UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
Conference of European Statisticians
Expert Meeting on Statistical Data Editing
3-7 October 2022, (virtual)

Producing admin-based property floor area statistics for England and Wales: methods, data and quality

Stephan Tietz, Emily Mason-Apps, Shannon Bull, Andreea Butnaru and Joseph Herson (Office for National Statistics, UK)
admin.based.characteristics@ons.gov.uk

I. Introduction

1. 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 space 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).

2. This research is a progression of previous research that looked at the feasibility of using Valuation Office Agency (VOA) data to provide a measure of property size (floor area). The previous research demonstrated that VOA data could provide a suitable measure of floor area to enable comparison of size between properties of the same type. Due to using two different methods to measure the floor area of different property types, comparing the property size of different property types is not possible using VOA data alone. This new research 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 II).

II. Measuring available living space across property types

3. 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.

4. Overcrowding is often measured using occupancy ratings, usually the room (defined in the Housing Act 1985) or bedroom standard (defined in 2012 by DCLG, now 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 past publication on number of rooms (Section 3) and number of bedrooms (Section 2).

A. Valuation Office Agency (VOA) property characteristics data

5. Valuation Office Agency (VOA) data covers 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.

6. 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 to 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 VI).

B. Energy Performance Certificate (EPC) data

7. 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. In contrast to the VOA data, EPC data only uses 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 VI.

C. Producing a single measure of available living space

8. 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.

9. This research explores the feasibility of producing a statistical model (see Section IV) 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.

III. Data used to harmonise VOA floor area measures

10. 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. Here we detail the data cleaning and linkage steps taken to ensure that the linked dataset was representative of all residential addresses across England and Wales.

A. Valuation Office Agency (VOA) property characteristics data

11. 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.

12. ONS receive 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.

13. 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 here.

B. Energy Performance Certificate (EPC) data

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

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

C. Linked VOA and EPC data

16. Both datasets were linked to the national statistics UPRN lookup (NSUL) to obtain additional geography variables. VOA data was 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.

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

18. 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” groups was too small for consideration in later models. Cook’s distance analysis before removing any further addresses produced values up to 46.22.

19. 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).

20. 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.

D. Representativeness and linkage rates of VOA and EPC addresses

21. 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 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 are planning to include a section about the representativeness of EPC data in their next annual Energy Efficiency of Housing publication.

22. 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 due to such properties not requiring 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

23. 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

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

E. Comparing VOA and EPC floor area and property type information

24. We undertook a series of steps to clean the linked dataset (see Section III), including removing linked addresses where the property type was “other” or missing. The cleaned and linked VOA and EPC data was used for all the following analysis.

25. 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 VI).

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

VOA property type	*VOA (m^2)*	*EPC (m^2)*	*Difference (VOA – EPC) (m^2)*
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

Note: Addresses where the VOA property type category was missing or “other” were removed from the linked and cleaned VOA and EPC dataset (see Section III).

26. 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.

27. 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 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".

28. 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.

IV. Method for harmonising floor area measures

29. We used regression modelling to test if we could use the two different floor area measures in Valuation Office Agency (VOA, see Section II) 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.

A. Exploring VOA variables that can predict EPC floor area

30. We explored a simple linear regression model using VOA floor area as the primary predictor for EPC floor area. This model produced an R^2 of 0.84, providing a benchmark for further analysis (see Table 5).

31. An important aim for the final model was to enable comparison of floor area across different property types (see Section II), and between England and Wales (see Section III). 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.

32. The R^2 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^2 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".

33. 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^2.

34. The results by country (England and Wales) show no change for England, and a slightly lower R^2 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^2, so only country was used to maintain model parsimony.

Table 5 Parameters of the simple linear regression models with VOA floor 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^2*
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

B. Best performing multiple linear regression model

35. 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 to the original simple linear regression model, this model produced an improved adjusted R^2 of 0.86. To evaluate the estimator performance of the model we performed k-fold cross-validation using k = 10. The R^2 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 (RCA or EFA) and country to predict EPC floor area

RMSE	*R^2*	*Adjusted R^2*		*Coefficients*		*Intercept*
14.61	0.86	0.86	VOA floor area	VOA floor area measure flag	Country	14.02
			0.93	-17.91	1.85	

Note: Reference category for VOA floor area measure was EFA (flats and maisonettes), and for country it was Wales.

36. 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 0sqm, 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, 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.

37. We therefore applied a log-transformation to the floor area variables prior to running our best performing multple 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^2 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.

C. Performance of final multiple linear regression model

38. 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 -7.16sqm (for England) to -5.51sqm (for

Wales), with medians of -7.55sqm for Wales and -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.

39. 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 it is small properties which 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.

D. Sources of variance in the floor area data

40. 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.

41. Research conducted by ONS' Methodological Research Hub using structural equation modelling to estimate the measurement error of VOA and EPC data estimates that EPC data tends to have a larger measurement error for the floor area variable (6.5%) compared to 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.

42. 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.

43. 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 useable 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 comes from a two-year period, whereas the floor area values on the VOA and EPC data could have been collected earlier, or more recently.

44. 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.

V. Future developments

45. Our previous research shows that Valuation Office Agency (VOA) data provides 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

can be compared for properties of the same type. Where appropriate, ONS will make use of VOA floor area to provide information about the housing stock in England and Wales in the future.

46. 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.

VI. Appendix: VOA and EPC floor area measures

47. 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. We refer to “VOA floor area measure flag” where we have grouped the data according to the floor area measurement method used.

48. The Council Tax Referencing Manual states that RCA includes “all the area covered within the external walls, measured externally”. The RCA *excludes* the following areas (those asterisked are measured separately but not included in the floor area measurement):

(a) eaves overhang
(b) open balconies
(c) covered ways and external passages
(d) unconverted loft areas
(e) attached and integral garages*
(f) washhouses* and fuel stores*/coal bunkers
(g) conservatories* and porches*
(h) any extension of a temporary nature or of significantly inferior quality to the main dwelling*

49. EFA is defined as “the useable 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:

(a) hallways, landings and passages (regardless of whether enclosed by structural or non-structural partitions)
(b) cupboards opening off excluded areas
(c) columns, piers, chimney breasts etc.
(d) bathrooms, toilets, and showers
(e) all areas with a headroom less than 1.5 meters
(f) areas covered by stud walls and partitions

50. Floor area in the Energy Performance Certificate (EPC) data is 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”

51. According to the Government’s Standard Assessment Procedure for Energy Rating of Dwellings (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 floor area measurement methods, the TFA *includes* the following areas (those marked with 1 are excluded from RCA, those marked with a 2 are excluded from EFA):

(a) Porches (if heated)^{1,2}
(b) Area under partition walls^{2}
(c) Hallways (if they are not shared)^{2}
(d) Conservatories (if they are not separated from the main dwelling)^{1,2}
(e) Utility rooms and storerooms (if they are connected to the dwelling)^{1}
(f) Bathrooms, shower rooms and toilets^{2}
(g) Garages (if heating is provided from the main central heating system)^{1,2}