Methodological Assurance Review Panel (MARP): Producing admin-based property floor area statistics for England and Wales: methods, data and quality

Version Number: *2.0*

Date: 19 August 2022

Authors: Shannon Bull, Andreea Butnaru, Joe Herson, Emily Mason-Apps, Stephan Tietz

Revision History

Version	Date	Owner	Summary of changes
1.0	15/02/2022	Stephan Tietz	Final version for MARP panel
2.0	19/08/2022	Stephan Tietz	Final version for publication

Note

Version 2.0 of this paper was created to incorporate our responses to feedback we received from MARP panel members before our presentation on 1st March 2022 and which we addressed in that presentation. This paper therefore represents an accurate account of the work on this date. Further work is underway and will be published on the ONS webpage in due course.

Changes to version 1.0 include:

- Rephrasing of the conclusion: We removed the phrase that our final "model would be best suited to predict EPC floor area at a household level across both VOA floor area measures as well as England and Wales." And added more clarity around the observed variance, data quality and how this impacts the suitability of this model to produce estimates of high statistical quality.
- Addition of Section 10 which summarises the results of our final model using log-transformed floor area variables.
- In order to rerun the final model using log-transformed floor area variables, the underlying analytical dataset was created using an improved analytical pipeline. For comparability, all of the analysis presented in the paper has been rerun and none of the original findings have changed.

1 Purpose

Further research demonstrating the potential of Valuation Office Agency (VOA) and Energy Performance Certificate (EPC) data to provide detailed information on property size (floor area). A multiple regression model is used to predict EPC floor area from VOA data. The aim is to harmonise the two different VOA floor area measures across property types.

2 Recommendations

Members of the group are invited to:

I. Provide feedback on the analysis underpinning the selection of the final regression model that predicts Energy Performance Certificate (EPC) floor area from Valuation Office Agency (VOA) data.
II. Suggest ways that could further improve the prediction of EPC floor area from VOA data.

3 Background

3.1 Measuring levels of overcrowding through bedroom standard and room occupancy rating

Housing policy is concerned with the availability and quality of housing. An important aspect 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 "crammed" or overcrowded (PDF, 681KB).

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). An occupancy rating is obtained by subtracting a hypothetical number of rooms (or bedrooms) recommended for a household from the actual number of rooms (or bedrooms) available. A household is considered overcrowded if it has fewer rooms (or bedrooms) available than recommended (negative occupancy rating), or under-occupied if it has more (positive occupancy rating). This makes the occupancy rating a straightforward way of measuring overcrowding and under-occupancy.

The current measures of overcrowding have some clear limitations. For example, two properties with the same number of residents and rooms (or bedrooms) could both be reported as overcrowded using the room occupancy rating (or bedroom standard). However, one of these properties could have over double the floor area compared to the other.

To our knowledge, there is no internationally harmonised definition of rooms or bedrooms, or methodology for how to count them, making international comparisons of living conditions challenging. Measuring available living space using floor area could be one way to better reflect the diversity of living conditions and offer an opportunity for international comparison.

The last decade has seen a move to more open plan living, which means that the lack of a separate kitchen is no longer a strong indicator of deprivation. More recently, Covid-19 has forced many people to "convert" parts of their houses into home offices, or classrooms for their children. This flexible use of living space is not reflected in the bedroom standard and is difficult to incorporate into the room occupancy rating as everyone's circumstances are slightly different. Again, measuring the available living space using floor area may therefore provide a more suitable measure to reflect living conditions.

3.2 Measuring overcrowding using floor area

Until now, published information about floor area has been produced through surveys such as the English Housing Survey. However, because of sample size, analysis of floor space for sub-regional geographies has been limited. This research explores the feasibility of using two administrative data sources, Valuation Office Agency data and Energy Performance Certificate data, to produce a harmonised floor area measure that is also capable of providing estimations at a more granular level across England and Wales.

3.3 Different floor area measures in administrative data

3.3.1 Valuation Office Agency (VOA) property attribute data

Valuation Office Agency (VOA) covers most residential properties within England & Wales and measures the floor area of them. Research on using VOA floor area has previously been published. The analysis found that floor area generally follows the expected distribution by property type, for example, detached properties are larger than those belonging to a terrace. However, the VOA have two distinct ways of measuring floor space depending on the type of property being measured. Reduced cover area (RCA) is used for houses and bungalows, whilst effective floor area (EFA) is used for flats and maisonettes. To enable comparisons across these two groups it is desirable to harmonise the two floor area measures.

The RCA measures the floor area of the whole dwelling (including external walls) and then reduces it by subtracting a range of areas that fall outside of the measure. The EFA on the other hand measures the useable area of the rooms within a dwelling measured to the internal face of the walls of those rooms. A breakdown of what is included in each measure can be found in Section 12 of the previous publication. Due to the differing measures of floor area, it is therefore not possible to effectively compare the floor area of the different property types using VOA data alone. Because the RCA includes external walls, and areas such as hallways, landings and passages in the measurements, we would expect this method to typically overestimate the “available living space” (see Section 3.3.3) compared to the EFA method.

It is worth also noting that VOA data are not updated until a property is sold, which could be problematic for properties that have had extensions or moderate modifications to increase floor area. This would not be recorded in the VOA data and may present a delay in updating the derived floor area variable. Further information about VOA data and its quality can be found here.

3.3.2 Energy Performance Certificate (EPC) data

An Energy Performance Certificate (EPC) provides a measure of the energy efficiency of a property. EPCs were introduce in 2007 and are a legal requirement for any property that is built, sold or rented. Once issued, an EPC is valid for ten years. The database of issued EPCs is maintained by the Department for Levelling Up, Housing and Communities (DLUHC, previously MHCLG). A list of buildings that do not require an EPC can be found here.

In contrast to the VOA data, EPC data only uses one measure of floor area for all property types: Total Floor Area (TFA). TFA is defined as, “the total of all enclosed spaces measured to the internal face of the external walls”. TFA only includes areas that are heated, habitable and internally accessible from the main dwelling, this makes the measure closely represent “available living space” (see Section 3.3.3).

There are some limitations to EPC data:

Representativeness: A recent ONS publication found that currently roughly half of all dwellings in England and Wales have an EPC. Because an EPC is only required for buildings that are sold, rented or newly constructed, it is possible that certain property types are underrepresented.

Timeliness: Because an EPC is valid for 10 years, it is possible that any changes to a property will not be captured within the data in a timely manner (for example, if a property was extended but never sold). This effect is likely to vary by property type and tenure.

Time series and quality: As noted in the EPC Statistical Release, changes in the EPC Scheme Operating Requirements came into effect in 2012. This was noted to result in an improvement to the quality of data and a recommendation that users consider this if they wish to use data collected before that date.

3.3.3 Available living space

To further our understanding of overcrowding in residential properties, the concept of "available living space" would benefit policy makers in understanding the amount of space people have to live in. The Total Floor Area (TFA) measured by EPC would be a suitable proxy of this concept and would provide a greater understanding as a consistent measure across all. The TFA is much more aligned to "available living space" than the VOA measures of EFA and RCA. Considering what the different measures include when calculating the floor area, 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.

A model that takes the completeness of VOA's measures and accurately predicts the EPC measure of TFA should be able to provide a high-quality measure of "available living space" for residential properties in England and Wales. To produce this harmonised measure of floor area, we will use VOA floor area alongside other VOA property characteristics to predict the EPC floor area. We will use linked VOA and EPC addresses (see Section 8) to create a statistical model which can then be used to impute the harmonised floor area for VOA addresses that do not have an EPC (i.e. they cannot be linked). This aligned measure could then be used at address level which would allow overcrowding measures such as people per floor area to be produced.

4 Statistical quality and coverage of VOA data linked to EPC data

4.1 Representativeness and quality of VOA data linked to EPC data

We linked VOA and EPC addresses (see Section 8). As shown in Table 1, 57.5% of VOA addresses in England were linked to an EPC address, and 53.7% in Wales. Similar rates were found in the energy efficiency of housing in England and Wales publications. Only 2.8% of cleaned EPC addresses failed to link to a VOA record. Records referred to as "Other" in Table 1 reflect those that could not be linked to the National Statistics UPRN Lookup (NSUL) to obtain additional geographical information. These records were removed at this point and are excluded from further analysis.

Table 1 Number of VOA and EPC addresses and linkage rate by Country

Country (assigned from the NSUL)	VOA addresses (n)	EPC addresses (n)	VOA addresses linked to EPC (%)
England	24,769,401	14,633,812	57.5
Wales	1,444,306	799,078	53.7
Other	1,075	2,146	29.7
Overall	**26,214,782**	**15,435,036**	**57.2**

Note: the "Other" category refers to records that were not able to be linked to the NSUL. For the EPC data, the "Other" category also includes records from Scotland.

To further understand the linkage and representativeness of addresses we looked at linkage rates of VOA data by property types in England and Wales. Table 2 shows that maisonettes and flats have the highest rates of linked addresses for both England and Wales, although Wales has slightly fewer than England. The lowest rate of linkage for each country is for "other" properties (such as caravans), this is likely to be 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). Overall, the coverage for the property types in England and Wales is good and will increase overtime.

Table 2 Linkage rates of cleaned VOA and EPC addressess 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

To check for possible biases in the data, we looked at the distributions of linked versus unlinked addresses by property type and country. As shown in Table 3, 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. For country, there is only a minimal difference in the percentages of linked and unlinked addresses.

Table 3 Distribution of linked and unlinked addresses by property type and country

	VOA addresses linked to EPC (%)	Unlinked VOA addresses (%)
VOA property type	**100.0**	**100.0**
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
Country	**100.0**	**100.0**
England	94.8	94.0
Wales	5.2	6.0
Total count	**15,004,525**	**11,209,182**

ONS's sub-national housing analysis team is collaborating with DLUHC to further understand the representativeness of EPC data. Meanwhile, we think that the VOA addresses linked to EPC data are reasonably representative of all residential addresses and can be used to produce a statistical model for harmonising floor area across property types.

4.2 VOA and EPC property type agreement rates

Before conducting any further analysis, we undertook a series of steps to clean the data which are detailed in Section 8. To assess the data quality and consistency, we also looked at agreement rates on property types between VOA and EPC. As can be seen in Table 4 , the agreement rate for most property types is high (over 90%), however maisonettes have a low agreement rate of 55.2%. This is likely to be due to the fact that guidance for EPC surveyors (The Government's Standard Assessment Procedure for Energy Rating of Dwellings, which includes the Reduced Data SAP for existing dwellings, RdSAP) states that "RdSAP makes no distinction between a flat and a maisonette as regards calculations; it is acceptable to select either type as definitions vary across the UK".

Table 4 Agreement rates between VOA and EPC data by VOA property type

VOA property type	Agreement rate (%)
House	98.6
Bungalow	93.2
Maisonette	55.2
Flat	94.3

When property type is grouped by the VOA floor measure used (RCA or EFA, see Section 3.3.1) for both VOA and EPC, the agreement rates are 99.7% for RCA (houses and bungalows) and 98.4% for EFA (flats and maisonettes), as displayed in Table 5. This improved agreement suggests that statistical models may benefit from grouping VOA property type according to the VOA floor area measure.

Table 5 Agreement rates between VOA and EPC data by VOA floor area measure

VOA floor area measure	Agreement rate (%)
RCA – houses and bungalows	99.7
EFA – flats and maisonettes	98.4

4.3 Comparing VOA floor area with EPC floor area

We looked at median floor values in meters squared for England and Wales to understand difference between the VOA and EPC floor area across VOA property types. Table 6 shows that for England and Wales houses have the highest floor area across both VOA and EPC. For both houses and bungalows the median VOA floor area was greater than EPC floor area measurements. Maisonettes and flats had the opposite pattern, where EPC showed greater floor area compared to VOA. This is in line with the definitions discussed in Section 3.3.

Separately, both countries follow the overall trend found for England and Wales combined. However, within country comparisons show that Wales has a larger median floor area for houses, bungalows, and maisonettes for both VOA and EPC measures. This suggests that our statistical model should include a geographical dimension.

Table 6 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

5 Regression modelling to predict EPC floor area

To enable us to develop and evaluate the best possible model for predicting EPC floor area, we first agreed a number of conditions that the model would have to satisfy. These were to:

- Produce the best predicting model at address level whilst minimizing prediction errors
- Enable comparison across property types (in particular, across the VOA floor area measure)
- Enable comparisons across geographies (in particular, across England and Wales)
- Maintain model parsimony

5.1 Exploring VOA variables that can predict EPC floor area

Here we summarise some of the analysis we conducted to narrow down the selection of VOA variables for our final multiple linear regression model. To begin, 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, demonstrating that VOA floor area alone accounts for a large proportion of the variance in EPC floor area.

We then conducted a series of simple linear regression models to assess if this relationship held when the data was split by VOA property type and geographic variables (Country, Government Office Region, and Rural Urban Classification). The R^2 for each subgroup when the data is split by property type (house, bungalow, maisonette and flat) varied between 0.54 to 0.85 indicating prediction is not equally successful across different property types. The results by country (England and Wales) show no change for England, and a slightly lower R^2 for Wales. Detailed results for the simple regression models can be found in Table 7.

Table 7 Parameters of simple linear regression models

VOA floor area	Coefficients	Intercept	R^2
	0.82	12.67	0.84
VOA property type			
House	0.93	-2.07	0.85
Bungalow	0.94	-0.18	0.81
Maisonette	0.90	24.46	0.54
Flat	1.05	10.16	0.67
Country			
England	0.82	12.76	0.84
Wales	0.82	10.17	0.82

5.2 Multiple linear regressions to predict EPC floor area using VOA data

5.2.1 Building multiple linear regression models

The next step involved building and evaluating multiple regression models to assess if including other predictor variables would improve the predictive power of the model whilst satisfying the conditions set out in Section 5. The overall R^2 metric of 0.84 obtained from the simple linear regression was used as a benchmark for future model comparisons.

Other than floor area, property type and county, several other property characteristic variables from the VOA dataset were considered for inclusion into the multiple regression models, these included: VOA floor area measure, number of rooms, number of bedrooms, number of bathrooms. These were deemed of importance as they would naturally have some bearing on the size of the property. To assess suitability for inclusion, a correlation matrix was conducted, and it was found that VOA number of rooms and bedrooms were too strongly correlated to VOA floor area and could therefore not be included in any models. Other variables were also explored as predictive variables in some models including number of bathrooms, Government Office Regions, and Rural Urban Classification. These models revealed no notable improvement or reduction in R^2, so in light of the criteria set out at the start of section 5, we decided to exclude these variables from consideration in later models.

Three models resulted in an improved R^2 of between 0.86 to 0.87 (see Table 8). The following sections will present the final model we have selected (MLR2), along with a detailed summary of the steps we took to deem this the best model to take forward.

Table 8 Displays the independent and dependant variables for each MLR Model

Model	Independent variables	Dependent variable
MLR1	VOA floor area, VOA floor area measure	EPC floor area
MLR2	VOA floor area, VOA floor area measure, country	EPC floor area
MLR3	VOA floor area, VOA property type, country	EPC floor area

5.2.2 Evaluation of multiple linear regression models

All models were shown to meet the assumptions for multiple linear regressions (linearity, multicollinearity, homoscedasticity and multivariate normality). Detailed results for the selected model can be found in Section 5)c). As all of the final candidate models satisfied the criteria of maintaining model parsimony, produced similar R^2, and had only marginally different k-fold cross-validation results, we focused on satisfying the remaining conditions:

- enable comparison between property types; and
- enable comparison between England and Wales.

This means that the model chosen would have to be one that minimises the under- and over-estimates across these boundaries. In order to determine this we examined the descriptive statistics of the residuals (mean, standard deviation and median), as well as the distributions.

5.2.3 Final linear regression model: VOA Floor Area, VOA Floor Area Measure and Country

Table 9 presents the results of the final multiple linear regression model we have selected as the best model to take forward. The model uses VOA floor area, VOA floor area measure (RCA or EFA) and country (England and Wales) as predictor variables and EPC floor area as the outcome variable. Compared to the original simple linear regression model, this model produced an improved (adjusted) R^2 of 0.86

Table 9 Results of multiple regression analysis using VOA floor area, VOA floor area measure (RCA or EFA) and country to predict EPC floor area

Reference category for property type was the flats and maisonette group, and for country Wales.

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

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.

5.2.4 Residuals for the final linear regression model

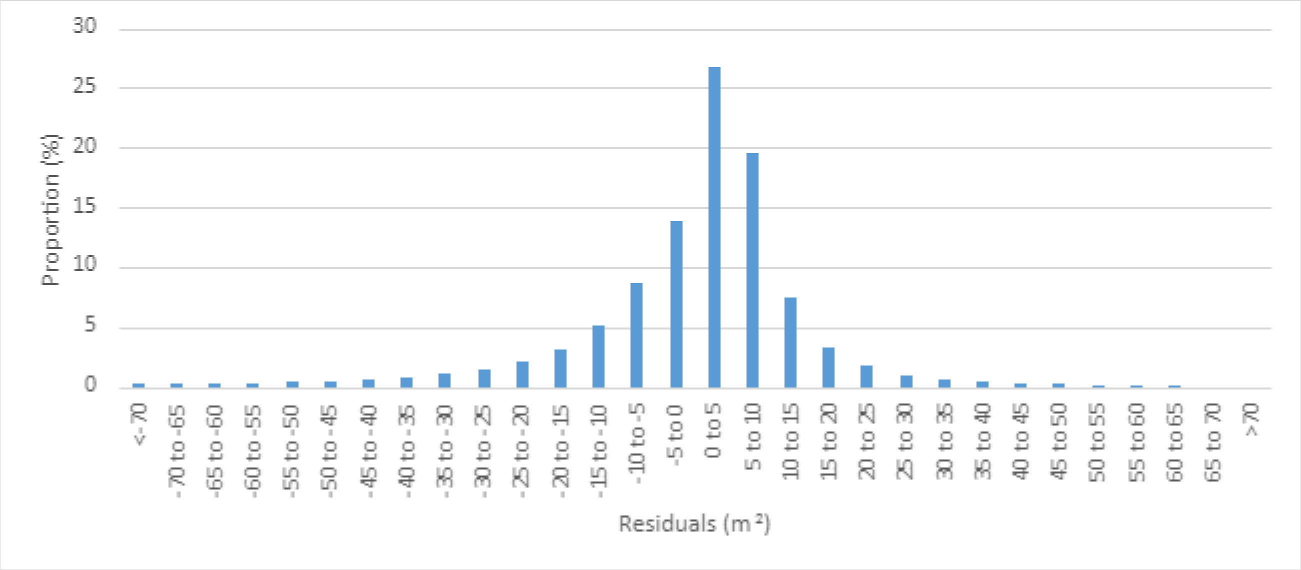
Residuals were calculated by taking the actual EPC floor area measurements (m^2) and subtracting them from the predicted EPC floor area measurements (m^2). A positive number indicates that the model is overpredicting the floor area, while a negative number indicates the model is underpredicting the floor area. The mean, standard deviation (SD) and median of residuals by property type, country and VOA floor area measure can be seen in Table 10. The means and medians being close to 0 suggests that the model predicts well for most property types, enabling comparisons across houses, flats and bungalows. However, comparisons to maisonettes are more challenging due to the larger differences in model predictions for this property group. The small overall differences in predictions between England and Wales also suggests that across country comparisons are possible. It should be noted however that the standard deviation for the mean is large in all groups, indicating a degree of variation within groups.

Table 10 Mean (SD) and Median of residuals (EPC prediction – EPC actual floor area)

VOA property type	Country	Mean (SD)	Median
House	**England**	0.33 (15.52)	3.41
	Wales	0.92 (16.09)	2.78
	England & Wales	0.36 (15.56)	3.38
Bungalow	**England**	-2.53 (15.47)	1.10
	Wales	-4.29 (17.06)	-0.60
	England & Wales	-2.64 (15.58)	1.03
Flat	**England**	0.54 (11.00)	1.12
	Wales	-0.99 (12.02)	-1.09
	England & Wales	0.49 (11.04)	1.04
Maisonette	**England**	-7.16 (16.52)	-8.21
	Wales	-5.51 (20.16)	-7.55
	England & Wales	-7.12 (16.6)	-8.21
House/Bungalow	**England**	-0.02 (15.54)	3.03
	Wales	0.25 (16.32)	2.26
	England & Wales	0.00 (15.59)	2.99
Flat/Maisonette	**England**	0.04 (11.59)	0.69
	Wales	-1.17 (12.49)	-1.22
	England & Wales	0.00 (11.63)	0.62
Overall	**England**	0.00 (14.54)	2.46
	Wales	0.00 (15.73)	1.70
	England & Wales	0.00 (14.61)	2.41

To further evaluate how the model was operating we also explored the distributions of the residuals. As can be seen in Figure 1, the distribution of residuals for the chosen model is reasonably evenly spread around 0m^2, but with a small left skew indicating that the model has a tendency to slightly overestimates floor area. The model predicts 40.6% of addresses within 5m^2, 68.7% within 10m^2 and 87.8% within 20m^2.

Figure 1 Residuals of predicted EPC total floor area minus actual EPC total floor area in 5m^2 banding for all properties.



To determine which model best satisfied our specified conditions, we explored and compared the distribution of residuals for each model across property types and country. On balance, the chosen model presented the best compromise in terms of enabling comparison both across property types and across England and Wales. Simplifying the model by removing country as a predictor (MLR1) variable led to greater overestimation of floor area for Wales. Including VOA property type as a predictor instead of VOA floor area measure (MLR3) did improve the floor area predictions for maisonettes, but led to overpredictions for both houses and bungalows.

6 Conclusion

In order to assess the feasibility of producing harmonised floor area statistics, VOA data was linked to EPC data. Initial checks and descriptive statistics highlight good linkage rates for England and Wales with good agreement on key variables between the two datasets. Initial simple linear regressions were run to select our predictor variables, followed by exploration of several multiple linear regression models. Of the several models assessed, a model was chosen that best satisfied our goals.

The chosen model has VOA floor area, VOA floor area measure and country as predictor variables and has an R^2 value of 0.86. Further analysis of distributions of residuals and the mean/median were undertaken to assess the models performance to predict EPC floor area at a household level across both VOA floor area measures as well as England and Wales.

We are inviting the board to provide feedback on the analysis underpinning the selection of the final regression model that predicts EPC floor area from VOA data and suggest ways the model could be further improved. We conclude that at this point the model does not produce address-level floor area estimates of high enough statistical quality for further analysis (e.g. overcrowding).

7 Next steps

Further work will explore if the observed variance of the modelled estimates is driven by data quality or differences in property structure that cannot be predicted using the data.

8 Annex: Method for linking VOA data to EPC data

The datasets used for the analysis were:

- VOA: April 2021 dataset
- EPC: March 2021 dataset
- AddressBase cross-reference table: EPOCH 84 - cut off 12th April 2021
- National Statistics UPRN lookup: April 2021 dataset

To link the datasets, the following method was used:

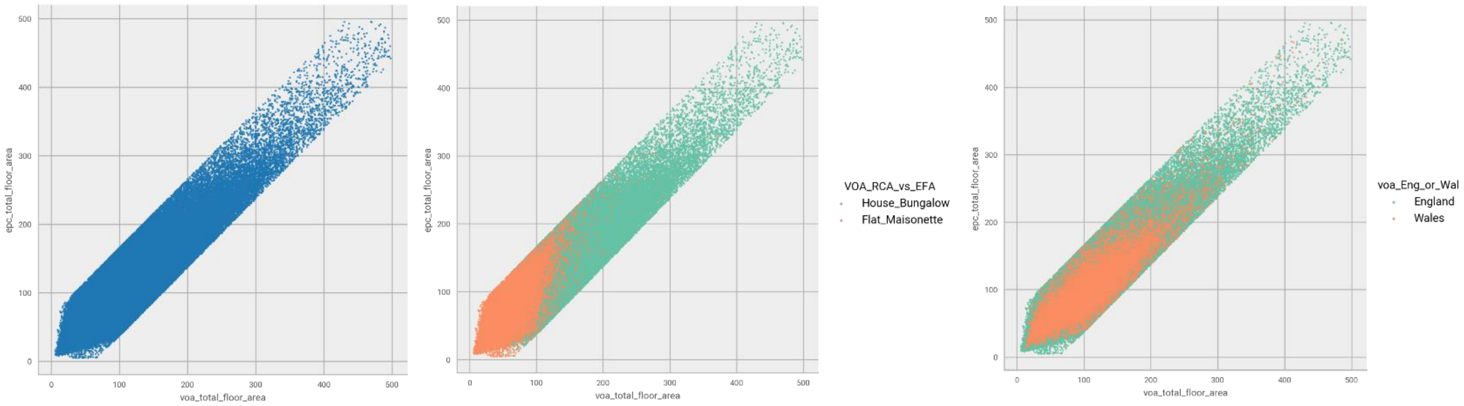
1) EPC data was prepared as follows:
   a) Primary deduplication: Records were ordered using building reference number and lodgement date, to take only the most recent record for each property.
   b) Secondary deduplication: Records were deduplicated by ONS Unique Property Reference Number (UPRN) to keep only those records with a unique UPRN. UPRN is assigned to EPC data using ONS's Address Index Matching Service (AIMS).
   c) EPC data was then linked to the National Statistics UPRN Lookup (NSUL) via UPRN to obtain additional geographical variables.
2) VOA data was prepared as follows:
   a) Primary deduplication: Records were ordered using Unique Address Reference Number (UARN) and lodgement date, to take only the most recent record for each property.
   b) VOA data was linked to an AddressBase cross-reference table, mapping the VOA's UARN to a corresponding UPRN.
   c) Secondary deduplication: Records were deduplicated by UPRN to keep only those records with a unique UPRN from AddressBase.
   d) VOA data was then linked to the NSUL via UPRN to obtain additional geographical variables.
3) EPC data was then linked to VOA data using UPRN.
4) Records that could not be linked to the NSUL were removed from the linked EPC-VOA dataset.
5) 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 3.65% of the linked dataset.
   a) Firstly, we removed cases where the property type on either the EPC or VOA was missing or not listed as either a house, bungalow, flat or maisonette. We also removed any cases with a missing floor area value on either the EPC or VOA. Removing missing values was essential to enable us to conduct the later regression analysis, and the very small size of the "other" groups were deemed too small for consideration in later models. Cook's distance analysis before removing any further cases revealed values up to 46.22.
   b) We then removed any cases with unfeasible (<=5m^2) floor area values on either the EPC or VOA, along with properties with especially large floor area values (>500m^2).
   c) 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 cases reduces the likelihood of outliers distorting later regression analysis.

9 Annex: Testing assumptions for multiple linear regression

This annex shows results for testing the assumptions for multiple linear regressions (linearity, multicollinearity, homoscedasticity and multivariate normality) for the final multiple linear regression model (MLR2, see Section 5.2).

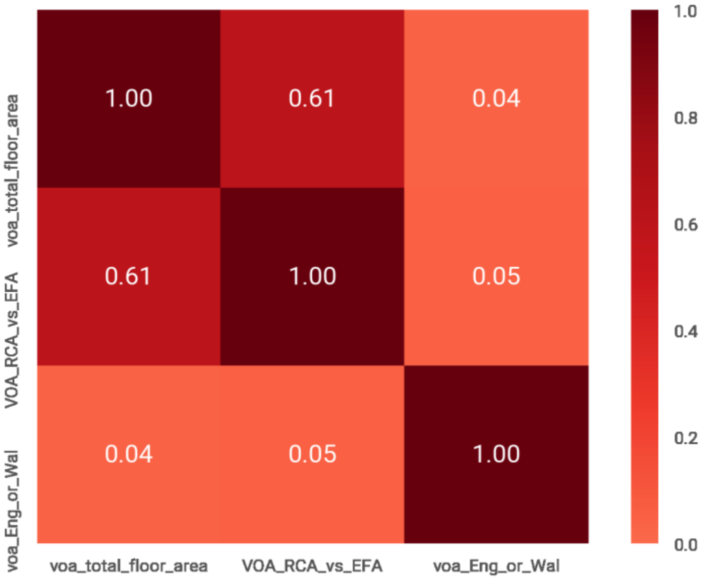
Figure 2 shows a linear relationship between predictor variables (VOA floor area, VOA floor area measure and country) and the outcome variable (EPC floor area), in line with the multiple linear regression assumption of linearity.

Figure 2 Relationship between the VOA floor area and EPC floor area (left), by VOA floor area measure (middle) and by country (right) to check the assumption of linearity.



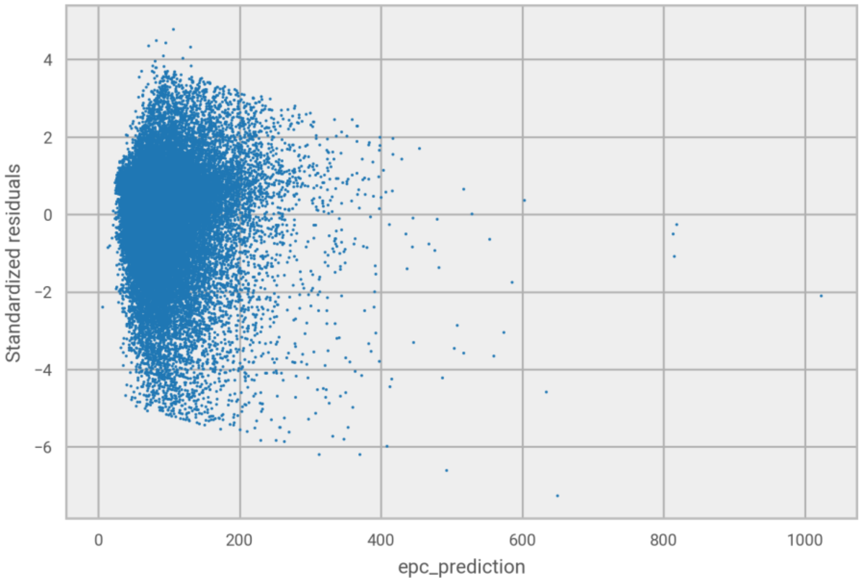
The assumption of multicollinearity was satisfied as shown by the little or no association between the three predictor variables in Figure 3.

Figure 3 Heatmap with associations between the three predictor variables, with 0.0 meaning no association and 1.0 perfect association



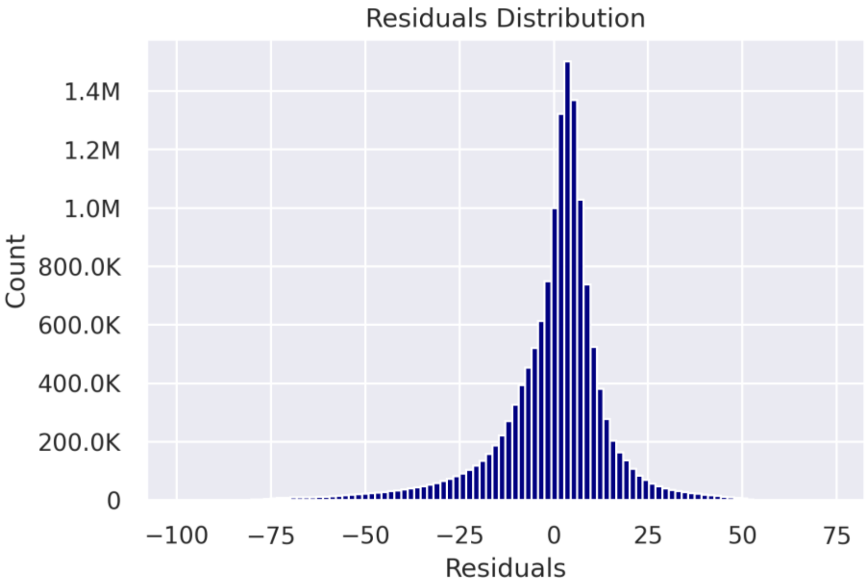
Further we checked homoscedasticity which assumes equal variance within error terms across all values of the independent variables. Figure 4 displays a plot of the standardized residuals versus predicted values which shows that there is some skew which means that larger properties might have an undue effect on the analysis. Following advice from the MARP panel, we rerun the final model after the EPC and VOA floor area variables were log-transformed. The results can be found in Section 10.

Figure 4 Standardised residuals and predicted EPC floor area relationship



Finally, we checked the multivariate normality which assumes residuals are normally distributed. A histogram of residuals can be found in Figure 5, which shows a normal distribution of residuals.

Figure 5 Distribution of residuals



10 Annex: Final linear regression model using log-transformed EPC and VOA floor area variables

Following advice from the MARP panel, this annex shows results for our final model after the EPC and VOA floor area variables were log-transformed.

Table 11 Results of multiple regression analysis using log-transformed VOA floor area, VOA floor area measure (RCA or EFA) and country to predict log-tranformed EPC floor area
*Reference category for property type was the flats and maisonette group, and for country Wales.*Table 11 presents the results of the final multiple linear regression model using log-transformed EPC and VOA floor area variables. The model uses VOA floor area (log-transformed), VOA floor area measure (RCA or EFA) and country (England and Wales) as predictor variables and EPC floor area (log-transformed) as the outcome variable. This model produced a slightly lower (adjusted) R^2 of 0.84 compared to not using log-transformed EPC and VOA floor area variables (R^2 = 0.86).

Table 11 Results of multiple regression analysis using log-transformed VOA floor area, VOA floor area measure (RCA or EFA) and country to predict log-tranformed EPC floor area
Reference category for property type was the flats and maisonette group, and for country Wales.

RMSE	R^2	Adjusted R^2	Coefficients			Intercept
0.07	0.84	0.84	VOA floor area (log transformed)	VOA floor area measure	Country	0.18
			0.95	-0.13	0.01	

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 0.83, which is consistent with the original model.

We checked the assumptions for multiple linear regressions (linearity, multicollinearity, homoscedasticity and multivariate normality) for the final multiple linear regression model using log-transformed EPC and VOA floor area variables.

As shown in Figure 6, the log transformation improved the linearity for large properties, with larger variance being seen for smaller properties. Figure 7 show that association between the predictor variables of VOA floor area and VOA floor area definition has increased.

Figure 8 displays a plot of the standardized residuals versus predicted values which shows that there is a reduction in the skew for larger properties and large variance for smaller properties. Figure 9 shows that the residuals are still normally distributed.

Figure 6 Relationship between the VOA floor area and EPC floor area (left), by VOA floor area measure (middle) and by country (right) to check the assumptions of linearity using log-transformed EPC and VOA floor area variables

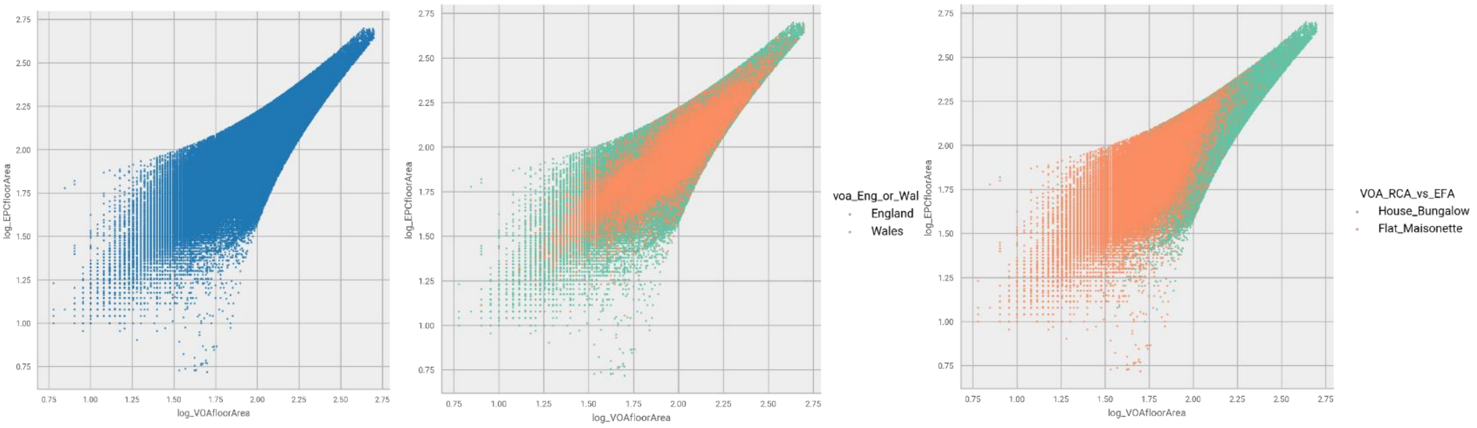


Figure 7 Heatmap with associations between the three log transformed predictor variables, with 0.0 meaning no association and 1.0 perfect association.

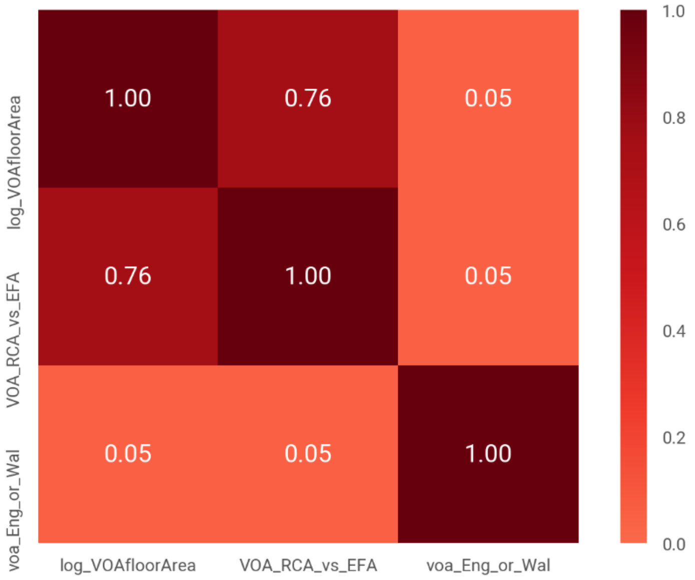


Figure 8 Standardised residuals and predicted EPC floor area relationship using log-transformed EPC and VOA floor area variables

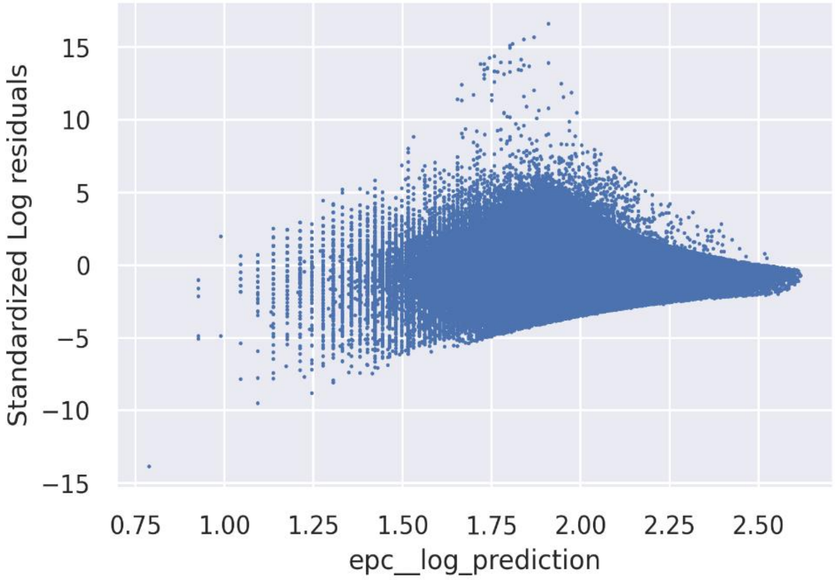
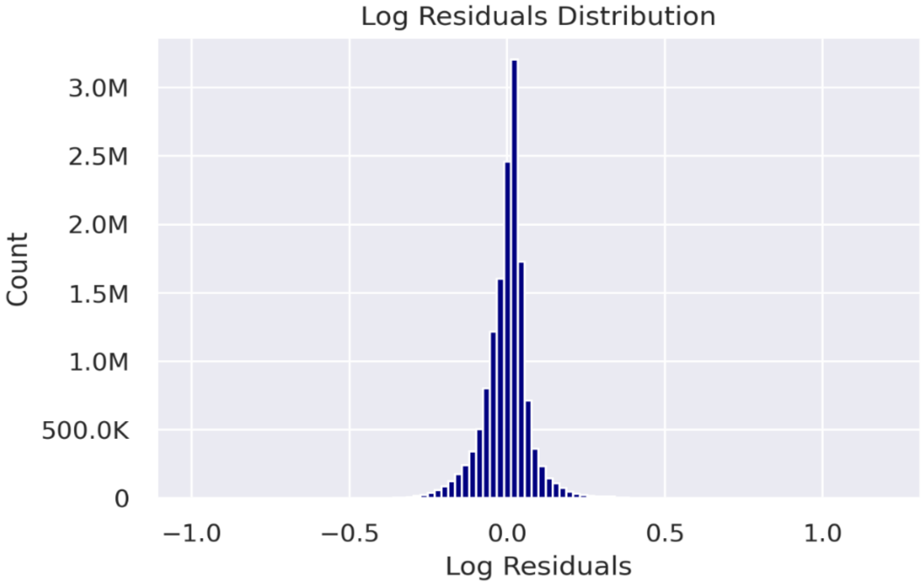


Figure 9 Distributions of residuals using log-transformed EPC and VOA floor area variables


Log Residuals Distribution
Count
3.0M
2.5M
2.0M
1.5M
1.0M
500.0K
Log Residuals