravans), likely because such properties do not require an EPC as they are exempt if used for holiday lets (for further exemptions see the EPC information page). Linkage rates by property type show a similar distribution for England and Wales, with slightly lower linkage rates for Wales. Overall, these findings indicate that the linked dataset is representative of all four of the main property types (houses, bungalows, maisonettes and flats) across England and Wales.

Table 1: Linkage rates of VOA and EPC addresses for England and Wales by VOA property type

VOA property type	VOA addresses linked to EPC for England (%)	VOA addresses linked to EPC for Wales (%)
House	54.9	52.4
Bungalow	53.3	49.3
Maisonette	57.7	54.5
Flat	68.8	66.7
Other	9.7	8.1
Missing	45.7	44.1
Total	57.5	53.7

Source: Valuation Office Agency and Department for Levelling Up, Housing and Communities

As shown in Table 2, there is a slightly higher proportion of houses and bungalows in the unlinked addresses than in the linked addresses, and a slightly lower proportion of flats, but the general pattern of distribution is similar. Looking at the distributions by country in Table 3, there is only a minimal difference in the percentages of linked and unlinked addresses between England and Wales, which suggests that the linked VOA and EPC data are reasonably representative of all residential addresses.

Table 2: Distribution of linked and unlinked addresses by VOA property type

VOA property type	VOA addresses linked to EPC (%)	Unlinked VOA addresses (%)
House	63.3	70.1
Bungalow	8.7	10.3
Maisonette	1.8	1.8
Flat	25.5	15.6
Other	0.1	1.3
Missing	0.6	0.9

Source: Valuation Office Agency and Department for Levelling Up, Housing and Communities

Table 3: Distribution of linked and unlinked addresses by country

Country	VOA addresses linked to EPC (%)	Unlinked VOA addresses (%)
England	94.8	94.0
Wales	5.2	6.0

Source: Valuation Office Agency and Department for Levelling Up, Housing and Communities

Comparing VOA and EPC floor area and property type information

We undertook a series of steps to clean the linked dataset (see Section 8, Data sources and quality), including removing linked addresses where the property type was "other" or missing. The cleaned and linked VOA and EPC data were used for all the following analysis.

For both houses and bungalows in England and Wales, the median VOA floor area in square meters (sqm) is greater than the EPC floor area (see Table 4). Maisonettes and flats in both England and Wales show the opposite pattern, with the median EPC floor area being greater than the VOA floor area. This is in line with what we would predict, considering what the floor area measurement methods include and exclude (see Section 7, Glossary).

Table 4: Median VOA and EPC floor area by VOA property type for England and Wales

Country	VOA property type	VOA (sqm)	EPC (sqm)	Difference (VOA – EPC) (sqm)
England and Wales				
	House	99.0	88.0	11.0
	Bungalow	78.0	71.4	6.6
	Maisonette	57.0	76.6	-19.6
	Flat	43.0	55.0	-12.0
	Overall	87.0	79.0	8.0
England				
	House	98.0	88.0	10.0
	Bungalow	78.0	71.0	7.0
	Maisonette	57.0	76.7	-19.7
	Flat	43.0	55.0	-12.0
	Overall	87.0	78.4	8.6
Wales				
	House	101.0	88.0	13.0
	Bungalow	85.0	78.0	7.0
	Maisonette	60.0	75.0	-15.0
	Flat	42.0	54.0	-12.0
	Overall	95.0	83.0	12.0

Source: Valuation Office Agency and Department for Levelling Up, Housing and Communities

Notes

1. Addresses where the EPC or VOA property type category was missing or "other" were removed from the linked and cleaned VOA and EPC dataset (see Section 8 Data sources and quality).

Overall, the median floor area on both the VOA and EPC is greater in Wales compared with England. For houses and bungalows, the median floor area tends to be greater in Wales than in England. For flats, the median floor area is greater in England compared with Wales on both the VOA and EPC. For maisonettes however, the median floor area is greater in Wales compared with England according to the VOA, but greater in England compared with Wales according to the EPC. These differences suggest that our model should include a geographical dimension (see Section 5, Method for harmonising floor area measures).

To explore these differences further, we looked at the agreement rates of property types between the VOA and EPC data. The agreement rates for houses, bungalows and flats were high (over 93% for all), however the agreement rate for maisonettes was low, at just 55.2%. This is possibly because the Government's Standard Assessment Procedure for Energy Rating of Dwellings (PDF, 2.48MB) states that EPC surveyors do not need to distinguish between a flat and maisonette regarding calculations, and can "select either type as definitions vary across the UK".

Because VOA measure the floor area of flats and maisonettes using the same method, Effective Floor Area (EFA), we also looked at the agreement rates when grouping the property types by the VOA method used. For property types measured using EFA, we found an agreement rate of 98.4%. For Reduced Cover Area (RCA), used to measure houses and bungalows, we found an agreement rate of 99.7%. This suggests that statistical models may benefit from grouping VOA property type according to the VOA floor area method that would be used.

5 . Method for harmonising floor area measures

We used regression modelling to test if we could use the two different floor area measures in Valuation Office Agency (VOA), (see Section 3, Measuring available living space across property types) data alongside other VOA property characteristics to predict Energy Performance Certificate (EPC) floor area. Our aim was to find a model that predicts EPC floor area accurately enough that it could then be used to calculate the EPC floor area for VOA addresses that do not have an EPC. This would provide a harmonised measure of floor area for residential addresses in England and Wales across all property types.

Selecting the best VOA variables to predict EPC floor area

We explored a simple linear regression model using VOA floor area as the primary predictor for EPC floor area. This model produced an R-squared of 0.84, providing a benchmark for further analysis (see Table 5). An R-squared of 1 would indicate that VOA floor area can explain all of the variability of the EPC floor area, while an R-squared of 0 would indicate that the model explains none of the variability.

An important aim for the final model was to enable comparison of floor area across different property types (see Section 3, Measuring available living space across property types), and between England and Wales (see Section 4, Data used to harmonise VOA floor area measures). We therefore produced a series of simple linear regression models to assess if this relationship held when the data was split by VOA property type and then by country.

The R-squared for each subgroup when the data is split by property type (house, bungalow, maisonette and flat) varied between 0.54 (for maisonettes) to 0.85 (for houses) indicating prediction is not equally successful across different property types. Improved R-squared values when grouping property types according to the VOA floor area measure used (EFA and RCA) again suggested that the model may benefit from grouping property types together (see Table 5). This grouping is referred to as the "VOA floor area measure flag".

Other than VOA floor area, VOA property type, VOA floor area measure flag, several other property characteristic variables from the VOA dataset, which could have some bearing on property size, were then considered for inclusion in a multiple linear regression model. Out of these, number of rooms and number of bedrooms were too strongly correlated to VOA floor area and were therefore not included in any models. The inclusion of number of bathrooms and property age led to no notable improvement or reduction in R-squared.

The results by country (England and Wales) show no change for England, and a slightly lower R-squared for Wales (see Table 5). Other geographical variables were also explored as predictive variables in a multiple linear regression model, such as Government Office Regions, Rural Urban Classification and Local Authority. These models revealed no notable improvement or reduction in R-squared, so only country was used to keep the model as simple as possible, maintaining model parsimony.

Table 5: Parameters of the simple linear regression models with VOA floor area as a predictor of EPC floor area for the full linked dataset, then split by VOA floor area measure flag and country

		Coefficients	Intercept	R-squared
VOA floor area		0.82	12.67	0.84
VOA floor area measure flag				
	RCA (houses and bungalows)	0.92	-1.36	0.85
	EFA (flats and maisonettes)	1.05	10.14	0.68
Country				
	England	0.82	12.76	0.84
	Wales	0.82	10.17	0.82

Source: Valuation Office Agency and Department for Levelling Up, Housing and Communities

Best-performing multiple linear regression model

The best-performing multiple linear regression model uses VOA floor area, VOA floor area measure flag (RCA or EFA) and country (England and Wales) as predictor variables, and EPC floor area as the outcome variable (see Table 6). Compared with the original simple linear regression model, this model produced an improved adjusted R-squared of 0.86. The adjusted R-squared should be used when comparing with the R-squared from the simple linear regressions as it accounts for the use of multiple predictor variables. To evaluate the estimator performance of the model we performed k-fold cross-validation using k = 10. The R-squared for all k-folds was consistent with the original model.

Table 6: Results of multiple linear regression analysis using VOA floor area, VOA floor area measure flag (RCA or EFA) and country to predict EPC floor area

			Coefficients			
RMSE	R-squared	Adjusted R-squared	VOA floor area	VOA floor area measure flag	Country	Intercept
14.61	0.86	0.86	0.93	-17.91	1.85	14.02

Source: Valuation Office Agency and Department for Levelling Up, Housing and Communities

Notes

1. Reference category for VOA floor area measure flag was EFA (flats and maisonettes) and for country it was Wales.

We checked if the assumptions for multiple linear regressions (linearity, multicollinearity, homoscedasticity, and multivariate normality) were met by the final model. Residuals appear reasonably evenly distributed around 0 square meters (sqm), with a small left skew indicating that the model has a slight tendency to overestimate floor area. A plot of the standardised residuals against the predicted EPC floor area revealed a violation of the homoscedasticity assumption (homogeneity of variances), suggesting that larger properties might be having an undue effect on the analysis. Scatter plots of all predictor variables against EPC floor area showed a linear relationship, but also revealed a skew implying larger properties may be having an unequal effect on the model.

We therefore applied a log-transformation to the floor area variables before running our best-performing multiple linear regression model again. This improved the linearity and distribution of residuals for larger properties, but increased the skew for smaller properties. The log-transformed model resulted in a slightly reduced R-squared of 0.84, as well as increasing the correlation between two of the predictor variables (VOA floor area and VOA floor area measure flag) from 0.61 to 0.76. For these reasons and the reduced interpretability of the results from the log-transformed regression model, the performance of the final model was assessed without log-transformation.

Performance of final multiple linear regression model

To assess the performance of the final multiple linear regression model in more detail, we looked at the mean, median, and standard deviation of residuals by property types and country. For houses and flats, the means and medians were all close to 0sqm, suggesting that the model would allow comparisons across these groups. However, for maisonettes the mean of residuals ranged from negative 7.16sqm (for England) to negative 5.51 sqm (for Wales), with medians of negative 7.55sqm for Wales and negative 8.21sqm for England. The means and medians for bungalows were also higher, making comparisons to these property groups more challenging. The small overall differences in predictions between England and Wales suggested that cross-country comparisons would also be possible. It should be noted, however, that the standard deviation for the mean was greater than the mean and median in all groups (ranging between 11.0sqm and 20.2sqm), indicating a large degree of variation within groups.

We also looked at the distribution of residuals and found that the model predicts 41% of addresses across England and Wales within 5sqm, 69% within 10sqm and 88% within 20sqm. It is important to note that our intended use of this model is ultimately to enable overcrowding analysis. The residuals would be too large to accurately assess the levels of overcrowding in a small property, and small properties are most likely to be of importance when assessing overcrowding. Only 31% of properties are predicted within 5% of their actual floor area, 60% within 10%, and 84% within 20% of their actual floor area. We conclude that, at this point, the model does not produce harmonised address-level floor area estimates of high enough statistical quality to provide an alternative measure of overcrowding that focuses on available living space per person and allows comparisons across all property types.

Sources of variance in the floor area data

We hypothesise that there are two primary factors causing the variance observed in the models: data quality; and differences in property structure that cannot be fully accounted for in the data available.

Research conducted by ONS' Methodological Research Hub using structural equation modelling to estimate the measurement error of linked VOA and EPC data estimates that EPC data tends to have a larger measurement error for the floor area variable (6.5%) compared with VOA data (2.5%). This pattern is observed across most local authorities (LAs) in England and Wales, apart from eight LAs within Greater London, Isles of Scilly, Isle of Anglesey and two more rural LAs (Gwynedd and South Hams). These findings suggest that improvements in the quality of the floor area variable in the EPC data would improve the performance of the model.

The English Housing Survey (EHS) collects and publishes data on floor area for England. The Welsh Housing Conditions Survey, 2017 to 2018 (WHCS) also collected information on floor area. Both the EHS and WHCS include two measures of usable floor area: "floorx" and "floory". "Floorx" is defined as the "original EHS definition" of usable floor area, and "floory" is defined as being aligned with the Building Regulations definition, in line with EPC floor area.

We linked a sample of EHS data (2017 to 2019) and WHCS data (2017 to 2018) to the linked VOA and EPC dataset. Both surveys use a statistical model to derive the usable floor area of an address from detailed measurements of the main rooms, meaning a direct comparison with VOA and EPC data is not possible. Correlation analysis showed that both floor area variables from the EHS and the WHCS surveys were more strongly correlated with EPC floor area than VOA floor area, which could be expected given that "floory" is modelled to mimic EPC floor area. The highest correlation between any of the floor area measures was 0.91, supporting the hypothesis that there may be differences in property structure that are too difficult to consistently account for in floor area measurements. This analysis was, however, based on very small sample sizes, and the survey data come from a two-year period, whereas the floor area values on the VOA and EPC data could have been collected earlier, or more recently.

Initial findings suggest that the variance observed in the models is caused by quality of the EPC floor area variable when used for statistical purposes, and differences in property structure that cannot be fully accounted for in the data available.

6 . Future developments

Our [methodology article, Valuation Office Agency property attribute data: quality assurance of administrative data used in Census 2021](#), shows that Valuation Office Agency (VOA) data provide property characteristics information for domestic properties across England and Wales, with a good degree of accuracy for statistical purposes. Without being able to harmonise the two different VOA floor area measures, caution needs to be taken when comparing available living space across different property types. However, floor area, as explained in our [previous methodology, Admin-based statistics for property floor space, feasibility research: England and Wales](#), can be compared for properties of the same type. Where appropriate, the Office for National Statistics (ONS) will make use of VOA floor area to provide information about the housing stock in England and Wales in the future.

We will also explore methods that use floor area to identify overcrowded or under-occupied addresses within the same property type. This would provide additional insight into the diversity of living conditions that are not captured by existing measures of overcrowding. Future research also aims to evaluate alternative ways of harmonising floor area across property types by using different statistical modelling methods such as machine learning, or through the inclusion of additional data sources.

Feedback

We welcome feedback on the method used to harmonise floor area measure and the planned future developments. We are very interested in understanding what uses floor area statistics have that are likely to be of interest in the future to inform policies, target schemes and monitor changes over time. This information will help us to ensure we meet user needs where possible. Please email your feedback to [admin.based. characteristics@ons.gov.uk](mailto:admin.based.characteristics@ons.gov.uk). Please include "Housing" in the subject line of your response.

7 . Glossary

Floor area

Floor area in the Valuation Office Agency (VOA) data is measured using two methods, depending on property type. Reduced Cover Area (RCA) is used for houses and bungalows and Effective Floor Area (EFA) is used for flats and maisonettes.

The [Council Tax Referencing Manual (PDF, 1.90MB)](#) states that RCA includes "all the area covered within the external walls, measured externally". The RCA excludes the following areas:

- eaves overhang

- open balconies

- covered ways and external passages

- unconverted loft areas

- areas with a headroom less than 1.5 metres (other than areas under stairs)

The RCA also excludes the following areas as they are measured separately:

- attached and integral garages

- washhouses, fuel stores and coal bunkers

- conservatories and porches

- any extension of a temporary nature or of significantly inferior quality to the main dwelling

EFA is defined as "the usable area of the rooms within a dwelling measured to the internal face of the walls of those rooms. It will not differentiate between structural and non-structural partitioning of rooms." It excludes:

- hallways, landings and passages (regardless of whether enclosed by structural or non-structural partitions)

- cupboards opening off excluded areas

- columns, piers, chimney breasts, etc

- bathrooms, toilets, and showers

- all areas with a headroom less than 1.5 meters

- areas covered by stud walls and partitions

Floor area in the Energy Performance Certificate (EPC) data are measured using a single method to provide a Total Floor Area (TFA) for all property types. The TFA is defined as "the total of all enclosed spaces measured to the internal face of the external walls."

According to the [Government's Standard Assessment Procedure for Energy Rating of Dwellings (PDF, 2.48 MB)](#) (SAP), "rooms and other spaces, such as built-in cupboards, should be included in the calculation of the floor area where these are directly accessible from the occupied area of the dwelling. However, unheated spaces clearly divided from the dwelling should not be included." Full information about what is included and excluded in the TFA measurement can be found in the SAP. To illustrate some of the differences between the VOA and EPC floor area measurement methods, the TFA includes the following areas (with notes to show exclusions from RCA and EFA):

- porches, if heated (excluded from RCA and EFA)

- area under partition walls (excluded from EFA)

- hallways, if they are not shared (excluded from EFA)

- conservatories, if they are not separated from the main dwelling (excluded from RCA and EFA)

- utility rooms and storerooms, if they are connected to the dwelling (excluded from RCA)

- bathrooms, shower rooms and toilets (excluded from EFA)

- garages, if heating is provided from the main central heating system (excluded from RCA and EFA)

Property type

The Energy Performance Certificate data group properties into five property (or dwelling) types: houses, bungalows, flats, maisonettes and park homes. Park homes were excluded from the linked dataset owing to small counts.

The Valuation Office Agency data group properties into 29 different property types plus a further three categories for "unidentified" houses, flats and bungalows. These were grouped into houses, bungalows, flats, maisonettes and other. Those listed as other were excluded from the linked dataset owing to small counts, along with a lack of comparable category in the EPC data.

Unique property reference number (UPRN)

A unique property reference number (UPRN) is a unique identifier for every address in Great Britain and is allocated by local government and Ordnance Survey (OS).

VOA floor area measure flag

VOA has two distinct ways of measuring floor area depending on the type of property being measured: Reduced Cover Area (RCA) is used for houses and bungalows; and Effective Floor Area (EFA) is used for flats and maisonettes (see "floor area" above for more detail). We refer to "VOA floor area measure flag" where we have grouped the data according to the floor area measurement method used. This splits the data into two groups:

- properties measured using the RCA measure, which includes houses and bungalows

- properties measured using the EFA measure, which includes flats and maisonettes

8 . Data sources and quality

Valuation Office Agency (VOA) property characteristics data

The VOA captures data about properties for Council Tax banding purposes, meaning that VOA data should cover all properties in England and Wales that are liable to pay Council Tax.

The Office for National Statistics (ONS) receives data from VOA on the second Monday of every month, so we used the April 2021 cut to best align with the EPC data. Unique property reference numbers (UPRNs) were mapped to the VOA's unique address reference number (UARN) for each address using the cross-reference table on AddressBase Premium. We removed 0.5% of addresses that had a duplicate UPRN, or where a UPRN could not be assigned.

It is worth noting that VOA data are not regularly updated until a property is sold, meaning that any modifications made to a property that would increase its floor area, such as extensions, may not be reflected in the derived floor area variable. Further information about VOA data and its quality can be found in our methodology article, Valuation Office Agency property attribute data: quality assurance of administrative data used in Census 2021.

Energy Performance Certificate (EPC) data

The EPC data are maintained by the Department of Levelling up, Housing and Communities (DLUHC). An EPC provides a measure of a property's energy efficiency, and since 2007, it has been a legal requirement for any property that is built, sold or rented. Once issued, an EPC is valid for ten years.

We used EPC data from March 2021. Only the most recent record for each property was used. UPRNs were assigned to EPC data using the ONS' Address Index Matching Service. We removed 7.9% of records with a duplicate UPRN.

Linked VOA and EPC data

Both datasets were linked to the national statistics UPRN lookup (NSUL) to obtain additional geography variables. VOA data were then linked to EPC data via UPRN, with 57.2% of VOA addresses (approximately 15.0 million) linking to an EPC address (57.5% for England, and 53.7% for Wales). Only 2.8% (approximately 430,000) of EPC addresses failed to link to a VOA address. A very small number of linked records that could not be linked to the NSUL were removed.

Before conducting the agreement rates and regression analysis we took steps to clean the linked EPC and VOA dataset. The following steps removed 3.7% of the linked dataset.

Firstly, we removed addresses with a missing floor area value on either the EPC or VOA and addresses where the property type on either the EPC or VOA was missing or not listed as a house, bungalow, flat or maisonette. Removing missing values was essential to conduct the regression analysis, and the "other" group was too small for consideration in later models. Cook's distance analysis, before removing any further addresses, produced values up to 46.22.

We removed any addresses with unfeasibly small (less than or equal to 5sqm) floor area values on either the EPC or VOA, along with addresses with especially large floor area values (greater than 500sqm).

Finally, we calculated the difference between the EPC and VOA total floor area values and removed the 1st and 99th percentiles. Post cleaning, all Cook's distance values reduced to less than 0.01 showing that removing these addresses reduced the likelihood of outliers distorting later regression analysis. The final dataset contained 14.5 million linked VOA and EPC addresses.

9 . Related links

[Using structural equation modelling to determine measurement error in different administrative data sources](#)
Article | 26 October 2022
This methodology explores how structural equation modelling can quantify the measurement error in administrative data sources, in this case floor area measured by Valuation Office Agency and Energy Performance Certificate data.

[Admin-based statistics for property floor space, feasibility research: England and Wales](#)
Methodology | Last revised 21 February 2020
Early research demonstrating measures of property size (floor space) from Valuation Office Agency data, which statistics can be generated from these data, what they can and cannot be used for, and a request for feedback on their usefulness.

[Administrative Data Census Research Outputs: Housing Characteristics](#)
Webpage | Last updated 18 January 2021
These outputs are part of our ongoing research into the possibility of combining administrative data and surveys to produce statistics for census-type information about housing characteristics as part of an Administrative Data Census.

[Valuation Office Agency property attribute data: quality assurance of administrative data used in Census 2021](#)
Methodology | Last revised 31 July 2020
Summary of the quality assurance undertaken on administrative data for Valuation Office Agency property attribute data used in Census 2021.

[Data source overviews: Valuation Office Agency data](#)
Webpage | Last revised 27 June 2017
Data source overviews intended to provide a high-level view of new data sources being evaluated by the Census 2021 Transformation Programme.

[Admin-based levels of overcrowding (using the bedroom standard and Valuation Office Agency number of bedrooms), feasibility research: England and Wales: January 2021](#)
Article | 18 January 2021
Exploration of differences in bedroom occupancy ratings derived from Census 2011 and linked Valuation Office Agency number of bedrooms data. The research demonstrates the potential of deriving levels of overcrowding using the bedroom standard from linked survey and admin data.

[Estimating the number of rooms in Census 2021: an update on deriving an occupancy rating from Valuation Office Agency number of rooms](#)
Article | 18 January 2021
Exploration of differences in room occupancy ratings derived from Census 2011 and linked Valuation Office Agency number of rooms data. The research aims to provide further assurance for the replacement of the number of rooms question in Census 2021.

10 . Cite this article

Office for National Statistics (ONS), released 26 October 2022, ONS website, article, [Developing admin-based property floor area statistics for England and Wales: 2021](#)