[OFFICIAL]

Developing a Hard to Count (HtC) Index for Scotland's Census 2022

Reviewed by External Methodology Assurance Panel, June 2020

Objective ID: A17754223

Table of Contents

1. Introduction

The Hard to Count Index is a list of planning areas (small geographic areas of 200-400 households) ranked by how difficult it is expected to be to enumerate the residents of that area via Scotland's Census 2022. Demographic factors associated with low levels of response to the 2011 Census and other more recent surveys were identified. Those factors were then used to rank planning areas in the index. The index is then split into five categories, and the category of a planning area is used to estimate the percentage of households in that area who will respond at different stages of the census process. The Hard to Count index is one of the drivers used to determine where to allocate census resources to boost response rates. For example, it is used to estimate how many households in different areas will need to be visited by field workers as part of non-response follow up.

2. Summary

This report describes the development of the Hard to Count (HtC) Index for Scotland's Census 2022. It introduces the concept of a HtC Index, and explains its importance in supporting the census. The analytical approach and statistical methodology used to develop the HtC Index are described. These are placed in the context of previous methodologies which have been used to develop previous HtC Indexes.

A group of four demographic measures (in statistical terminology, 'predictor variables') were chosen through analysis of several options, and were recommended to form the underlying data for the HtC index for the National Records of Scotland (NRS) Census Rehearsal in 2019, and for the census in 2022. The results of testing the HtC index as a predictor of return rates in the NRS Census Rehearsal in 2019 validated the HtC index in its current form.

3. The 2011 Hard to Count Index

The HtC Index for Scotland's Census 2011 was a categorisation developed to inform the Data Collection operation. The categorisation identified geographical areas which were expected to be difficult to enumerate (in other words to contain a relatively high proportion of non-responding households).

The 2011 HtC Index was also used to develop strata for the 2011 Census Coverage Survey (CCS). This CCS informed the 2011 Census Estimation and Adjustment process, through which individuals representative of non-responding households were imputed (either by placing them into existing households or placeholder addresses, or through the creation of new households).

Development of the 2011 HtC index applied a rank to each of the 6,505 Data Zones[1] in Scotland. The Data Zone with the lowest rank (1) was predicted to be the most

[1] Data Zones are the core geography for dissemination of results from Scottish Neighbourhood Statistics (SNS). The Data Zone geography covers the whole of Scotland and nest within local authority boundaries (as they were in 2011). Data Zones are groups of 2011 Census output areas

difficult to enumerate, while the Data Zone with the highest rank (6,505) was predicted to be the easiest. The ranking assigned to each Data Zone was based on demographic variables available at the time of planning for the census. These variables were converted to standardised scores, which were then combined additively with equal weightings to give a single HtC 'score' for each Data Zone. The Data Zones were ranked based on their HtC scores and placed into five categories (HtC 1-5; Table 1) based on their relative magnitude.

Table 1: 2011 HtC Index[2] categories

HtC 1 (easiest)	HtC 2	HtC 3	HtC 4	HtC 5 (hardest)
2,602 DZs (40%)	2,602 DZs (40%)	651 DZs (10%)	520 DZs (8%)	130 DZs (2%)

The decision on which variables to include in the 2011 HtC Index was based in part on evidence and insights gained from the 2009 Census Rehearsal prior to the 2011 Census. A five-part index was chosen in order to be sufficiently simplified for Census operations purposes, as different follow up resource was allocated to different HtC groups based on predicted response levels. The hardest to count 2% (HtC 5) was a small segment of the population expected to require larger investment of resource to prompt returns.

The use of a 40:40:10:8:2 split for the five hard to count categories in 2011 was partly designed to be consistent with the 2001 three category split (which was 40:40:20). It was also consistent with observed non-response rates in Scotland's Census in 2001 and 2011.

which have populations of around 500 to 1,000 residents. The 2011 HtC Index was based on Data Zones. Planning Areas (PAs) will be the basic geographical unit used in construction of the 2019 (Census Rehearsal) and 2022 (Census) HtC Indexes.

[2] The term 'HtC Index' refers both to the HtC category (1-5) assigned to a Data Zone and to the ranked list of all categorised Data Zones in 2011.

4. Comparison of Scotland, England and Wales approaches

The Office for National Statistics (ONS) has produced a Hard to Count Index for the 2021 Census in England and Wales. That Hard to Count index combines factors linked to low engagement with the census ("willingness") with a digital exclusion component. Details of the index have been published[3].

The NRS Hard to Count index is based on demographic factors that have been associated with lower response rate in previous, paper-based censuses. Scotland's Census 2022 will be "digital first", with collection methods aimed at garnering a high level of online responses. A separate Digital Exclusion Index has been developed to capture areas where there may be low response due specifically to difficulties with using an online census questionnaire. NRS has separated Hard to Count and Digital Exclusion into two indices because there are some interventions under consideration that may specifically target digitally excluded populations who may have relatively high engagement with the census, such as people aged over 75. Conversely, there may be some groups who are highly digitally literate but have in the past had low engagement with the census, such as university students. Separating the indices may allow us to better target interventions to these groups.

5. 2022 Hard to Count Index: Statistical analysis of demographic variables

This section describes the method used to identify variables for inclusion in the HtC Index for Scotland's Census 2022. A statistical model is used to test the ability of Data Zone level variables to predict the[4] household census return rates observed during Scotland's Census 2011. Both the significance and explanatory power of variables are explored.

5.1. Measuring Census Non-Response

In planning for Scotland's Census 2022, there is a need to carry out a formal statistical analysis of variables to identify whether they are associated with household non-response. A methodology for doing this is described in ONS's paper, "Predicting patterns of household non-response in the 2011 Census" (Hopper, 2011). Their study used a logistic regression model to test the association between area-level demographic variables and the household imputation rate (as a proxy for under-coverage) in the 2001 census in England and Wales. The probability of a household not responding to the census was modelled as the dependent variable, and the model fit was assessed by comparing the observed and predicted rates of non-response for each Lower Layer Super Output Area. In particular, in order to

[3] Statistical design for Census 2021, England and Wales - Office for National Statistics https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP102-Hard-to-Count-index-for-the-2021-Census.docx

[4] The raw household census return rates are the household return rates observed prior to Estimation and Adjustment.

capture as much societal change as possible, the model considered predictor variables which were able to be updated at least yearly[5] (Hopper, 2011).

Here we describe the model used to test the association between area-level demographic variables and household non-response to Scotland's Census 2011. An ordinary least-squares (OLS) multiple linear regression model was used for the analysis. Other modelling approaches, such as logistic regression, were considered but discounted (see Section 5.3 for more information). The dependent variable is the household return rate, calculated for each Data Zone in Scotland. The household return rate is defined by equation (1) and is bounded in the range [0, 1]:

$$Household\ return\ rate = \frac{(D + C)}{(A - B + C)} \quad (1).$$

Here:

A = the number of households on the pre-Census Address List
B = the number of households on the pre-Census Address List with no response expected[6]
C = the number of Found in Field household returns
D = the number of household returns matched to the pre-Census Address List

Full descriptions of the model and demographic variables are provided in Appendices A and B.

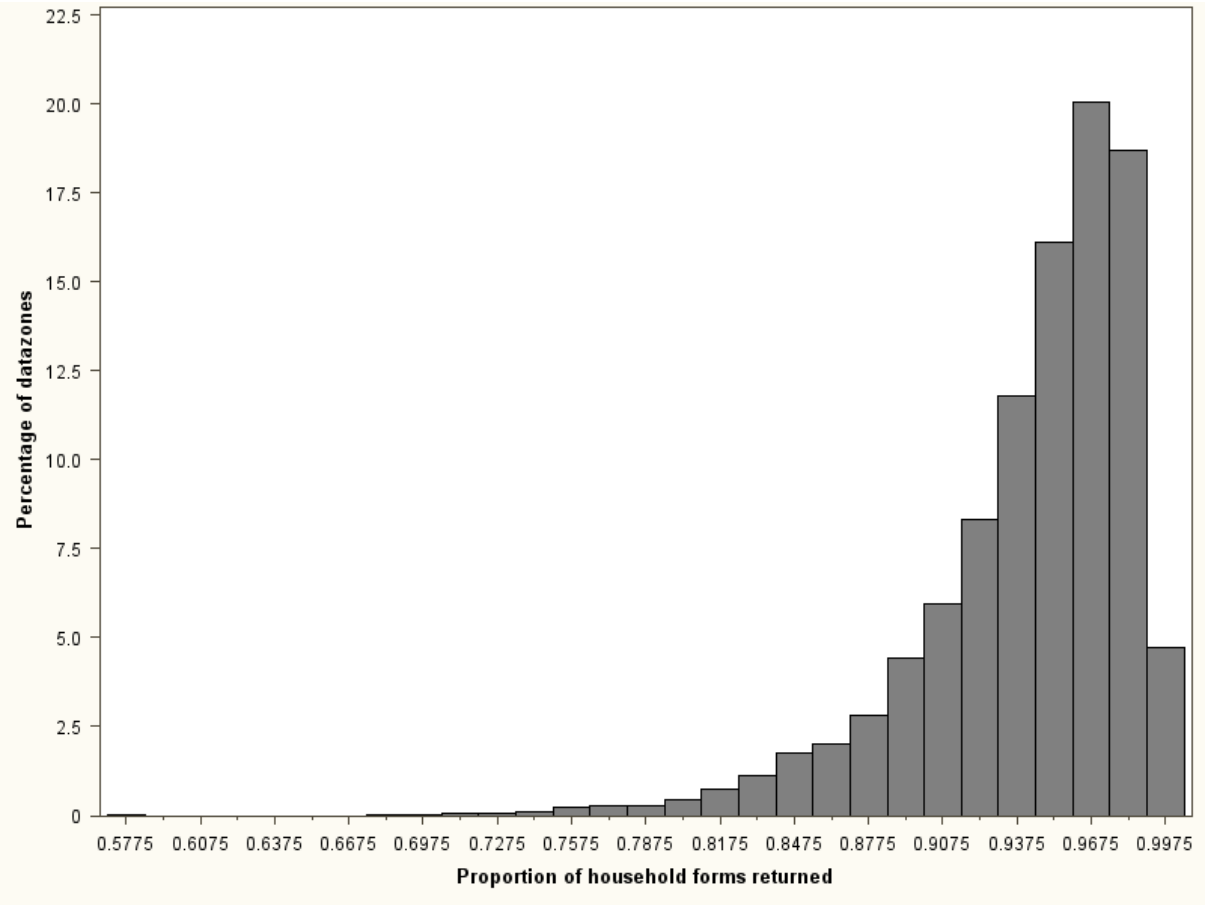
The inclusion of Found in Field[7] addresses in equation (1) accounts for addresses which were not included on the pre-census Address List but which returned a valid census questionnaire. Equation (1) also subtracts the number of addresses from which no return is expected from the denominator[8].

Figure 1 shows the distribution of household return rates (1), across the Data Zones in Scotland. Each of the column bars has a width of 0.015 as measured on the horizontal scale; the vertical scale gives the percentage of Data Zones. For example, the bar centred at 0.9075 has a width of 0.015 and a height of approximately 6.0. This means that approximately 6% of Data Zones had a household return rate between 0.9 and 0.915.

[5] The same principle informed the selection of variables comprising the HtC Index for Scotland's Census 2011.
[6] 'No response expected' refers to specific categories recorded on a placeholder form; e.g., the property is derelict, or a second/ holiday home.
[7] Found in Field addresses found by field workers during address check work and the non-response follow up period, which we were not on the pre-Census address list but which returned a valid questionnaire by the end of the Census period.
[8] If an enumerator classified an address as 'no response expected' but a valid household return was received, this was counted towards the total household returns.

The distribution of household return rates is negatively skewed, with a maximum of 1 and a minimum of 0.58 (Figure 1). The average household return rate is 0.94 (*result not shown on graph*).

Figure 1: Frequency distribution of household return rate in Scotland's Census 2011 by percentage of Data Zones



5.2. Statistical Methodology and Variables

A regression model is fitted to the data, described in Appendix A. The expected household return rate is fitted as the response (dependent) variable.

The predictor variables[9] tested for a significant association with household return rate are shown in table 2.

[9] The name of the variable as included in the model is given in brackets. An apostrophe following the variable name (e.g., pr_16to29') indicates that a transformation is applied to the variable. The variables and data sources are described in Appendix B in Table B1. Transformations are listed in Table B2.

Table 2: Predictor variables tested for association with return rate

Variable	Name of variable in model
The proportion of the population aged 16-29 years	(pr_16to29')
The proportion of the population who are full time students and not living either in a communal establishment or parental/carer's home during term-time	(pr_Student')
Scottish Index of Multiple Deprivation score	(SIMD_sr')
The proportion of occupied dwellings which are privately rented	(pr_Rents)
The proportion of dwellings classed as flats	(pr_Flats)
A binary variable (Urb_Rur) where '1' corresponds to a rural area	(Urb_Rur)
A binary variable (EAL) where '1' corresponds to a high proportion of pupils recorded as having English as an additional home language (an adjustment was applied to the data to account for pupils being taught in Gaelic; see Appendix B for details).	(EAL)

There are high quality national datasets with recent data available at a small geographic area level for all of the above variables. Additional variables including the proportion of properties on the electoral register, local turnout in general elections, employment status, long-term illness and disability were considered but have various issues. Some are subject to rapid change at low geographic area level, others have poor quality data, do not have any potential link to low census response, do not have reliable, consistent, or regular Scotland-wide datasets at a low geographic area level, or do not have such a dataset that is accessible for our use due to confidentiality concerns. Other variables considered such as level of formal education and the proportion of the population in older age groups are used in the separate Digital Exclusion Index development work to predict digital exclusion-related non-response, rather than non-response related to low engagement with the Census, as discussed in Section 4 above.

Note 1:
The model includes as a variable the proportion of *dwellings* classed as flats, rather than the proportion of *occupied dwellings* classed as flats. Estimates of the proportion of occupied dwellings within each Data Zone are available. However, this proportion is calculated across all dwellings and is not partitioned based on type of dwelling, so the data necessary to use *occupied flats* as a variable is not available.

Note 2:
The analysis described below assumes that the structural relationship between predictor variables and household return rate remains relatively constant over time. Hence, if a Data Zone level variable is associated with household return rate for the 2011 Census, a similar relationship will exist at the time of Scotland's Census 2022. One reason to question this assumption is the 'Digital First' requirement for Scotland's

Census 2022; and more generally, the implications of digital technology as relates to the census. Digital exclusion is addressed in an additional and separate index to the Hard to Count Index.

For example, it may be hypothesised that geographical areas containing a high proportion of students or young adults may show a higher household return rate in Scotland's Census 2022 compared with the 2011 census, due to the convenience of responding online for such groups. In contrast, areas with a demographic profile corresponding to relatively low internet access or usage, or with inadequate broadband provision, may show a drop in household return rate compared with 2011. The degree to which the latter effect predominates may depend on the public assistance measures put in place to facilitate and encourage participation in the online census. The potential geographic distribution of this issue is analysed in the Digital Exclusion Index.

5.3. Statistical Assumptions

The multiple regression approach assumes that there is a linear relationship between household return rate (the response variable) and each of the *n* predictor variables; when the values of the other (*n-1*) predictor variables are held constant. It is further assumed that the residuals[10] are normally distributed and exhibit constant variability when plotted against the fitted values (a condition known as homoscedasticity). Log transformation of the household return rate and some of the predictor variables is carried out to meet these assumptions (detailed in Table B2). With an average household return rate of 0.94, inspection of the residuals indicated some problems with the model predicting values outside the valid range (0, 1) for the return rate. However, OLS regression seemed the best choice given the other limitations. The other models which were initially considered and then discounted are now described.

The alternative models considered in response to external methodology assurance panel feedback included a beta regression model, which we used for confirmatory analysis in section 9.3 of this paper. In line with the ONS 2011 methodology paper (described in Section 5.1), we also considered a binary logistic regression model, modelled at the household level. This model allowed a household's response or non-response to the census to be taken as the dependent variable (e.g., 0 = 'return received', 1 = 'no return received'). Due to considerations of data confidentiality, the demographic variable of interest are only available at Data Zone level, and therefore the area-level value has to be assigned to each household in the Data Zone. This creates a spatial misalignment problem, which may introduce errors. Furthermore, when we tried to fit such a model to the data, the Hosmer-Lemeshow statistic indicated that the model was not a good fit ($p < 0.05$).

We note that a degree of spatial autocorrelation is seen in the underlying variables for the OLS regression model, but this reflects area characteristics (for example, flats

[10] The difference between the observed value of the response variable and the value predicted by the model; i.e., Residual = Observed value - Predicted value.

are likely to be near other flats). The underlying variables are all known at a low geographic level, and spatial autocorrelation reflects their distribution in the real world and is not a clustering artefact. Spatial autocorrelation was tested for in rehearsal data and a low level of autocorrelation was found. We are not intending to work to eliminate this because a) the index in its current form has good predictive power, b) we are attempting to predict an area-level characteristic rather than an individual one, and c) it would require additional resource for no clear gain.

6. 2022 Hard to Count Index: Model selection and fit

6.1. Linear regression with one variable

Firstly, each of the predictor variables is fit individually[11] to the data to gauge whether there exists a significant association with household return rate. All of the predictor variables are found to be highly significant ($p < 10^{-4}$).

The large number of data points may make statistical significance easier to achieve[12]. Particular attention is given to the R^2 values (Appendix C, Table C1). R^2 is the coefficient of determination: it measures the proportion of variation in the response variable explained by the values of the predictor variables.

6.2. Linear regression with multiple variables

The multiple regression model fits all of the predictor variables to the data together. This is done using a stepwise method (implemented in SAS Enterprise Guide, Version 5.1). The stepwise method sequentially adds and then removes individual predictor variables so as to maximise the adjusted R^2 statistic[13].

An alternative method (elastic net) was later proposed by external methodology assurance panel reviewers that would potentially improve the selection of model parameters. However, redoing this analysis would duplicate previous efforts and variable changes could not be tested in the field. The index produced using the variables selected below has been tested in the NRS census rehearsal 2019 (detailed in section 9 below) and found to be predictive of response rates and therefore suitable for use in Scotland's Census 2022. As any subsequent changes in variables could not be tested in a census rehearsal, the variables originally chosen using the stepwise method will be kept.

The maximum value of the adjusted R^2 (0.61) is achieved by including all seven predictor variables, each of which have highly significant p-values ($p < 10^{-4}$). However, from the list of models generated, it is apparent that models including three or four variables explain the data nearly as accurately as the seven-variable model. As restricting the number of variables to four does not appear to impact accuracy and including fewer variables decreases the work needed to maintain and update the HtC index, subsequent analysis is restricted to models which have a maximum of four predictor variables. In addition, the variables UrbRur and pr_Student' are removed from the model due to their low values of R^2 in the single-variable regression (Table C1).

Within this limited variable set, there are a number of plausible models of similar explanatory power. The best-fitting models as measured by the adjusted R^2 are shown in Table 3.

[11] I.e., excluding the other predictor variables.
[12] 6,500 Data Zones are included in the analysis.
[13] The adjusted R^2 statistic is equivalent to R^2 with a penalty applied due to including more parameters in the model.

Table 3: Proposed predictor variables and coefficients of determination

Model	Variables (standardised β) as listed in Table 2	Adjusted R^2 statistic
1	pr_16to29', pr_Rents, pr_Flats, SIMD_score	0.61
2	pr_Rents, pr_Flats, SIMD_score, EAL	0.60
3	pr_Rents, pr_Flats, SIMD_score	0.60
4	pr_16to29', pr_Flats, SIMD_score, EAL	0.58
5	pr_16to29', pr_Flats, SIMD_score	0.58
6	pr_16to29', pr_Rents, SIMD_score, EAL	0.57
7	pr_Flats, SIMD_score, EAL	0.56

6.3. Correlation analysis

The highest correlations between any of the variables included within models 1-7 (Table 2) are between pr_16to29' and pr_Flats (r = 0.56) and between pr_Flats and EAL (r = 0.50). These correlations do not seem high enough to warrant removing any variables from the model. This conclusion is supported by checking the Variance Inflation Factors and Condition Indices[14] for models 1-7. The largest variance inflation factor observed is 2.0 for the pr_Flats variable (model 2), and the largest condition index is 23.2 (model 1). Montgomery (2001) argues that variance inflation factors should be less than 5 and the condition index less than 100, to avoid serious problems with multicollinearity. As the values quoted above for models 1-7 fall within these bounds, it is concluded that the fitted predictor variables are not seriously affected by multicollinearity.

The correlations between the individual predictor variables are shown in Table 4.

Table 4: Values of Pearson's correlation coefficient

	pr_16to29'	pr_Rents	pr_Flats	pr_Student	SIMD_score	Urban_Rural	EAL
pr_16to29'	1						
pr_Rents	0.41	1					
pr_Flats	0.56	0.41	1				
pr_Student	0.58	0.57	0.48	1			
SIMD_score	0.31	-0.13	0.43	-0.16	1		
Urban_Rural	-0.36	0.08	-0.36	-0.20	-0.16	1	
EAL	0.39	0.33	0.50	0.44	0.11	-0.21	1

7. Recommendation of variables

The explanatory power of the models listed in Table 3 is very similar regardless of which variables are fitted. It is difficult to make a definitive judgement about which set of variables provide a better predictor of household return rate.

However, an additional factor in ensuring that future iterations of the index remain predictive is ensuring that up-to-date predictor variable information is available. The

[14] Variance Inflation Factors and Condition Indices were checked in SAS Enterprise Guide, Version 5.1, using the COLLINT and VIF commands.

distribution of predictor variables will change over time; for example as more flats are built in an area, the proportion of flats will increase. It is therefore important that chosen variables come from regularly updated datasets with data relevant to 2022.

The following data sets are produced yearly: proportion of 16-29 year olds, proportion of flats, and proportion of pupils for whom English is an additional home language. The Scottish Index of Multiple Deprivation is also produced regularly (every four years).

It is therefore straightforward to produce HtC scores from these four variables and rank scored planning areas to form a HtC Index which reflects the current demographic picture across Scotland. To do this we drop the Urban/rural binary variable, the proportion of privately rented flats and the proportion of students. Using these variables allows us to produce a model that uses up to date values of demographic variables that were found to be correlated with 2011 response rates. Model 4 in Table 3 captures these variables. This was recommended as the basis of the Hard to Count Index.

8. Method of producing the 2022 Hard to Count Index

The method used to derive the HtC index for Scotland's Census 2022 was adapted from work done by ONS prior to the 2011 Census in England and Wales to provide an "Enumeration Targeting Categorisation". The following procedure describes how the HtC Index is calculated for the 2022 census. All data sets were at the Data Zone level and had to be transformed to Planning Area[15] geographic data groups. Then each of 8,939 Planning Areas was ranked for each variable. To reduce the effect of extreme values each ranking was then scaled using the following formula:

23*LN(1-Rank/Number of ranks*(1-Exp(-100/23)))

To produce the HtC score for each planning area the scaled rankings for each of the four variables were added up without weighting. The Planning Areas were then sorted by these scores to make a ranked list.

The segmentation for the HtC index was applied to this ranked list. This is the division of the ranked list into 5 unequal segments. Segmentation was decided based on 2011 census HtC divisions and based on operational needs for targeting interventions.

- The easiest to count 40% make up the section called HtC1,
- The next 40% make up HtC2,
- The next 10% make up HtC3
- The next 8% make up HtC4
- The hardest to count 2% make up HtC5.

This segmentation was applied in order to be able to apply broad assumptions on response rates for modelling, and also to allow interventions to be directed at groups

[15] Planning Areas are small geographic areas of 200-400 households (see Glossary)

of areas. For example, HtC5 areas are expected to have a lower rate of response to initial contact, and so additional interventions to boost response rate are expected to be needed in these areas. The segmentation of the 2022 HtC index is the same as the segmentation of the 2011 HtC index, described in Table 1 in section 3 above.

The most recent refresh of the HtC index was carried out in 2020, when new Scottish Index of Multiple Deprivation (SIMD) data was released. Because the census date has moved to 2022, there may be scope to refresh the index in 2021 if more up-to-date source datasets are available, can be accessed and updating does not disrupt other census processes. Below in Table 5 are the datasets used, the year the dataset was published and what was used from the dataset.

Table 5: Datasets used in calculating the HtC Index

Dataset	Data applying to year	Description of area variable used
Small Area Population Estimate	2018 (published 2019)	Proportion of 16-29 year olds in population
Scottish Index of Multiple Deprivation	2020 (published 2020)	SIMD score
School Pupil Census	2017 (supplied 2018)	Proportion of pupils for whom English is an Additional Language (EAL)
Small Area Statistics on Households and Dwellings	2017 (published 2018)	Proportion of flats (flat categorisation is absent from 2018 and 2019 datasets)

9. Further analysis of Hard to Count Index

9.1. Validation of Hard to Count Index in 2019 Census Rehearsal

National Records of Scotland carried out a major public rehearsal exercise that tested some of the systems and services that will be used in the 2022 census.

The rehearsal took place between 7 October and 7 November 2019, using a reference day of 13 October, and was conducted in three local authority areas; namely parts of Glasgow City, Dumfries and Galloway and Na h-Eileanan Siar. These areas were selected to allow National Records of Scotland to test approaches in rural, urban and diverse communities. Participation in the rehearsal was on a voluntary basis. Over 72,000 households were contacted with nearly 18,500 responding. More detail on the rehearsal is available at https://www.scotlandscensus.gov.uk/news/census-rehearsal-evaluation-report-published

The HtC index was tested using the results from the 2019 census Rehearsal. This analysis indicated that both the 2022 HtC index and the planned segmentation of that index are supported by the geographic distribution of return rates.

9.1.1. Rehearsal Data

The paper and online return rates within each planning area were used in conjunction with planning area HtC scores to investigate and validate the HtC index.

Completing the rehearsal survey was voluntary, while filling out the census is not. Therefore rehearsal data does not predict return rates in the actual census, as voluntary respondents will be a smaller and demographically skewed sample of census respondents. We are assuming that patterns of response in rehearsal will broadly mirror those in the census as this has been the case in previous rehearsals and census years, even though return rates are very different. However, the relative levels of return seen in different areas can be used to evaluate the rankings in the HtC index, as even in a voluntary survey we would broadly expect a higher level of response from areas lower on the HtC index.

9.1.2. General Results

Table 6 compares the HtC scores and the HtC Index values between the rehearsal dataset and the Scotland-wide HtC index dataset prepared for Scotland's Census 2022. The table shows that the mean HtC scores and HtC index values are considerably different between the rehearsal dataset and the main census dataset, meaning that the rehearsal areas were on average harder to count than the average of all areas across Scotland. These areas were chosen using the HtC index prepared for census 2022 to test census processes and so a high proportion of harder to count areas were included by design. This is also borne out by t-tests which are significant for both variables: scores ($t = 8.0131$, $p < 0{,}01$) and index categories ($t = 8.3291$, $p < 0.01$). This means that any conclusions drawn from the rehearsal dataset about the HtC Index may not directly apply to the main census, and should be considered indicative only.

Table 6: Aspects of census engagement in census and rehearsal datasets

	All planning areas	**Rehearsal planning areas**
Mean score HtC	86.75	140.96
Mean index value HtC	1.92	2.80
Median index value HtC	2	3
Standard Deviation of HtC score	61.90	96.64

The key variables analysed are listed below.

- The *Return Rate* is the proportion of households that returned a completed questionnaire, either online or on paper.

- The *Proportion of Online Returns* is the proportion of returned rehearsal questionnaires that were completed online.[16] This is also a factor in the examination of the Digital Exclusion Index, which was evaluated separately.
- The *Online Return Rate* is the proportion of households that completed the questionnaire online. This is also a factor in the examination of the Digital Exclusion Index, which was evaluated separately.
- The *Paper Request Rate* is the proportion of households that requested a paper version of the census questionnaire. Many but not all paper questionnaires requested were then returned as completed paper questionnaires.
- The *Paper Return Rate* is the proportion of requested paper questionnaires that were returned.

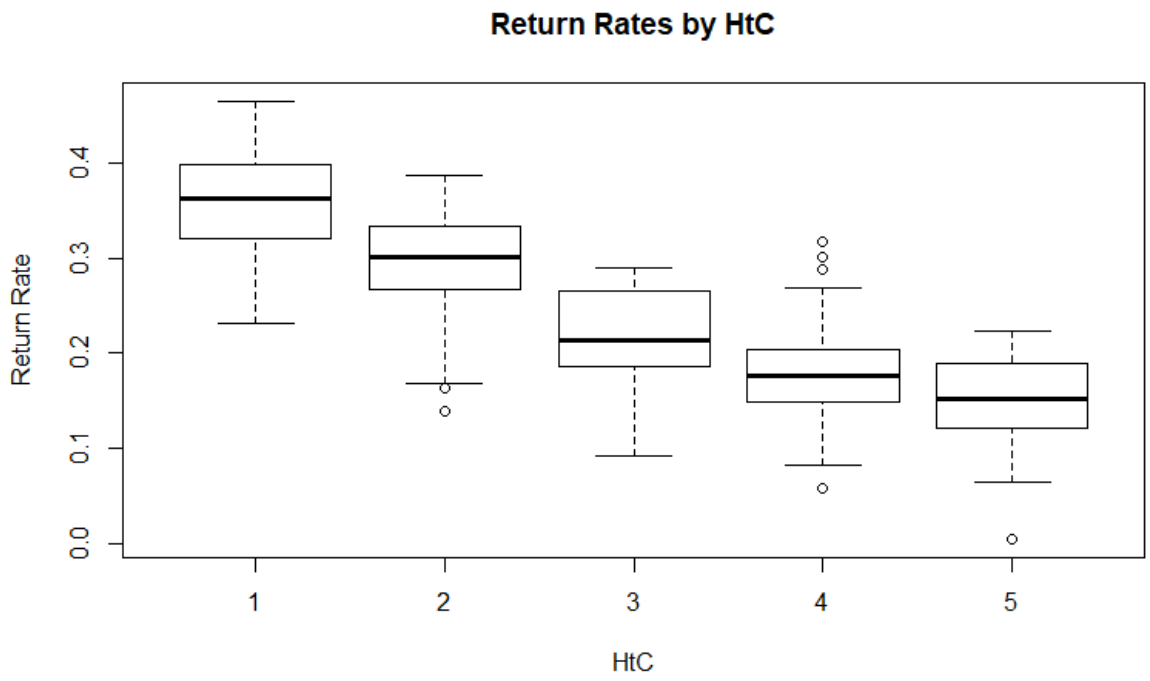
Table 7 shows the correlations between different types of rehearsal return rates and HtC scores. Figure 2 shows the relationship between HtC Index group (1-5) and return rates in rehearsal planning areas.

Table 7: Correlation between HtC Index group and return rates

	Return Rate	Online Return Rate	Paper Request Rate	Proportion of Online Returns	Paper Return Rate
HtC scores	-0.878***	-0.864***	-0.536***	0.061	-0.552***

[16] A note on the difference between Proportion of Online Returns vs Online Return Rate: If 100 households returned 25 census questionnaires of which 15 were online and 10 paper, the proportion of online returns would be 60% (15 of 25 returns) while the Online Return Rate would be 15% (15 of 100 households returned questionnaires online).

Figure 2: Planning area level return rates by Hard to Count index groups in 2019 rehearsal



The Return Rate is highly negatively correlated with HtC scores, meaning that the higher the HtC score of a planning area is, the lower the return rate of that area is. This indicates that the HtC score is a good representation of the relative levels of response to the rehearsal seen in different planning areas, and thus that HtC index rating is a reasonable predictor of relative return rate.

The Online Return Rate, and to a lesser extent, the Paper Request rate, are also negatively correlated with HtC score. This also indicates that HtC index rating is a reasonable predictor of return rate.

Making a paper request requires engagement with the census process, and so it makes sense that people in easier to count areas were more likely than people in harder to count areas to request paper questionnaires when they chose not to complete the questionnaire online.

The Proportion of Online Returns is not correlated with the HtC scores which suggests that HtC score does not influence the proportion of people in an area who choose to complete the questionnaire online or on paper. As this is a factor more linked to the respondent's use of digital services than to their engagement with the census, it is logical that it is not captured in the HtC score. It is instead captured in the Digital Exclusion Index. The interventions targeted to digitally excluded areas may be different to the interventions targeted to HtC areas, so differentiating between the two may be useful.

9.1.3. Evaluating the Hard to Count Index segmentation

Segmentation of the HtC index is important as the categories are used to inform census planning and resource allocation, rather than the underlying scores. To create the categories the planning areas were sorted by their HtC scores in ascending order and the divided into five categories with the following pattern: 40%, 40%, 10%, 8%, 2%. To explore whether the segmentation is an appropriate representation of the HtC index to use in allocating follow up resources, a correlation analysis was carried out as seen in Table 8 below.

Table 8: Correlation between HtC raw scores, HtC index group and return rates

	HtC_Score	HtC_Index
Overall_Return_Rate	-0.878	-0.855
Online_Return_Rate	-0.864	-0.838
Paper_Request_Rate	-0.536	-0.532
Proportion_Of_Online_Returns	0.0675	0.0761
Paper_Return_Rate	-0.552	-0.546
HtC_Score	1.000	0.988
HtC_Index	0.988	1.000

The correlation coefficient for the index values and the scores with the observed return rates are similar. This indicates that the segmentation is a good representation of the underlying structure of the HtC scores data. This means that the segmentation currently applied to simplify the index does not greatly reduce the correlation of the index with observed return rates, compared to using the scores. The segmentation applied to allow the index to be used for planning purposes appears to be appropriate. This indicates that it will also be appropriate for the larger Scotland-wide dataset, with the caveats mentioned above.

The linear relationship between the index and the overall return rate was also tested using ANOVA. Results are shown in table 9.

Table 9: Correlation between HtC index and return rate

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
HtC Index	1	1.189	1.819	633.7	<2e-16 ***
Residuals	234	0.67	0.003		

9.2. Intercorrelation of variables: further testing

To further ensure that the four variables underlying the 2022 HtC Index were not overly intercorrelated, both Pearson's and Spearman's rho were calculated, to compare with the Pearson's coefficient calculations in Table 4 (section 6.3). The results can be found in Tables 10 and 11. It should be noted that both correlation tests were conducted on the scaled variables, rather than the proportions as in Table 3. This is due to the fact that the unscaled SIMD variable was not available. The two

tests compared indicate that the intercorrelation of the four predictor variables used is not large and should not affect modelling.

Table 10: Values of Pearson's rho correlation coefficient

	Age	SIMD	Flats	EAL
Age	1			
SIMD	0.250	1		
Flats	0.697	0.295	1	
EAL	0.634	0.194	0.691	1

Table 11: Values of Spearman's rho correlation coefficient

	Age	SIMD	Flats	EAL
Age	1			
SIMD	0.374	1		
Flats	0.580	0.362	1	
EAL	0.544	0.178	0.573	1

9.3. Beta Regression confirmatory analysis

In response to external methodology assurance panel feedback, to further validate the variables selected by linear regression to underlie the Hard to count Index, beta regression analysis was carried out on the four variables selected in section 7 using the same datasets as the original linear regression. Results are shown in Table 11 below.

The analysis indicates that all four variables are still significantly associated with low household return rate, with significance below the p=0.001 threshold. English as additional language is used as a categorical variable (high versus low proportion of pupils) as was done in the linear regression analysis in section 6 above.

The negative numbers in the Estimate column mean that higher values of these variables are associated with lower census household return rates.

Table 11: Beta regression analysis of Hard to Count variables

Variable	Estimate	Z-value	P
Proportion of 16 to 29 year olds	-2.02	-21.83	<0.001
Proportion of flats	-0.88	-33.10	<0.001
Pupils with English as additional language	-0.10	-6.60	<0.001
Scottish Index of Multiple Deprivations score	-0.01	-37.62	<0.001

10. References

Durrant, G. B. and Steel, F. (2009) "Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK government surveys". J. R. Statist. Soc. A, 172, 2, pp. 1-21.

Foster, K. (1998) "Evaluating non-response on household surveys: Report of a study linked to the 1991 Census". GSS Methodology Series, No. 8. ISBN: 1857742729

Freeth, S. and Sparks, J. (2003). "Summary of the 2001 Census-linked studies of survey non-response". Unpublished, internal ONS document.

Hopper, N. (2011) "Predicting patterns of household non-response in the 2011 Census". Survey Methodology Bulletin No. 69 - September 2011, pp. 9-22. http://www.ons.gov.uk/ons/guide-method/method-quality/survey-methodology-bulletin/smb-69/index.html

Montgomery, D. C. and Vining, G. G. (2001) "Introduction to linear regression analysis, 3rd edition, Wiley, New York.

Rahman, N. and Goldring, S. (2006) "Factors associated with household non-response in the Census 2001". Survey Methodology Bulletin, 59, pp. 11-24.

11. Appendix A: Data Zone level model of household return rate

A linear regression model is fitted to the data, which has the form:

$$E(Y_i) = \alpha + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_k X_{i,j} \quad (2)$$

Here Y_i is the Data Zone level household return rate, which lies between 0 and 1. The model predicts the expected household return rate, $E(Y_i)$, by fitting the intercept (α) and j predictor variables. Here $X_{i,j}$ denotes the value of the j-th variable for the i-th Data Zone; the β_j are parameters which measure the effect of the j-th variable on the expected household return rate $E(Y_i)$. Equation (A.1) assumes that the predictor variables combine linearly in determining the expected household return rate.

Each Data Zone is also associated with a random error ε_i. The errors are assumed to be independent normally-distributed random variables with mean 0 and constant variance σ^2. Spatial autocorrelation of variables has been considered (see section 5.3).

Model fit is assessed using the method of least squares estimation, by looking at the differences (residuals) between the observed values of the response variable Y_i and the values predicted by the model (A.1). This is done by calculating the variance ratio:

Mean Squares (Regression) / Mean Squares (Residual) (3)

Under the null hypothesis (H_0), the variance ratio has an F-distribution specified by the degrees of freedom for the regression (i.e., the fitted model) and the degrees of freedom on the residuals. Under H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$, and variance ratio ~ F (degrees of freedom (Regression), degrees of freedom (Residual)). From the F-statistic (A.2), a 'p-value' is calculated to test whether the null hypothesis can be accepted. Low p-values (< 0.05) suggest that there is evidence to reject the null hypothesis, and to conclude that the parameters (the β_j) do not all equal zero and that the associated predictor variables (the $X_{i,j}$) predict the expected household return rate.

12. Appendix B: Variables included in the Data Zone level model

The variables included in the model were selected on the basis that they provide a broad range of social and economic characteristics. Rahman & Goldring (2006) found that both household tenure and householder age affected the likelihood of participation in the 2001 Census Coverage Survey. Householder unemployment has been associated with non-response in the census and in social surveys, as has living in cities (Foster, 1998; Freeth and Sparks, 2003; Rahman and Goldring, 2006; Durant and Steele, 2009). Finally, Rahman and Goldring (2006) found that households with single occupants or which paid rent were twice as likely not to respond in the 2001 census compared with control households.

The model for the expected Data Zone level household return rate is:

$$E[(pr\_Return')_i] = \alpha + \beta_1(pr\_16to29')_i + \beta_2(pr\_Rents)_i + \beta_3(pr\_Flats)_i + \beta_4(pr\_Student')_i + \beta_5(SIMD\_sr')_i + \beta_6(Urb\_Rur)_i + \beta_7(EAL)_i \quad (4)$$

The variable 'pr_16to29' refers to the proportion of the population within the Data Zone who are aged 16-29 years. The variable 'pr_Rents' refers to the proportion of privately rented properties. The variable 'pr_Flats' refers to the proportion of properties within a Data Zone which are classed as flats (this statistic does not consider occupancy and therefore includes empty properties). The variable 'pr_Student' refers to the proportion of full-time students within the population of the Data Zone, with students assigned to Data Zones based on their term-time address[17]. The variable 'SIMD_sr' refers to the Scottish Index of Multiple Deprivation (SIMD) score for the Data Zone. SIMD score is based on the characteristics: income, employment, education, health, access to services, crime, and housing[18].

The above variables are all defined on a continuous scale. The model also includes two binary variables (equal to 0 or 1). The variable 'Urb_Rur' is derived directly from the Urban Rural 2-Fold Index. Urban Data Zones are coded as Urb_Rur = 0 and rural Data Zones are coded as Urb_Rur = 1.

The variable 'EAL' is derived from the field 'English as an Additional Language (EAL)' recorded in the annual School Pupil Census. Pupils who were recorded as 'New to English', 'Early Acquisition', 'Developing Competence' or 'Competent' are classed as having English as an additional home language, unless the pupil was also recorded as being taught in Gaelic, e.g. through Gaelic Medium Education. On this basis, the proportion of pupils for whom English is an additional home language is estimated[19].

Table B1 gives the full list of variables included in the model.

Table B1: Description of variables.

[17] Only full-time students who were not living in either a communal establishment (whether an institution-maintained property or private hall) or their parental/ carer's home during term-time are counted.

[18] A higher score indicates a relatively higher level of deprivation.

[19] The categories 'Limited Communication' and 'Not Assessed' were excluded from the calculation.

Variable	Description	Data and source(s)
Y_i	Household return rate for i-th Data Zone	-
$E(Y_i)$	Expected household return rate for i-th Data Zone	-
$X_{i,j}$	Value of j-th variable for i-th Data Zone	-
α	Intercept	-
β_j	Effect of j-th variable on expected household return rate	-
pr_Return	Proportion of household questionnaires returned	Scotland's Census 2011 (unpublished data); NRS
pr_16to29	Proportion of population aged 16-29	Mid-year population estimates (2010); NRS.
pr_Student	Proportion of population who are full time students (term-time address)	HESA (2010/11); Education Analytical Services, SG.
		Mid-year population estimates (2010); NRS.
SIMD_sr	Scottish Index of Multiple Deprivation (SIMD) score.	SIMD 2012; Office of the Chief Statistician, SG.
pr_Rents	Proportion of occupied dwellings being privately rented.	Private landlord register (extract July 2008); SG Housing.
		Small area dwelling estimates (2008); NRS.
pr_Flats	Proportion of dwellings which are classed as flats.	Household Estimates (2010); NRS.
Urb_Rur	Binary variable ('1' if rural, '0' if urban)	Urban Rural Index 2-Fold 2009/10; SG.
EAL	Binary variable ('1' if proportion of school pupils with English as an additional home language > 0.05; else = '0').	School Pupil Census (2010/11); Education Analytical Services, SG.

Table B2: Transformations applied to variables.

Variable	Transformed variable
pr_Return	pr_Return' = Log_e (1.01-pr_Return)
pr_16to29	pr_16to29' = Log_e (0.01+pr_16to29)
pr_Student	pr_Student' = Log_e (0.01+pr_Student)
SIMD_sr	SIMD_sr' = (SIMD_sr)^(1/2)
pr_Rents	pr_Rents
pr_Flats	pr_Flats
Urb_Rur	Urb_Rur
EAL	EAL

13. Appendix C: Single-variable Regression Model

Table C1: Fitted parameters, p-values, and coefficients of determination.

Variable	Standardised β	p-value	R^2 statistic
pr_16to29'	0.52	$<10^{-4}$	0.27
pr_Rents	0.34	$<10^{-4}$	0.11
pr_Flats	0.67	$<10^{-4}$	0.45
pr_Student'	0.31	$<10^{-4}$	0.09
SIMD_sr'	0.57	$<10^{-4}$	0.32
Urb_Rur	-0.18	$<10^{-4}$	0.03
EAL	0.37	$<10^{-4}$	0.14

14. Glossary of terms and acronyms

Acronym	Term	Definition
CCS	Census Coverage Survey	The Census Coverage Survey is a voluntary, independent, post-enumeration, representative, sample survey used during coverage adjustment to produce population estimates.
DZ	Data Zone	Data Zones are the core geography for dissemination of results from Scottish Neighbourhood Statistics (SNS). The Data Zone geography covers the whole of Scotland and nest within local authority boundaries (as they were in 2011). Data Zones are groups of 2011 Census output areas which have populations of around 500 to 1,000 residents.
EAL	English as an Additional Language	Used in the Scottish Government's Pupil Census to describe pupils who have a primary language other than English. https://www.gov.scot/publications/pupil-census-supplementary-statistics/
HtC index	Hard to Count index	An index indicating how willing households within a Planning Area will be to respond to the census. The Hard to Count Index is created from a ranked list of planning areas based on factors found to be associated with census non-response.
PA	Planning Area	Planning Areas (PAs) are geographic areas built from groups of postcodes and averaging between 200-400 residential addresses, although there are a small number outside that range. PAs nest within Local Authorities and will be used during the production of the Hard to Count Index and for the prioritisation of enumeration addresses to follow-up.
SHS	Scottish Household Survey	The Scottish Household Survey (SHS) is an annual survey of over 10,000 households. The survey covers a wide range of topics relating to the composition, characteristics, behaviour and attitudes of Scottish households and adults http://scottishhouseholdsurvey.com/
SIMD	Scottish Index of Multiple Deprivations	The Scottish Index of Multiple Deprivation is a relative measure of deprivation across 6,976 small areas (called Data Zones). If an area is identified as 'deprived', this can relate to people having a low income but it can also mean fewer

		resources or opportunities. SIMD looks at the extent to which an area is deprived across seven domains: income, employment, education, health, access to services, crime and housing. https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/
SNS	Scottish Neighbourhood Statistics	Scottish Neighbourhood Statistics is the Scottish Government's on-going programme to improve the availability, consistency and accessibility of small area statistics in Scotland. https://www.nrscotland.gov.uk/files/statistics/seminars/ftsyn-sns.pdf