Alternative Household Estimate 2021

1 – Introduction

1.1 – Counting Households

Households are defined as properties containing usual residents that live together and share facilities. Excluded from this definition are communal establishments. A communal establishment is a place providing managed residential accommodation. "Managed" here means full-time or part-time supervision of the accommodation. Communal establishments don't include sheltered accommodation, serviced apartments, nurses' accommodation, and houses rented to students by private landlords. These are considered to be households.

The main method for enumerating households in the country is simply to count the number of properties that contained usual residents, as found through a Census return. For some addresses, valid responses were received stating that there were no usual residents of the property, for instance if the address was someone's holiday home.

The remaining addresses in the country are *non-responding* and therefore further work must be carried out to determine how many of these did in fact have usual residents. This is calculated using a dual-system estimation (DSE) approach through the Census Coverage Survey (CCS), which samples postcodes across the country with the aim of estimating how many addresses with usual residents were missed by the Census.

1.2 – The need for dependence bias adjustment

The DSE has a key assumption – that the probability of a property responding to the CCS is independent of the probability that the property responds to the Census. The breakdown of this assumption leads to the potential for bias in its outputs. This *dependence bias* must be corrected for if present, and the method for doing this is to enumerate the number of occupied addresses in an alternative way (a so-called Alternative Household Estimate or AHE). These AHEs are created for subsections of the population and used to adjust DSE results accordingly. The methodology for using AHEs to adjust for dependence bias is given by EAP160 - Adjusting for dependence bias in coverage estimation.

1.3 – Alternative Household Estimate

Simply put, an AHE is a count of the number of properties with usual residents that are not communal establishments. This figure is the sum of two contributions:

- A count of the properties known to be occupied from a Census return
- An estimate of the properties thought to be occupied, but where Census return is either missing or inconclusive.

The method of estimating the number of addresses that should be included in the latter of these groupings is the challenge this paper aims to address.

1.4 – The key assumption: Inferring occupancy of non-responders from responders

Any method of estimating the number of occupied non-responding properties will require some *source of truth* that can be used to make informed judgments. In 2011 and all methods suggested in the paper, this source of truth has been responding addresses. The key assumption of the AHE aspects that rely on this is therefore that responding addresses are sufficiently representative of non-responding addresses to draw inferences on the occupancy of the latter from "similar" properties in the former.

This assumption has some obvious flaws. The most prominent of these being that intuitively one would expect a vacant property to be less likely to have a Census form returned, especially in the context of Covid-19 restrictions preventing visits to a second home.

We will look at validating this assumption by using the available data in different ways and assessing the feasibility of the outcomes. The subsequent sections set out the first two approaches mentioned below in detail, but essentially, we can produce alternative numbers and sense-check the results against each other in the following ways:

- A logistic modelling approach, trained on data using the assumption above
- A method that is similar to 2011, but makes use of available administrative data, which also makes use of the assumption above
- Just believing the field observations, which makes no reference to received responses
- Just believing the admin sources, with no reference to received responses
- Aggregate levels of occupation from council tax data (the 2011 QA method)

1.5 – Improvements in the last decade

Since the production of AHEs at the last Census, improvements have been made to data quality and linkage that may aid the production and accuracy of the estimates. The Census Intelligence Datastore (CID) is one of these key improvements. CID is an address level collation of all information that Census has access to, including administrative data, Census and Field responses, and Census collection management information, for example the number of forms that have been requested.

A further benefit of CID is that it contains address level information for every property that interacts with Census across England and Wales. Previously, information used to create the AHE was limited only to samples of key information in specific areas (CCS postcodes). This wealth of data should therefore improve the quality of decision making due to an increase in data size, even if no other changes are made to the AHE methodology.

The primary focus of this work has been to determine the best way of incorporating the intelligence from these data sources into assessments of the address's occupancy.

In this paper, we investigate two possible ways to incorporate the intelligence from CID into an improved AHE. The first of these is conceptually similar to the approach taken in 2011, while the second uses modelling to estimate the likely occupancy of addresses.

2 – Idealised approach: Alternative Household Count

2.1 – Dummy forms as a source of occupancy information

The goal of finding the number of non-responding households seems to have an obvious solution: field officer *dummy forms*. Field officers complete these for non-responding addresses, providing assessments for key characteristics including occupancy. Ideally, we might therefore be able to create simplistic Alternative Household Counts (AHCs) that count the number of addresses assessed as occupied by either a field officer or a census form.

Dummy forms would ideally be completed for all addresses where a census form does not indicate occupancy, categorising them as either occupied or vacant. In practice, addresses were split across the below categories in 2011:

1. Properties assessed by field officers to be occupied *(dummy occupied)*
2. Properties assessed by field officers to be vacant *(dummy vacant)*
3. Properties where field officer visited but was unable to assess occupancy *(dummy non-contact)*
4. Properties where field officer visited but the dummy form available did not contain a valid value for the occupancy of the address (*dummy invalid).*
5. Properties where no field officer assessment of occupancy was recorded. These are described as unaccounted for addresses (UFAs)
6. UFAs where a blank questionnaire was submitted. Although these properties are technically not *non-responding*, we do not have sufficient information to determine their occupancy for the purposes of an AHC.

Note that raw dummy form data does not provide these categories; they are instead created by the aggregation of different *reasons for dummy completion* options, as shown in *annex A*.

For an AHC to correctly assess the number of non-responding addresses that are occupied, categories 1 & 2 (occupied and vacant addresses) should be as accurate as possible, with minimal addresses in categories 3 – 6.

2.2 – Issues with AHCs in 2011

For an AHC to be a feasible way of counting the number of occupied addresses in an area, the true occupancy rate for addresses assessed as occupied by field officers would need to approach 100% while the corresponding proportion for allegedly vacant addresses would need to roughly approach 0%.

To assess whether this were the case, a source of truth for occupancy was required. Fortunately, some addresses that responded to the Census also had occupancy recorded by field officers on a dummy form – these are subsequently referred to as *doubly-captured addresses*. A key assumption was that the known occupancy of *doubly-captured addresses* are representative of the non-

responding addresses with which they shared the same assessment of occupancy by a dummy officer (*dummy-only addresses*). The breakdown of these known occupancy rates for each category of dummy completion are given below in *table 1*:

Table 1 - Occupancy for each category of doubly-captured addresses in 2011

Category	Occupancy proportion (correction based on observed responses and clerical work)
Census form - Occupied	100%
Dummy form - Occupied	81%
Dummy form - Vacant	50%
Dummy form – No contact/non-response	86%
Dummy form - Invalid	73%
UFA	47%
Blank form submitted; no dummy information	5%
Additional addresses	100%

As the above shows, there was a substantial degree of misclassification by dummy officers. This meant that the production of a simplistic AHC was not possible and hence an Alternative Household Estimate (AHE) was required.

2.3 – Issues with AHCs in 2021

Before attempting to replicate the 2011 AHE methodology for the 2021 Census, it is worth verifying that an AHC approach remains inappropriate. As previously noted, data is not yet available to assess this, but we will conduct this analysis when possible.

One factor which may increase the difficulty of dummy form occupancy is the effect of the pandemic. Properties owned by residents with a second home may have been vacant around Census day due to restrictions but would still be classed as having usual residents. When data become available, this could be observed in an increased misclassification rate of dummies compared to 2011.

3 – Approach followed in 2011

3.1 – Adjustments for data quality in 2011

Although an AHC was not judged to be possible in 2011, the AHE developed was still heavily based on the field officer assessment of occupancy from dummy forms. As stated in section 2.2, the Census occupancy of *doubly-captured addresses* for each dummy occupancy value in an Estimation Area

(EA) was assumed to be representative of the actual occupancy of the corresponding *dummy-only addresses*. This assumption formed the basis of the 2011 AHE calculations.

The known occupancy rate (according to Census returns) for each class of *doubly-captured address* was used as a scaling factor when determining how many corresponding *dummy-only addresses* should be included in the final AHE for the EA. A worked example of this is given in section 3.2.

For addresses where no dummy information was recorded, clerical work was carried out to determine their occupancy. This has not been carried out for 2021 work and is therefore not discussed further within this paper. Another issue that was similarly present in 2011 but not in 2021 is addresses that submitted blank Census forms. In 2011, these properties did not receive the same follow-up as other non-responding addresses – a problem that has been resolved in 2021. A final complication present in 2011 that is no longer relevant for 2021 was the discovery of new addresses that existed on Census Day, but which did not receive an invitation to respond. In 2021, addresses were dynamically added to the address frame throughout the collection period and followed up accordingly.

For each EA, an AHE was created as the adjusted count of properties in each category of dummy response. The contribution of addresses in each category across the entirety of England & Wales in 2011 is shown in *table 2*, while *figure 1* shows the frequency of each occupancy rate across EAs.

Figure 1 - Histogram showing the distribution of occupancy rates across EAs in 2011.

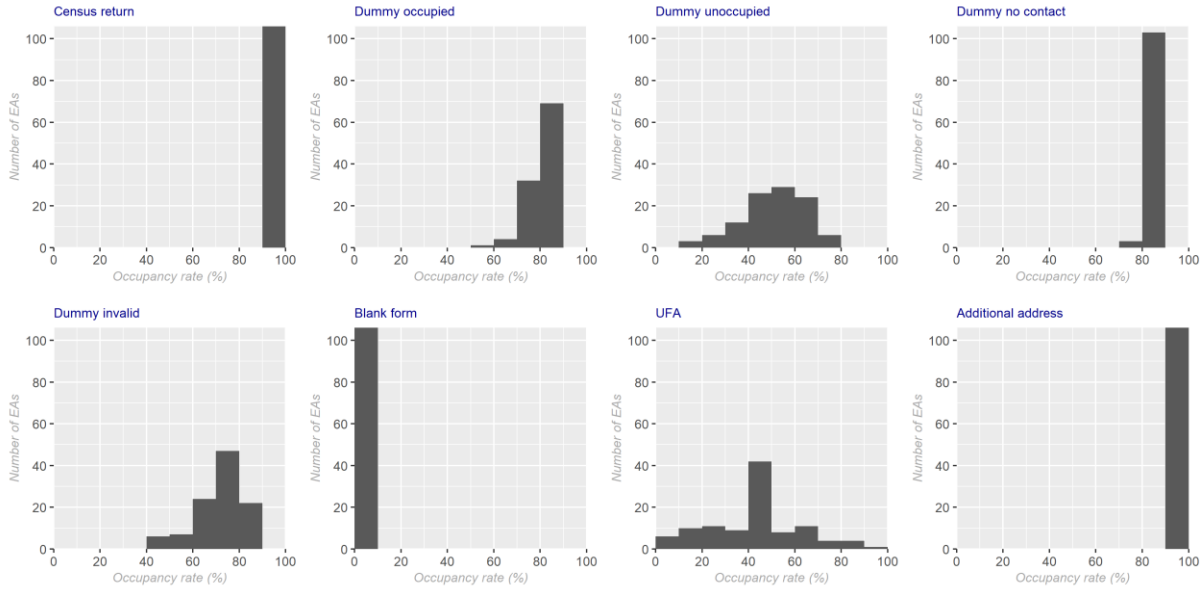

Table 2 - Contributions from each category of dummy response to the 2011 AHE (summed across all of England and Wales). Note that these are not the same as the correction factors given above but are the final number of households in each category added to the AHE calculated using the occupancy rate to scale down the total count of addresses in the category.

Category	Contribution to Alternative Household Estimate
Census form - Occupied	93.7%
Dummy form - Occupied	0.4%
Dummy form - Vacant	1.2%
Dummy form – No contact/non-response	4.2%
Dummy form - Invalid	0.0%
UFA	0.3%

Blank form submitted; no dummy information	0.0%
Additional addresses	0.1%

3.2 – 2011 worked example

Table 3 provides a worked example that shows how the number of occupied non-responding addresses was estimated for a fictitious EA. As discussed previously, the occupancy adjustments for UFAs, blanks and "additional" addresses were determined clerically in 2011. These have been included for the sake of completeness, but this approach is not necessary in 2021.

Table 3 - 2011 AHE worked example

Stratum (dummy form status)	Number of Addresses			Occupancy adjustment	Estimated number of occupied Non-responding addresses
	Census return indicates occupied	Census return indicates vacant	No census return		
Dummy form occupied	950	50	300	95%	285
Dummy form vacant	500	300	100	62.5%	62.5
Dummy form – No contact/non-response	400	100	200	80%	160
Dummy form invalid	80	20	50	80%	40
UFA	-	-	*100*	*50%*	*50*
Address filled in blank questionnaire	-	-	*100*	*5%*	*5*
"Additional" address	-	-	*20*	*100%*	*20*
Total non-responding properties estimated to be occupied:					**622.5**

The AHE for this fictitious EA would then be calculated by adding the estimated component calculated above to the number of properties that declared their occupancy through a Census form.

4 – 2021 Improvements

4.1 – Census Intelligence Datastore

CID links the Census outcomes for a property to its administrative data and management information (MI) held by Census. MI comes in two forms - data on interactions with field officers (Fieldwork Management Tool, FWMT) and data listing any other interaction the address has had with the census process (Response Management data, RM). The full list of variables available in CID is given in *annex b.*

4.2 – Reframing AHEs as probabilities

Another change to the AHE for 2021 is the level of aggregation of the final figures we are producing. Reinterpreting the calculations from 2011 provides a way of producing this.

We have previously defined the 2011 methodology as applying correction rates across each stratum, i.e. "keeping 70% of 1,000 properties in the stratum". An alternative way of interpreting this approach is assigning each of the properties in the stratum with a probability of being occupied, i.e. "each of the 1,000 properties has a 70% chance of being occupied". This is illustrated below:

4.3 – New addresses

In 2021, addresses not included on our frame at the start of collection, including newly registered properties, were added weekly to the field workload, and so into our base data. We will validate the occupancy of these new additions by searching for newly found addresses that were present on the census period cut of the PDS register that were not present on the cut of this at the start of the year.

4.4 – Approach to UFAs

The majority of non-responding addresses for which an estimate of occupancy must be made will have received a dummy form. The estimation of occupancy for these addresses is the main focus of this paper. For the remaining properties (so-called Unaccounted for Addresses, UFAs) a different approach is required. *Table 5* lays out how these plan to be estimated, as well as including the treatment of other properties for the sake of completeness. The treatment of UFAs may change as the AHE is developed, depending on the outcome of research into this particular group of addresses.

Table 5 – AHE categories and their assumed occupancy (AHE probability). Note that the following categories are mutually exclusive and populated in the order given below.

Ref	Category	AHE Probability
0.	**Excluded addresses**	
0.1	Confirmed invalid addresses (e.g commercial properties, duplicates included in original frame)	Exclude from data
0.2	Communal Establishments	Exclude from data
1.	**Responses**	
1.1	Census responding households containing usual residents	100%
1.2	Census responding households with no residents, or visitor-only	0%
2.	**Non-responding addresses, with dummy collected by Field**	
2.1	With dummy indicating occupied (absent, refusal)	Calculated through approaches described in remainder of paper
2.2	With dummy implying occupied (no response, no contact)	
2.3	With dummy indicating vacant (vacant, 2nd residence, holiday home)	
2.4	Dummy, but reason for dummy missing	To be confirmed. Small numbers in this category so

		investigating potential of treating as UFAs as below.
3.	**Unaccounted for Addresses (no response or dummy form)**	
	3.1 Have some form of sign of activity from RM/FWMT data	
3.1.1	RM indication that household refused (eg from calls to the Contact Centre)	100% (unless reason for refusal implies otherwise)
3.1.2	Addresses that had at least one field visit that made contact, or there was any fulfilment request in RM	100%
3.1.3	Addresses where every field visit was no contact, and there are no signs of activity in RM	0%
	3.2 UFAs that didn't get a field visit in last 3 weeks, so didn't get a dummy response	
3.2.1	Paper-first areas removed from follow-up, that haven't responded and not included in the above	To be confirmed
3.2.2	Areas where there wasn't enough field staff	To be confirmed
3.2.3	Caravan parks marked as locked during lockdown, valid but vacant	0%
3.3	Additional addresses valid as at census day (potential new builds not captured in original address frame)	Rate based on new patient registrations found

5 – 2021 Method A: 2011+

5.1 – Building on the 2011 method

As outlined in *section 4*, there is now significantly more address level information which might be useful in creating AHEs in 2021. The challenge comes from finding a way in incorporating these new data sources.

The 2011 approach separated our dataset into strata (dummy form outcomes) and assigned a probability of occupancy to each based on an occupancy rate for the strata as calculated with a source of truth (census returns). The two improvements that could be made to the 2011 methodology without substantially altering it can therefore improve one of two things:

- the source of truth
- the choice of strata

5.2 – CID indicators as sources of truth

CID has been deliberately designed to include all address-level indications of occupancy that it was possible to link to our Census data. Although many of these indicators had high correlation with the Census occupancy indicators (significance levels given in *section 5.4)*, none had a high enough level of agreement with Census return occupancy to aid the creation of an AHC as described in section 2.2.

It is possible that discrepancies between the occupancy indicated by an admin data source and a census return are caused by inaccuracy of the latter, for instance due to spurious returns for a property. For the purposes of the AHE we discount this possibility and treat Census returns as an absolute source of truth.

5.3 – CID indicators as improved strata – the 2011+ approach

As none of the available administrative data is sufficiently robust when predicting occupancy to use as a source of truth, we have focussed on better stratification as our primary approach to directly improving the 2011 methodology. We refer to this method throughout the paper as the 2011+ methodology. The 2011+ approach therefore follows the 2011 approach but splits addresses into a greater number of smaller categories. The choice of these categories is explained in *section 5.4*.

The 2011+ method also simplifies complexities in the 2011 method. UFAs are treated as another category of dummy form status (and combined with *dummy form invalid* addresses), while so-called "*additional" addresses* and properties returning blank forms were followed up elsewhere in the 2021 collection process removing the need for explicit correction in the AHE. One further change that has been explored is the limiting of stratum size. If stratum size dips below a given threshold, the occupancy probability for the stratum is recalculated with one fewer variable. Currently this threshold is set as 5 for both the number of occupied and vacant responding addresses. A future area of work would be to tune this parameter to optimise results. Feedback at MARP suggested a threshold of 20 would be more robust and sensitivity analysis should be undertaken to check the impact of different strata sizes.

For the purpose of the worked example in *table 6*, the data has been stratified by two variables - dummy status and a combined sign-of-life flag (as detailed later in *section 5.4*).

Table 6 - Worked example of 2011+ AHE methodology

Stratum		Number of Addresses			Occupancy adjustment	Estimated number of occupied non-responding addresses
Dummy category	Combined sign-of-life (see below)	Census return indicates occupied	Census return indicates vacant	No census return		
Occupied	Occupied	495	5	150	99%	148.5
Occupied	Vacant	300	200	50	60%	30
Occupied	RM interaction	200	70	90	74%	66.7
Occupied	No indication	40	10	20	80%	16
Vacant	Occupied	40	40	150	50%	75

Vacant	Vacant	200	200	50	50%	25
Vacant	RM interaction	200	30	110	87%	95.7
Vacant	No indication	75	25	100	75%	75
No contact	Occupied	80	20	120	80%	96
No contact	Vacant	30	20	40	60%	24
No contact	RM interaction	60	20	60	75%	45
No contact	No indication	80	5	20	94%	18.8
UFAs		*N/A – calculated differently*		5	100%	5
Total non-responding properties estimated to be occupied:						**720.7**

Note that UFAs are not included in this treatment. These are treated separately as laid out in section 4.4.

5.4 – Choice of variables to stratify by

Following this methodology immediately presents us with a challenge: the choice of which indicators to stratify by. Due to the wealth of information in CID, if we used all available variables, we would quickly create tiny strata containing very few addresses. Indeed, strata size proved challenging for some local authorities even in 2011 when stratifying only by dummy outcome. Due to this, decisions need to be made on the variables to stratify by. We are investigating several possibilities for this:

1. Stratify only by the condensed *dummy reason* variable. This is as close to the 2011 methodology as can be repeated without the existence of clerical checks and can therefore be used as a benchmark to see whether we are actually improving on the previous methodology. Strata that previously had occupancy rates determined clerically are now calculated consistently with other strata, through occupancy of comparable *doubly-captured addresses*.
2. Using *dummy reason* (as in 2011) in combination with a derived variable (DV) for a so-called *sign of life*. This allows us to condense administrative signs of life from a variety of sources into one concise variable, mitigating the issues from stratum size that would occur if treated individually. The value in this column is assigned in the order given below. Note that an indication of vacancy is taken as a stronger indicator of the absence of usual residents than the admin signs of life indicator, which in turn is taken as a stronger indication of occupancy than the RM indicator. This is based on analysis of preliminary data, but these assumptions will be verified once finalised data is available.
    a. Vacant - indicated by flags from Council Tax (CT), PDS or utilities data.
    b. Admin data sign of life – from all of the above sources as well as English School Census.

c. RM interaction
d. No indication

3. Stratify by the best combination of variables. Best is defined using *Logistic Loss* as outlined in *section 7.1*. Most common "best variables" were presence of PDS registration, occupancy according to dummy form and reason for dummy completion. We do not intend to utilise this method in the final version of the AHE due to it resulting in less consistency across local authorities

A final point to note on stratum choice is the need for quality assurance. We have taken the decision to incorporate all admin and RM indicators into our sign of life variable, but this means that there are no obvious sources that could be used to independently verify our results. An alternative stratification method might therefore be to deliberately leave out a good indicator for use in later QA.

As outlined in *section 4.2*, the occupancy adjustment value could be interpreted as a probability of occupancy for all non-responding addresses in the given stratum, allowing the production of an address-level dataset. This dataset could then be filtered to produce final AHEs at any level of granularity – not just for the variables we chose to stratify by.

5.5 – Dealing with small stratum sizes

We have previously noted a consideration for the 2011+ approach is not stratifying our dataset too much, which would cause occupancy rates to be calculated using low counts. One way to minimise the effect of this is to define a minimum observed count for vacant and occupied addresses within a stratum. If the counts fall below this limit, we have instead applied the occupancy rate for a more generic stratum of which the small stratum is a subset. We have investigated doing this in two ways, namely by geography and by decreasing the number of stratification variables. Results are given in *section 8* for both. A worked example for the former of these approaches is given in *table 7* below. As with our data, the stratum size limit has been set at 5.

Table 7 - example of stratum aggregation in 2011+ methods

UPRN	Occupied addresses in most specific stratum	Vacant addresses in most specific stratum	Occupancy rate for most specific stratum	Occupied addresses in middle stratum	Vacant addresses in middle stratum	Occupancy rate for middle stratum	Occupied addresses in least specific stratum	Vacant addresses in least specific stratum	Occupancy rate for least specific stratum	Final AHE prob.
1	20	4*	83%	40	10	**80%**†	22,000	3,000	88%	80%
2	250	80	**76%**†	500	300	63%	22,000	3,000	88%	76%
3	9	1*	90%	20	4*	83%	22,000	3,000	**88%**†	88%

** value below threshold*
† final AHE value origin

In the above case, the three possible strata for each row are our desired variables to stratify by (for instance *dummy occupancy* and *sign of life*), all but one of these variables (just *dummy occupancy*) and all but two of these variables (i.e., the whole LA without subsetting). When performed geographically, all stratification variables are used for all strata, but the basis used to calculate these rates vary. For the most specific stratum the basis is local authority (the default for other approaches). The other two stratum sizes use delivery group and NUTS1 region as our basis datasets.

We are also investigating the use of ONS area classifications as an alternative for "similar geographies" that can be used when generating aggregated strata.

A final change made to the 2011+ methodology to prevent stratum size issues has been to merge "invalid dummy form" into our UFA category. This is due to the low numbers involved making many strata with this variable too small even at the highest level of geographical aggregation.

5.6 – Treatment of addresses without a valid dummy form

Recall our key assumption for our AHE, namely that we are aiming to infer occupancy of non-responding addresses from *similar* responding addresses. This poses a problem when we attempt to apply the above methodology to non-responding addresses that did not receive a dummy form assessment of occupancy – (unaccounted for addresses, UFAs). UFAs are unlikely to be well represented by responding properties that did not receive a dummy form. For this reason, the estimated occupancy probability for UFAs is calculated through a modified 2011+ approach where we do not subset by our reason for dummy form completion. This is currently an area of investigation while we await final data that will allow us to assess the characteristics of UFAs.

6 – 2021 Method B: Modelling

6.1 – Modelling probabilities of occupancy

The reframed approach to the AHE of computing the probability of an address being occupied lends itself to the use of modelling.

Due to its similarity to the reframed 2011 methodology, we have focussed on logistic regression as our modelling approach. This finds how each variable affects the likelihood of an address with that variable being occupied and allows a probability to be computed for new data. Recursive feature elimination has been used to find the best combination of variables to include in our model, alongside cross-validation of train/test splits.

Analogously to the 2011+ approach, we have trained different models for each local authority and predicted occupancy probabilities for all non-responding addresses. These probabilities are summed to provide the estimated number of *occupied* non-responding addresses and subsequently added to the number of addresses known to be occupied through a Census form to give an overall AHE for the Local Authority (or any other subset of the data as required).

6.2 – Modelling-based AHE example data journey

The data journey to creating a modelling based AHE is given below.

1. Initial data in CID is filtered for the local authority of interest (*df0)*

2. This data is split into three groups – responding addresses (*df1*), non-responding addresses (*df2*), and UFAs
3. A logistic regression model is trained to predict the known occupancy values for *df1*. Recursive feature elimination iteratively finds the best combination of variables to use, with cross-validation used to ensure the model is less likely to randomly overfit to a specific subset of our data.
4. This model will produce coefficients for each predictor. For each address in *df2,* these coefficients are used to generate a probability that the address is occupied. These results are stored in a new data frame (*df3).*
5. Filter *df3* for required level of granularity of final AHE. For instance, keeping only records where Hard to Count index is 3 and Accommodation type is Detached. This filtered data frame is saved as *df4*.
6. Sum the probabilities from *df4* to estimate number of occupied non-responding households ($AHE_{non\text{-}resp}$). For instance, if we have three addresses in *df4* with probabilities of occupancy of 0.2, 0.8 and 0.9, $AHE_{non\text{-}resp}$ would be 1.9.
7. Count the addresses in *df1* where census form indicated property was occupied (AHE_{resp})
8. Steps 2-7 are repeated for UFAs. This means that our *dummy reason* category is removed as a predictor from *df1* when training a new model to be applied to *df3*
9. The final AHE is created by summing the components calculated in steps 6 & 7 (AHE = $AHE_{non\text{-}resp}$ + AHE_{resp})

7 – Comparison of methods

7.1 – Validation of results with Logistic Loss

One advantage of the reframing of the 2011 AHE in terms of probabilities is that it provides us with a way of assessing the quality of its predictions. When given a list of probabilities of an event occurring alongside a source of truth for what actually occurred, *Logistic Loss* is a metric which adequately assesses the quality of those predictions. This is given by the below formulae:

$$i) \quad LL_{address} = \left[occ_{census} \times log\left(p_{occ}\right)\right] \; + \; \left[(1 - occ_{census}) \times \log\left(1 - p_{occ}\right)\right]$$

$$ii) \quad LL_{method} = -\frac{1}{n}\sum_{1}^{n} LL_{address}$$

$occ_{census} = 1 \; when \; census \; return \; indicates \; occupied; 0 \; when \; return \; indicates \; vacant$
$p_{occ} = probability \; that \; address \; is \; occupied$
$n \; = number \; of \; addresses \; with \; known \; census \; occupancy \; in \; test \; dataset \; (not \; used \; to \; train \; the \; method \; itself)$

Formula i shows how Logistic Loss values are computed for each address, assessing how similar the predicted and actual occupancy values are to one another. Formula *b* shows how each of the Logistic Loss values from formula i are combined together to give a metric for the method as a whole. This is a simple average of individual values, multiplied by negative one to fit with the convention that lower numbers indicate a better prediction.

As shown by the above formulae, the logistic loss for a given method is dependent only on the results it outputs and is independent of the method's complexity.

A key point to note about Logistic Loss as a concept is that it does not judge the ability of a method to label the outcome of an individual event, i.e., whether a given address is occupied or vacant. This sets it apart from many other commonly used metrics for addressing predictive power, such as accuracy and other approaches found through the use of a confusion matrix. This suits analysis of the AHE where we do not care about the occupancy label for any given property but rather the impact it has on the aggregate number of occupied properties of any group it is contained in.

7.2 – Test/train splits

One crucial modification we must make to any AHE methodology before assessing its predictive power is to assess the quality of fit on unseen data. For the modelling approach, this occurs in addition to cross-validation as employed in model training. For the 2011 and 2011+ approaches, we randomly split the data 10 times with one portion used to generate AHE probabilities and the remaining portion used to assess this quality. We are exploring the possibility of averaging predictions made across each of these splits but currently intend to generate AHE probabilities using 100% of the available *doubly-counted* addresses.

7.3 – Other metrics

We plan to conduct likelihood ratio tests to assess whether or not our fitted models are significantly better at predicting the occupancy of a property than the method used in 2011. We will also compare the AHE values estimated by each method to assess the relative counts given for each approach and compare these to estimates from administrative data sources to verify the outputted values remain feasible.

8 – Summary

Once data become available, we will apply the above methods and associated metrics to determine which approach is suitable for calculating AHEs. We will conduct further analysis to achieve the best quality results from each of the methods. One such area to be investigated is the limit to the size of each strata for the 2011+ approach, currently set at a minimum of 5 addresses in each category of Census response (occupied and vacant).

Unless results indicate that the modelling approach is sufficiently higher quality than the 2011+ approach, we suggest using the latter to generate our final AHEs due to the following benefits:

- It is more consistent across local authorities than the modelling approach where different predictors would be used for each.
- It is analogous to the approach followed at the previous Census
- It is less complex and easier to conceptualise for users.

Despite these points, the inclusion of a greater wealth of record level information in the modelling approach implies the potential for significantly improved predictions relative to the 2011+ methods. We therefore will attempt both methods and use the metrics outlined above to make an evidence-based decision on our final approach.

Annex A – Dummy occupancy lookup

The familiar dummy occupancy strata used throughout the paper are not present in the original data but are instead formed through the simplification of the more complex *reason for dummy completion* variable, as well as the presence of a refusal as recorded in RM. The components of these categories are shown below:

Reason for dummy completion (raw data)	Dummy reason derived variable 2011	Dummy reason derived variable 2021
Holiday accommodation	Dummy vacant	Dummy vacant
Second Residence		
Vacant household		
Vacant property		
Absent Household	Dummy occupied	Dummy occupied
Extraordinary Refusal		
Hard Refusal		
Non-return or no contact	Dummy no contact	Dummy no contact
Null value in column (dummy form collected)	Dummy invalid	UFA
Null value in column (no dummy form collected)	UFA	UFA
Null value in column. Refusal recorded through RM	UFA	Dummy occupied

Annex B – Other variables in CID

Source	Variable	Explanation	Possible values
2011 Census	Occupancy indicator	States whether household contained usual residents in 2011	1. Occupied, 2. Vacant, 3. Unknown
Council Tax (CT)	CT Occupancy	Consolidates council tax discount and exemption codes into one occupancy variable	1. Occupied 2. Vacant 3. Unknown
English School Census (ESC)	ESC occupancy	Flag that says whether the address has a record in the ESC	1. Occupied 2. Not on register
Fieldwork Management Tool (FWMT)	Dummy accommodation type	Field officer assessment of accommodation type.	Split into 8 categories of response as well as "unknown" value. Aligned with Census definition.
	Dummy reason	*See Annex A*	
	FWMT inaccessible reason	Reason a field officer was unable to access the property	1. Concierge 2. Other 3. Secure Entry 4. Gated community 5. Not present
	FWMT outcome	Derived variable from field outcomes. Split into 7 categories. Looking to remove in future iterations	1. Contact made 2. No contact made 3. Vacant 4. Impossible to occupy 5. Contact made; occupied 6. Information other than occupancy 7. Unknown
	Officer ID	Field officer who filled in dummy form	One per officer in data
Higher Education Statistics Authority Dataset (HESA)	HESA student quartile	Number of students in postcode area per address with that postcode in CID. Expressed as a quartile across E&W	*N/A - numeric*
NHS Personal Demographics Service (PDS)	PDS occupancy	Looks at PDS register to see addresses that have recent indication of occupancy or vacancy. If neither present then variable states whether property is on register or not.	1. Occupied 2. Vacant 3. On register 4. Off register
Response Management (RM)	Hard to Count	Geography variable. Directly from RM	All values 1 --> 5
	Hard to Count (Digital)	Geography variable. Directly from RM	All values 1 --> 5

RM continued	MSOA	Geography variable. Directly from RM	All MSOAs in data
	RM address type change count	Sum of all such events recorded by RM	*N/A - numeric*
	RM invalidated address count	Sum of all such events recorded by RM	*N/A - numeric*
	RM PQ Requested	Sum of all such events recorded by RM	*N/A - numeric*
	RM UAC	Sum of all such events recorded by RM	*N/A - numeric*
	RM undelivered mail count	Sum of all such events recorded by RM	*N/A - numeric*
Electralink, XOServe and ECOES datasets	Utility occupancy	Condenses indications from three utility datasets into one occupancy variable	1. Occupied 2. Vacant 3. Unknown
Valuation Officer Agency	VOA property type	Property type according to admin data source, condensed into more general categories	1. House 2. Bungalow 3. Maisonette 4. Flat 5. Mobile Structure 6. Annex