Imputing linked 2011 Census and administrative
Valuation Office Agency (VOA) data: A feasibility study

# Introduction

The 2021 Census white paper, ‘[Help Shape Our Future](https://www.gov.uk/government/publications/the-2021-census-of-population-and-housing-in-england-and-wales)’ (1), outlined the Office for National Statistics’s intention to integrate other data sources into the 2021 Census to enhance outputs. [Previous research](https://www.ons.gov.uk/census/censustransformationprogramme/questiondevelopment/housingcommunalestablishmentsandvisitors/estimatingthenumberofroomsandbedroomsinthe2021censusanalternativeapproachusingvaluationofficeagencydata) (2) on administrative data sourced from the Valuation Office Agency (VOA) identified that they are a primary data source to be incorporated into the design and development of the 2021 Census statistical processing strategy.

This report details the findings of feasibility research into whether administrative data variables from the VOA are suitable to undergo editing and imputation (E&I) within the standard census framework when linked to 2011 Census questionnaire data. The focus of this research was on E&I of the number of rooms variable present in the VOA data. This follows the proposal in the [white paper](https://www.gov.uk/government/publications/the-2021-census-of-population-and-housing-in-england-and-wales) (1) that the number of rooms question would be removed from the 2021 Census due to quality issues, as shown by the 2011 [Census Quality Survey](https://www.ons.gov.uk/file?uri=/census/censustransformationprogramme/consultations/the2021censusinitialviewoncontentforenglandandwales/2011censusqualitysurveyreport.pdf) (3), and the availability of equivalent variables from administrative data. We considered this as a good starting point to examine the potential challenges linked administrative data may make to the assumptions underpinning the E&I process, such as concerns which may arise from linking datasets, missing data mechanisms, and definitional changes and time-frame issues.

Following linkage of the datasets, we considered the VOA rooms variable to potentially require imputation in the following scenarios:

1. Where there is missingness due to failure to link a 2011 Census record to a VOA record
2. Where there is missingness present in the VOA data prior to linkage
3. Where VOA data are inconsistent with observed 2011 Census data

This report examines how linked administrative and survey data challenge assumptions underlying the E&I process and we attempt to impute VOA data using 2011 Census data to predict the missing and inconsistent values. This includes the question of whether we should allow questionnaire data to be changed where they are inconsistent with administrative data.

Please note that following the feedback received from panel members (action 50 (4)), the methodology used to investigate the relationship between VOA number of rooms and other census household variables was changed from Cramer’s V to a simple regression (see Section 7). This has not had an impact on the findings.

# Data linkage and current datasets

Figure 1 Preparatory steps prior to the linkage of the Valuation Office Agency (VOA) data to pre-editing and imputation (pre- E&I) 2011 Census data

The test platform for the current project was 2011 Census data linked with Valuation Office Agency (VOA) data. Unique property reference numbers (UPRNs) were assigned to records in both datasets, which were then linked via these UPRNs. Figure 1 depicts the linkage process and numbers of records in the relevant datasets. As UPRNs are based on idiosyncratic address information, the false positive linkage rate was assumed to be low. Records with poor quality address information (which therefore could not be assigned a UPRN) and records with duplicate UPRNs, which may represent, amongst other things, houses in multiple occupation (HMOs), were removed from the data for this feasibility study and will be addressed in further research. The extract of VOA data was taken from July 2016; properties built from 2012 onwards were removed in order to make these data more comparable with 2011 Census data.

Note that the 2021 Census address frame will be built around UPRNs meaning they will not need manual assignment following data collection. Only additional addresses found during fieldwork will potentially not be assigned a UPRN so these numbers should be lower in 2021. Similarly, the coverage of VOA addresses with an assigned UPRN has improved. For 2021 we expect the largest challenge to arise from linking/mapping the address-level VOA dataset onto the household-level Census dataset i.e. accounting for HMOs. Further research on this topic is underway.

# Comparing number of rooms as measured by the 2011 Census and Valuation Office Agency (VOA)

[Previous research](https://www.ons.gov.uk/census/censustransformationprogramme/questiondevelopment/housingcommunalestablishmentsandvisitors/estimatingthenumberofroomsandbedroomsinthe2021censusanalternativeapproachusingvaluationofficeagencydata) (2) detailed the definitional differences between 2011 Census and the Valuation Office Agency (VOA) rooms variables. Essentially, the census version included kitchens, conservatories and utility rooms in its definition (which the VOA did not), whereas the VOA included storage rooms. As most properties contain cooking facilities, number of rooms was generally smaller in the VOA data compared to the census data (see Figure 2).

Figure 2 Distribution of number of rooms as measured in the 2011 Census and Valuation Office Agency (VOA) data

The distributional (and definitional) change does not necessarily represent a problem for E&I, however it is crucial to highlight this difference for continuity reasons when producing outputs for number of rooms.

# 2011 Census editing and imputation (E&I) methodology

The primary aim of the [2011 Census editing and imputation (E&I) strategy](http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/quality/quality-measures/response-and-imputation-rates/item-edit-and-imputation-process.pdf) (5) was to produce a fully complete microdata set with no erroneous values, missing values or inconsistent responses (also known as a utility dataset). For the data under consideration in the current report, a record would be inconsistent if number of bedrooms was greater than total number of rooms.

The 2011 Census E&I strategy used a donor-based imputation method where records with errors (‘recipients’) are assigned values from an error-free record (‘donors’) in the target variable (here, number of rooms). Potential donors are selected based on having similar characteristics on ‘auxiliary variables’ (here, other census household variables such as accommodation type, tenure, or number of bedrooms) as the recipients. This method aims to retain the multivariate structure of the data. Sections 5, 6 and 7 discuss potential challenges to the assumptions underlying this method posed by linked administrative data. These assumptions include all data being sourced from the same population and that missingness in the data is predictable from other variables in the dataset. Linked administrative data may also pose definitional and time-frame issues. The three broad challenges considered here are:

1. Record linkage failures (Section 5)
2. Missing data mechanisms (Section 6)
3. Relationships with other variables (Section 7)

# The impact of record linkage failures on editing and imputation

An important assumption underpinning editing and imputation (E&I) is that all data are sampled from a single population to ensure the pool of donors is as large and representative of the recipients as possible (unrepresentative potential donors can lead to biased or failed imputations). However, because one million 2011 Census records failed to link, these records have no corresponding Valuation Office Agency (VOA) rooms data. These records are known as Census residuals (see Figure 1). All of these records require VOA number of rooms to be imputed and donors can only be sourced from the linked data. For an unbiased imputation, the auxiliary variables used to inform the imputation of the Census residuals ideally need to resemble the auxiliary variables in the linked data (i.e. the residuals do not represent a wholly distinct subpopulation not found in the observed data). We therefore explored the distributions of auxiliary variables in both the residual and linked data.

Figure 3 Distribution of accommodation type in the residual and linked data

Figure 3 is representative of how auxiliary variable distributions differed between the Census residuals and linked data. Generally, the linked data contained higher proportions of larger terraced and semi-detached properties which were more likely to be owned outright or with a mortgage or loan, whereas the Census residuals contained higher proportions of properties such as caravans, rented converted houses and flats with small numbers of bedrooms (this likely reflects the inefficiency of the matching algorithm to capture and match address information for smaller or complex properties). Therefore, there was some evidence that particular subpopulations are overrepresented in the residual records. Crucially, however, there was overlap between residual and linked data in the distributions of all categories in all auxiliary variables. This means that unbiased imputation should be possible by including variables which explain distributional differences in the imputation model.

Despite the different method used for building the 2021 Census address frame, we still expect smaller or complex properties to be more likely to be missing from the address frame and hence to not get linked in 2021.

# Missing data mechanisms

As described in Section 1, failure to link is not the only source of missingness in the final dataset. The target Valuation Office Agency (VOA) data has both unsystematic missingness (from errors present in the data prior to linkage) and systematic missingness (from records failing to link). The editing and imputation (E&I) strategy treats all missing data in the same way, so the overall missing data mechanism also needs to be considered. Successful imputation of a target variable depends on missingness and observed values being predictable from other variables in the dataset. If missingness is non-random but cannot be predicted from observed data (i.e. predicting missingness depends on some factor not measured in the dataset), imputation is likely to yield poor results (6). We therefore examined whether missingness in VOA rooms could be predicted from auxiliary variables in the combined linked + residuals datasets (containing both systematic and unsystematic missingness, as these are the data which will ultimately require imputation).

Adding the one million Census residuals records to the data resulted in 5.4% missing VOA number of rooms. This compared with 1.1% missingness in the VOA variable when only considering the linked data (and 3.1% in the rooms variable captured on the 2011 Census questionnaire). Despite this higher level of missingness, if the mechanism(s) by which data are missing can be predicted from auxiliary variables, unbiased imputation should be possible.

Figure 4 Percentage of missing data in VOA number of rooms by accommodation type in the linked and linked + residual datasets

A similar pattern of missingness was observed in both the linked and linked + residuals datasets: higher levels of missingness were observed in generally smaller properties and buildings classed as commercial. However, this was accentuated in the linked + residuals data (compared to just the linked). Nevertheless, missingness was non-random and therefore predictable based on other observed variables. This means imputation should be feasible (although note the very high levels of missingness in some categories, for example 40% of rooms records were missing for commercial buildings in accommodation type).

Next, we examined missingness in the linked + residuals data according to local authority. Census data are divided into processing units containing one to seven contiguous local authorities for imputation. If certain local authorities contain disproportionate percentages of missing data, the proportion of suitable donors to recipients could be low and compromise the quality of the imputation within that local authority or processing unit. The three local authorities with the highest percentage of missing VOA number of rooms data were Isles of Scilly (41%), Kensington and Chelsea (35%), and Hammersmith and Fulham (35%). Although there is no established cut-off point for where the proportion of missing data becomes problematic for the validity of statistical estimates, inferences or imputation, higher levels of missingness lead to a greater risk of bias in the imputed data, and end-users interested in small area estimates may be concerned if high levels of the data have been imputed in certain local authorities.

# Relationships with other household variables

The 2011 Census and Valuation Office Agency (VOA) define number of rooms differently and as such do not measure precisely the same statistical concept as the 2011 Census number of rooms question. For our purposes, it was necessary to accept the VOA definition of rooms as this variable is intended to replace the census version. However, it is vital for the quality of the current editing and imputation (E&I) method that VOA number of rooms has similar statistical relationships with other household variables as 2011 Census number of rooms. If these relationships are of a different magnitude, then replacing the census version with the VOA version will mean the weightings and the imputation model would no longer be valid and the imputation could be biased.

Furthermore, the definitional change means use of VOA data led to an increase in records which were inconsistent with 2011 Census number of bedrooms. In the census data, 0.2% of rooms values were higher than bedrooms; in the VOA data this was 1.2%. This is in line with the number of edit rule failures within the census in general and is not problematic for imputation to resolve. However, the number of VOA rooms records which were equal to number of bedrooms as captured on the census (1.6% of 2011 Census records; 6.1% of VOA records) has implications for the continuity of outputs. Estimates regarding number of rooms from future linked census-VOA data are likely to be lower than those published previously and contain more properties where the number of rooms is equal to the number of bedrooms, presenting a possible discontinuity.

The source and time-frame of the VOA rooms data is also different from the census. The 2011 Census data are a snapshot of England and Wales on the day it was taken via self-report; the VOA data have been collected in various ways from different sources, updated over time until 2016. Coupled with the change in definition, this may impact on the level of association with auxiliary household variables.

Therefore, we examined the relationship between both census and VOA number of rooms and auxiliary household variables using several simple regression analyses. These consisted of separate analyses where census household variables were used as predictors and the number of rooms variables as separate outcome variables. Nominal household variables were dummy coded. The results for accommodation type are reported in Table 1. We report the unstandardized beta coefficient (*Β*) of the intercept and both unstandardized and standardised beta coefficients (*β*) of the predictor variables. If the obtained coefficients representing the relationship between rooms and other variables is similar for both versions of the variable, it can be inferred they are measuring similar statistical concepts. This was conducted on the linked data to keep a similar sample size between census and VOA number of rooms.

Table 1 Results of regression analyses predicting number of rooms from accommodation type

|  |  |  |
| --- | --- | --- |
|  | **2011 Census number of rooms** | **VOA number of rooms** |
| **Unstandardized beta coefficient (Β)** | **Standardised beta coefficient (β)** | **Unstandardized beta coefficient (Β)** | **Standardised beta coefficient (β)** |
| **Intercept** | 4.60 |  | 3.68 |  |
| **Detached** | 2.42 | 0.52 | 1.59 | 0.45 |
| **Semi-detached** | 1.04 | 0.25 | 0.68 | 0.21 |
| **Terraced** | 0.64 | 0.14 | 0.45 | 0.13 |
| **Purpose built flat** | -1.06 | -0.19 | -1.01 | -0.24 |
| **Flat within a converted or shared house** | -1.02 | -0.08 | -1.00 | -0.11 |
| **Above/within commercial building** | -0.63 | -0.02 | -0.62 | -0.03 |

*Reference category: Caravan or mobile temporary structure*

The difference in the value of the intercept between 2011 Census and VOA number of rooms was approximately 1. Additionally, the standardised beta coefficients for the household variables obtained from these analyses show remarkable consistency when predicting either 2011 Census or VOA number of rooms. Similar results were obtained for all census household variables.

Although the value for VOA number of rooms was generally lower than 2011 Census number of rooms (see Figure 2), this difference was fairly consistent (likely because most properties contain a single kitchen or other cooking facilities not measured by the VOA variable) but the relationship coefficients were generally similar. Therefore, it appears that the relationships with other variables is of a similar magnitude. Because of this, auxiliary variables used to inform the imputation can be weighted similarly for both variable versions.

# Imputing number of rooms

Data were imputed using a similar strategy as detailed in the [2011 Census Item Edit and Imputation Process report](http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/quality/quality-measures/response-and-imputation-rates/item-edit-and-imputation-process.pdf) (5), where comprehensive details of the process can be found. The software used was [CANCEIS](https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2009/wp.15.e.pdf) (7). In the current report, we focused on imputing the 2011 processing units which contained the ten local authorities with the highest percentages of missing Valuation Office Agency (VOA) data (as these data have the most potential to become biased following imputation). These processing units represent 1.7 million addresses covering both urban and rural geographies.

Initially, we used a joint imputation method where all household variables (including number of rooms) were imputed simultaneously. This method should preserve the multivariate structure of the data. Firstly, we conducted separate joint imputations on the linked + residuals data; once including 2011 Census number of rooms (12.5% missing data) and once including VOA number of rooms (16.7% missing data). Because of the focus on imputing rooms, auxiliary variables were weighted in the imputation based on their relationship with number of rooms. Results of these imputations were compared. Note that when using this imputation method, where the edit rule that number of bedrooms should be less than number of rooms is violated, either the census bedroom or VOA room variables could potentially be altered to make the data consistent. This means it is possible that census number of bedrooms could be lowered to make it consistent with VOA number of bedrooms, which means that administrative data could change survey data.

In order to assess the accuracy of imputation, a number of criteria should be met:

First, the conditional pre-imputation distributions should be preserved following imputation with any deviations explained by known missing data biases.

Second, multivariate relationships in the dataset should also have been preserved.

Third, post-imputation VOA number of rooms distribution should have a lower peak than the census equivalent, due to the different definitions of these variables (as in Figure 2).

Figure 5 Pre- and post-imputation distribution of 2011 Census and VOA number of rooms

Figure 5 shows the overlapping distributions of the two number of rooms variables. Census number of rooms tended to be greater than VOA number of rooms, as anticipated from the differing definitions of the variables (that is, the VOA variable had a lower peak than the census variable). The post-imputation census and VOA distributions were also similar to their respective pre-imputation distributions. The largest differences were in two- and five-bedroom properties in the VOA data at approximately one percentage point change. This is higher than the maximum difference between pre- and post-imputation distributions for household variables in the 2011 Census, but missingness was higher in the VOA rooms variable than any of the 2011 Census variables and the changes are consistent with known non-response biases in the data. For example, we know that flats and other smaller properties were more likely to be missing number of rooms (Figure 4) so would expect the proportion of records with fewer rooms to increase after imputation to reflect this if the imputation was successful. We can therefore conclude that CANCEIS was able to successfully impute the data using census household variables to predict VOA rooms and account for the characteristics of households with missing data.

Following imputation of the VOA number of rooms variable, 12.4% of records had the same number of rooms as bedrooms. This was 2.6 percentage points higher than in the pre-imputation distribution. In comparison, 4.7% of census records had the same number of rooms as bedrooms following imputation, which was 0.4 percentage points higher than the census pre-imputation distribution. This is likely due to more smaller properties (including bedsits) requiring imputation, where VOA number of rooms is likely to equal census number of bedrooms.

We also examined the regression coefficients between the two post-imputation number of rooms variables and auxiliary variables as in Section 7. All relationships were retained, and again census and VOA number of rooms exhibited similarly sized relationships.

This imputation was then repeated but constrained so that donors had to be sought from the same local authority as recipients, due to the uneven distribution of missingness between local authorities. However, this made little difference to results and is not reported further here. Sufficient donors were found even in local authorities where VOA missingness was high.

Next, we performed an imputation which reflects an alternative census processing strategy. The 2011 Census data were jointly imputed first then the VOA data were imputed separately as their own module using the imputed census variables in the matching variable set to condition imputation. This protects the observed census data and ensures they cannot be changed on the basis of what has been observed in the administrative source (e.g. where there is an edit rule failure). This differs from the previous imputation where census number of bedrooms could be changed on the basis of VOA number of rooms. This is in line with the edit rule principle that where there are any conflicts between census and admin data, the census variables should be favoured due to the inherent uncertainty in the linkage mechanism and the fact that the census has been specifically designed to capture the statistical concepts being analysed. It also prevents number of bedrooms being altered to be consistent with the VOA, rather than census, definition of rooms for some records. This prevents there being two slightly different definitions at play within one variable which is quite clearly unacceptable.

Figure 6 Post-imputation VOA rooms distributions for the standard joint and separate post-census processing imputations

Figure 6 compares the distribution of VOA number of rooms from the two imputations. The method used to impute these data made little difference to the obtained distributions; either imputing VOA rooms simultaneously with, or subsequently to, other household variables resulted in similar post-imputation distributions. Following resolution of data inconsistencies, 14.3% of records had equal values for VOA number of rooms and census number of bedrooms. This was 4.5 percentage points larger than the pre-imputation distribution.

We also verified there was little change in in the unstandardized beta coefficient (*Β*) of the intercept and standardised beta coefficients (*β*) of the predictor variables for the simple regressions predicting VOA number of rooms from the census household variables between the pre- and post-imputation distributions. This confirms that either the standard joint imputation or post-census processing approaches appear equally viable for the VOA data, with the post-census approach additionally aligning with the principle to favour census data over admin data where there are conflicts between the two.

# Conclusions, recommendations and future work

The results suggest that it is feasible that linked Valuation Office Agency (VOA) number of rooms can be imputed for 2011 Census household records, despite some assumptions being partially violated (definitional and time-frame issues, the possibility of subpopulations in the data). This shows that it is possible to predict VOA rooms from census variables both when there were missing records prior to linkage and when missingness was due to a data linkage failure. Additionally, the results show that the current census editing and imputation (E&I) method is robust enough to handle administrative data. In particular, it is also possible to follow a design principal to favour survey data over alternative data where the two are inconsistent.

Note, however, that some local authorities had large percentages of imputed data, and there was an increase in the percentage of properties where VOA number of rooms was equal to census number of bedrooms. This affects end-user interpretation of the data. While the research was carried out using 2011 Census data, the findings should largely hold true for the 2021 Census as no major changes are expected to either the census household data or VOA data.

We can also be confident in the linking procedure i.e. that the correct census and VOA addresses were matched for records where a unique unique property reference number (UPRN) could be assigned. This confidence should increase in 2021 where UPRNs will be present on the source datasets and will only need to be assigned by a separate process for a small proportion of addresses.

Crucially, our results apply to only this one variable from this one administrative source. They should not be seen as a more general endorsement that all linked survey-administrative data can be effectively treated by standard E&I procedures. We recommend this kind of feasibility research should be conducted on any administrative variable which is intended to be linked to the census, or which is under consideration to replace a census variable. Every administrative variable from every potential source will have its own particular issues to consider.

Although this report serves as proof-of-concept of imputing administrative data, note that we had to remove records with no, or duplicate, UPRNs prior to analyses. This means the E&I strategy to provide a fully populated and consistent post-imputation set of microdata could not be fulfilled at this time. In future, the census and VOA will use the same address register to govern coverage, which should minimise the number of addresses with no UPRN. The current research should be repeated with 2019 Census rehearsal data to verify this assumption. However, the census distinguishes when a property contains multiple households which the VOA does not (hence the difficulty with duplicate UPRNs). Determining the best method of dealing with the differences between households and addresses requires further research.

# References

1. Office for National Statistics (2018). Help shape our future: The 2021 Census of population and housing in England and Wales. Available online [02/06/2019]: <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/765089/Census2021WhitePaper.pdf>
2. Office for National Statistics (2017). Estimating the number of rooms and bedrooms in the 2021 Census: An alternative approach using Valuation Office Agency data. Available online [02/06/2020]: <https://www.ons.gov.uk/census/censustransformationprogramme/questiondevelopment/housingcommunalestablishmentsandvisitors/estimatingthenumberofroomsandbedroomsinthe2021censusanalternativeapproachusingvaluationofficeagencydata>
3. Office for National Statistics (2014). 2011 Census Quality Survey. Available online [02/06/2020]: <https://www.ons.gov.uk/file?uri=/census/censustransformationprogramme/consultations/the2021censusinitialviewoncontentforenglandandwales/2011censusqualitysurveyreport.pdf>
4. Office for National Statistics (2019). 3rd Meeting of the Methodological Assurance Review Panel - Agenda & Minutes. Available online [02/06/2020]: <https://www.statisticsauthority.gov.uk/wp-content/uploads/2020/05/MARP03-Minutes.pdf>
5. Office for National Statistics (2012). 2011 Census Item Edit and Imputation Process. Available online [02/06/2020]: <http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/quality/quality-measures/response-and-imputation-rates/item-edit-and-imputation-process.pdf>
6. De Waal, T., Pannekoek, J., and Scholtus, S. (2011). Handbook of statistical editing and imputation. Wiley, New Jersey.
7. Bankier, M. and Crowe, S. (2009). Enhancements to the 2011 Canadian Census E&I System. Conference of European Statisticians, Work Session on Statistical Data Editing, UNECE.