**Maximising Response – Response Profiles**

**Victor Meirinhos**

Contents

[Contents 1](#_Toc10633580)

[1. Introduction 1](#_Toc10633581)

[2. Moving to an online-first census 2](#_Toc10633582)

[3. Context in terms of ONS research 3](#_Toc10633583)

[4. Exploratory analysis and statistical techniques 4](#_Toc10633584)

[5. Preliminary results 5](#_Toc10633585)

[6. Discussion 14](#_Toc10633586)

[7. Appendix 17](#_Toc10633587)

[8. References 21](#_Toc10633588)

# Introduction

Response profiles are a best estimate of the volume of responses we expect to receive during the census period. They will model when, where and how people are going to respond to the census, and they will be vital to inform and support operational decisions during the 2021 Census (Meyer, Mok, & Sullivan, 2015; Salganik, 2017).

Before the census, the response profiles will feed into the Field Operation Simulation Model (FOS) which will be used to help make decisions about the number of field staff required in each area, where to send reminder letters, and where to send paper questionnaires as the initial contact. During the live census period, the response profiles will be used by the Response Chasing Algorithm tool (RCA) as a basis for comparison against live return data, so that we can tell where we need to put in extra effort to meet our quality targets of 94% overall response and at least 80% response in every local authority.

Our aim is to build response profiles for returns by day, from the start of the census period until the end of the collection period. We aim to do this for each mode of response – online and paper. Our objective is that each lower super output area (LSOA) will have a response profile which is associated with its influential demographic variables and Hard-to-Count (HtC) rating. Groups of LSOAs that have similar response profile shapes and demographics, will be clustered together. This means that we can deal with the smallest possible number of response profiles, rather than 34,753 (the number of LSOAs).

We anticipate that a cumulative response profile might look like the graph in figure 1. The grey line shows the overall response. This is broken down to show the daily expected online (orange line) and paper (blue line) responses.



Figure 1. Cumulative response profile example

# Moving to an online-first census

A challenge in building the 2021 Census response profiles, is that in 2021 we are moving to an ‘online-first’ strategy. This means that online engagement will be actively sought and prioritised ahead of non-online modes. This objective makes the task of predicting responses for 2021 more difficult. Comparisons to the 2011 Census or other paper-based censuses are limited due to this fundamental change in approach. Therefore, a realistic approach to the development of the response profiles needs to consider not only the willingness of the public to respond to a census, but also their willingness to respond online.

An additional challenge to our ambitions comes from the fact that we know that sending a unique access code (UAC) and invitation to complete the census online, does have a negative impact on response when compared with sending a paper questionnaire. This was shown in the 2017 Test when the sample who were sent paper questionnaires had a response rate of 39% compared with the 26% response rate achieved by the comparative sample who were sent invitations to complete online.

However, of the households who were sent paper questionnaires first in the test, 30% of the responding households did respond online. Comparing this with the 16% online submission rate for the 2011 Census, we can see that people are becoming more willing to complete the census questionnaire online. Evidence from the 2017 Test also shows that people tend to respond in the mode first offered to them. We therefore do not want to provide too much paper in the first instance, but would rather predict where this is the most effective approach so that we can achieve the quality target of 75% online response.

In addition, we must bear in mind the need to control variance in response – keeping variation low whilst simultaneously seeking high response rates. A relatively high response rate is needed to prevent a substantial increase in variance when adjusting for non-response (Racinskij, 2015).

To achieve the 75% online response target, whilst pursuing the highest response rate and low variance, Operations will be using the response profiles as a source of reliable information about the underlying characteristics of the population for each geographical area (LSOA or LA) before and during the census collection period. Evidence from the 2016[[1]](#footnote-1) Test regarding the preferred mode of census completion by age distribution for online respondents suggests that there should be a focus on the age group 75 and over to achieve the desired target for this segment of the population (when asked about the preferred mode of response to census only 48% of the sample in this group mentioned online response as their choice to reply to the census).

By considering results from the 2017 Test, large scale surveys that have moved to an online approach, and online-first censuses in other countries, we aim to predict how much of the response will be online and how much on paper, so that we can split the response profiles into online and paper as in figure 1.

# Context in terms of ONS research

The response profiles are part of the Maximising Response Design product, and will be used in combination with other epics, such as the Field Operations Simulation model (FOS), the Hard-to-Count (HtC) index and the Response Chasing Algorithm tool (RCA). The FOS will be used to prepare for follow-up, for example, by simulating different scenarios to test the effect of follow-up strategy on response. The HtC Index will help us decide how to chase non-response at the start of follow-up. In comparison, the RCA is a decision support tool that will help the operational community make the best decisions about deploying interventions to maximise the return rate, helping ONS reach the key response targets during both the 2019 Census Rehearsal and the 2021 Census. It is expected that the response profiles will feed into the FOS and the RCA. They will provide the baseline returns during the census collection period (self-response) to the FOS, and will be used by the RCA as the reference data to flag shortages in response from all LSOAs in England and Wales.

# Exploratory analysis and statistical techniques

1. **Analysing the response shape** – Analysing and plotting the self-response shape using a clustering analysis. This analysis will consider LSOA response during the 5 weeks prior to the start of field follow-up in 2011. This analysis aims to link response behaviour to geography, under the assumption that LSOA classifications reflect the characteristics of the underlying population.
	1. Perform a **cluster analysis (two-step cluster)** to segment 2011 Census returns. The cluster analysis is used as an exploratory analysis to identify structures within the census self-response data. This technique is considered robust for clustering homogenous groups of cases when the grouping is not previously known. However, because it is exploratory, it does not make any distinction between dependent and independent variables and therefore we cannot infer any causal relationship between variables. Despite this limitation, clustering has been used in a large variety of fields and applications to produce data segmentation and explore associations between variables in large datasets (Dietrich et al., 2014, 2016; Sarstedt & Mooi, 2014; Tkaczynski, 2017).
	2. Use a **discriminant analysis** to determine whether the clusters are significantly different from each other. This technique is used to analyse complex data when the criterion or the dependent variable is categorical (cluster classification) and predictors or the independent variables (clustering variables) are continuous. The main advantage of this technique is that it examines whether significant differences exist among the groups, in terms of the predictor variables, evaluates the accuracy of the classification and reassigns observations to new clusters (Silverman, 2018).
2. **Model specification** - Build up a model using respondent’s characteristics (e.g. demographics) from the 2011 Census and the HtC for 2021 to assess the influential variables in the prediction of the cumulative response rate by day 10 after census day in 2011 (6 April 2011). This will be followed by the development of a longitudinal model using observations from 5 consecutive weeks. This stage of the analysis does not make any distinction between online and paper responses, considering only the overall self-response for each of the 34,753 LSOAs in England and Wales.
	1. Use a **multiple regression model** with the entire sample of 34,753 LSOAs. Predictor variables to be tested in the first model will be HtC rating, predominant age group of household reference person (HRP), predominant gender group of HRP, predominant household size, predominant qualifications of HRP. The outcome variable will be the response rate by day 10 after census day.
	2. Using a **random sample** of 3,488 LSOAs, perform a multiple regression with the same predictor and outcome variables to evaluate significant main and interaction effects.
	3. Conduct ***Post Hoc Tests* with Bonferroni corrections** to confirm where the differences occurred between groups. These follow-up tests are usually performed when an overall statistically significant difference in group means is found and allows control of the experiment wise error rate (usually alpha = 0.05).
	4. **Modelling continuous longitudinal data** - using the predictor variables from stage 1 (model specification) and weekly response rates (with 5 time points) for all LSOAs in England and Wales perform a Mixed-effects Regression Model. Assess the model fit and evaluate significant main and interaction effects.

# Preliminary results

**Cluster Analysis**

A cluster analysis (two step cluster) was performed using SPSS version 23 using the categorical variable HtC and the continuous variables: response in the week before, during and after the Census in 2011.

Results from this suggest the existence of 5 different response clusters (figure 2 and table 1).

Figure 2. 5 clusters derived from the two-step procedure using HtC as categorical variable and LSOA response during 3 consecutive weeks (before, during and after Census)



Figure 2. Clusters (5) derived from the two-step procedure using HtC as categorical variable and LSOA response during 3 consecutive weeks (before, during and after Census)



Table 1. Distribution of LSOAs for each cluster.

**Cluster solution**

From table 1 we can see that the 1st Cluster (entirely composed of LSOAs in HtC 1) includes more than 37% of the sample of all LSOAs in England and Wales (13,009). In terms of respondant behavior, this cluster is characterized by a significant peak in the Census weekend (nearly 51% of all self-response occurs in this period), followed by the week after census (around 12%) and the week before Census (around 8.25%). This very willing and engaged cohort appears to be mainly driven by the Census reference day.

The 2nd Cluster reveals a very different pattern. It represents a group of 1,791 LSOAs extracted from HtC 1 (6.4%), HtC 2 (5.6%), HtC 3 (2.8%) and HtC 4 (0.7%). The response shape for this cluster shows that the peak in census week fades (32.5%) when compared with cluster 1 and transfers response (nearly 30%) to the week after census.

The remaining clusters, number 3 (3,380 LSOAs), 4 (3,456 LSOAs) and 5 (13,117 LSOAs), show a very similar pattern during the week before and the week after Census with around 9% before and 12% after. The main difference between these clusters appears to be the amount of response reached in the Census week (see figure 2 for further information).



Table 2. Frequency table with the distribution of cluster by HtC.

Following the cluster analysis, a discriminant analysis was performed using the cluster classification as criterion. This analysis confirmed the cluster solution (with 5 different clusters) and the accuracy of the classification. This confirms the existence of significant differences between groups of LSOAs considering the variable HtC and the response during the two weeks around Census. We also found that other models, which included more predictor variables, did not converge within the recommended number of ten iterations using the cluster analysis.

**Multiple Regression**

A limitation of the cluster analysis is that it does not provide information about the possible effects from other variables. Therefore, a multiple regression (GLM) was used subsequently, with the aim to predict the variance in response due to the influence of established predictor variables. This procedure is appropriate at this stage, as it provides information about the overall fit (variance explained) of the model, the relative contribution of each of the predictors to the total variance explained, and produces estimates based on group distinctions.

The first model included the entire sample of LSOAs (34,753) from the 2011 Census. The outcome variable was the response rate at day 10 after census day (before the field force started – 6th April 2011) aiming to evaluate the impact of the predictor variables only. This selection strategy was employed to remove the possible effect from confounding variables associated with the maximizing response interventions used during the 2011 Census collection period.

Using several demographic variables from the 2011 Census and the new HtC classification for the 2021 Census, a regression model was developed. Equation (1) summarizes the proposed regression model:

 *Yi = α + β1Xi,1 + · · · + βpXi,p + Єi  (1)*

Equation (1) contains a continuous dependent variable representing the response rate by day 10 after Census in 2011 for each LSOA in England and Wales. The predictor variables in the model include the following:

* HtC 2021
* predominant age group HRP
* predominant gender group HRP
* predominant household size
* predominant qualifications for HRP

Interaction effects between predictor variables were also considered in the model as presented in equation (2).

*Yi = β0 + X1 β1 + X2 β2 + X3 β3 + X4 β4 + X5 β5 + X1X2 β5 + X1X3 β6+ … X1X2X3X4X5βi + Єi  (2)*

Formula description with main factors, relevant coefficients and examples of two interactions is presented next.

*Yi* = response rate by day 10 after Census in 2011

*β0* =Intercept

*X1* =HtC 2021

*β1* = Coefficient for variable HtC 2021

*X2* =predominant age group HRP

*β 2* =Coefficient for predominant age group HRP

*X3 =* predominant gender group HRP

*β3* = Coefficient for predominant gender group HRP

*X4* = predominant household size

*β4* = Coefficient for predominant household size

*X5* = predominant qualifications for HRP

*β5* = coefficient for predominant qualifications for HRP

…

*X1X2* = interactionbetween HtC 2021 and predominant age group HRP

*β5 =* Coefficient for the interaction between the variable HtC 2021 and predominant age group HRP

…

*X1X2X3X4X5* = interaction between all predictor variables

*βi* = Coefficient for the interaction between all predictor variables

*Єi* = error term

To perform this analysis the set of all LSOAs in England and Wales (34,753) was used to identify the most influential variables in the prediction of the self-response (response without any intervention from field force or reminder letters) by the end of the self-response period (day 10 after census day).

The analysis revealed that all predictor variables in the model were significant in predicting the change of the self-response by day 10 after Census Day. It also shows that the most influential variable in the model was the Hard-to-Count (HtC) rating assigned to each LSOA. The second most influential variable in the model was the predominant age group for the HRP, followed by all the remaining variables (i.e. gender, household size and qualifications).

**Multiple regression with a random sample**

A second Multiple Regression was performed using a random sample of 10% of the initial observations. Research suggests that in very large samples, p-values are often below .001, and therefore relying on p values can lead the researcher to claim support for results of no practical significance (Royall, 1986; Lin, Lucas & Shmueli, 2013; Knapp; 2017).

The regression formula with interactions using a random sample of 3,488 LSOAs is presented in equation (3).

*Yi = β0 + X1 β1 + X2 β2 + X3 β3 + X4 β4 + X5 β5 + X1X2 β5 + X1X3 β6+ … X1X2X3X4X5βi + Єi  (3)*

Following the same approach as the previous analysis, equation (3) contains a continuous dependent variable representing the response rate by day 10 after Census in 2011 for a random sample of 3,488 LSOAs in E & W (see table 3 for descriptives). The predictor variables in the model are presented next with the description of all possible levels:

* HtC (one of 5 levels – HtC 1; HtC 2; HtC 3; HtC 4; HtC 5);
* predominant age group HRP (one of 5 levels – up to 24; 25 to 44; 45 to 64; 65 to 74; 75 and over);
* predominant gender group HRP (one of 2 levels – female; male);
* predominant household size (one of 6 levels – HH1; HH2; HH3; HH4; HH5; HH6 and over);
* predominant qualifications for HRP (one of 7 levels – No qualifications; level 1; level 2; apprenticeship; level 3; level 4; other qualifications).

Table 3. Between-subject factors and number of observations for each factor level in the model with a random sample of 3,488 LSOAs.



Table 3. Between-subject factors and number of observations for each factor level in the model with a random sample of 3,488 LSOAs.

As for the previous analysis, the model with the random sample of 10% of the set of LSOAs showed a similar pattern as the one using the entire sample (see appendix 3). We can see that all the main effects were still significant and the magnitude of the effects followed the same structure with HtC and age being the most influential variables in the model. We can also see that HtC significantly interacts with all the other variables (age, gender, household size and qualifications).

The model presents an R2 (R Square) of 0.80, meaning that 80% of the proportion of variance in the dependent variable (response rate by day 10 after Census day) can be explained by the independent variables in the model.

***Post Hoc Tests***

Detailed *Post hoc tests* for each independent variable were also performed to provide insight about the relationship between the predictor and outcome variables. For instance, an analysis of the different HtC levels revealed that they are significantly different from each other (sig <.0001). More precisely HtC 1 is significantly associated to higher self-response rate (D10 after Census day) compared with the other HtC levels. The opposite trend is observed in HtC 5 response that is found to be significantly lower compared with all the other HtCs.



Table 4. Post hoc test (multiple comparisons) for HtC levels using a Bonferroni correction in the model with a random sample of 3,488 LSOAs.

The second *Post hoc tests* considered the variable predominant age group for HRP. An analysis of the different levels of the variable showed significant differences between each one of the levels of the variable (sig <.0001). For instance, age group 25 to 44 was associated with the lowest self-response whilst age group 75 and over presented the highest value for response in day 10 after Census.



Table 5. Post hoc tests for predominant age group for HRP in the model with a random sample of 3,488 LSOAs.

Regarding the variable predominant household size *Post hoc tests* also provided significant differences between all levels of the variable (sig <.0001). More precisely, households with 5 occupants provided a significantly lower response than households with 1 or 2 occupants. Also, households with 1 occupant were associated with significantly lower responses when compared to households with 2 occupants.



Table 6. *Post hoc tests* for variable household size in the model with a random sample of 3,488 LSOAs.

The last step included evaluating the interaction effects between predictor variables in the model. Several interactions were found significant (see appendix 3) suggesting that profiles might need to consider these effects as they are influential in predicting response patterns. For example, a detailed analysis of the interaction between HtC, age, gender and HH size revealed significant effects for males’ response to Census but not significant for female HRP as shown in Figure 3.



Figure 3. Plotting interaction between HtC, age, gender and HH Post hoc tests in the model with a random sample of 3,488 LSOAs.

**Modelling continuous longitudinal data**

The last analysis consisted of performing a Generalized mixed model (GLIMMIX procedure) using the predictor variables HtC and predominant age group for HRP, and the time variable as a random effect. Weekly response rates (with five time points) for all LSOAs in England and Wales were included in the analysis to evaluate the effect of the two predictors (between-subject variables) on time (repeated measures variable).



Table 7. Fit Statistics for mixed model regression using the GLIMMIX procedure in SAS

The model shows that both between subject variables (HtC and Age group) are significant predictors of the period of response. It also shows that the repeated measures variable (Time) and the mixed effects interaction are significant (see table 8). This analysis suggests that HtC and Age are influential for the distribution of the response rate by LSOA across the five different time periods (see table in appendix 4).



Table 8. Type III effects for mixed model regression using the GLIMMIX procedure in SAS

# Discussion

This paper discusses research developed to create the response profiles and associated daily estimates for the Census in 2021. The comprehensive examination of the 2011 Census data and the 2017 Test data is necessary to develop statistical models that will enable the production of such estimates for the future Census.

The first approach considered the use of cluster analysis to identify patterns of self-response during the Census in 2011. This analysis, which considers the variables HtC and response in the week before, during and after the Census, has revealed 5 different clusters with distinct response patterns. It has also confirmed that the cluster aggregation (i.e. profile groups) consistently differs from the HtC categorization. More precisely, the same cluster (response behavior) might have LSOAs from different HtC classifications.

In fact, the cluster analysis clearly demonstrates that the use of the HtC classification alone is inadequate to predict patterns of response, and therefore, it should be used in conjunction with other variables that are influential predictors of census self-response. However, we will need to acknowledge the fact that this analysis does not provide any information about the predictor/outcome relationship between variables and the model failed to converge when more variables were included.

Subsequently, with a view of testing further predictor variables of self-response, a multiple regression (GLM) was used as the next exploratory analysis. Using a range of variables extracted from the Census 2011, we aimed to identify the most influential variables in the prediction of the self-response (response without any intervention from field force or reminder letters) by the end of the self-response period (Day 10 after census day).

Two regression analyses (one using the entire sample and the second a 10% random sample) identified all predictor variables in the model as significant. Furthermore, the analysis accounted for 80% of the variation of self-response by day 10 after Census Day. Once more, both analyses confirmed HtC as the most influential variable, closely followed by predominant age of the HRP. Other variables such as gender, qualifications and household size were also significant in the model but had very limited effect sizes.

Theses analyses were followed by *post hoc tests* to clarify the relationship between the predictor and outcome variable. Pairwise comparisons for the variable age group revealed that age group 25 to 44 was associated with the lowest self-response whilst age group 75 and over presented the highest value for response in day 10 after Census day.

Importantly, while the regression approach provides extensive information about the impact of the variables on response, this method can only explore a single outcome variable. Furthermore, the data needs to have a linear distribution for each level of the independent variables. These two factors constitute a limitation of this technique.

Therefore, a further exploratory analysis was deemed relevant: a Generalized mixed model approach using the predictor variables HtC with predominant age group for HRP, and the time variable as a random effect. This model confirmed that both between subject variables (HtC and Age group) were significant predictors of the period of response and that both are influential for the distribution of the response rate by LSOA across the five different time periods.

The final model was designed to consider a maximum of 25 different response profiles (5 HtC levels \* 5 age groups) and provide estimates for 5 different time periods during the Census in 2021 (3 weeks before, Census week, and the week after). However, due to the reduced number of observations in some profiles it was only possible to generate significant estimates for 17 of the potential 25 profile groups.

Nevertheless, we anticipate that a similar model (including as a minimum HtC and Age group) might indeed be the best approach to generate the daily estimates that we will require for the 2021 Census.



Table 9. Estimates for the initial 10 days of self-response during the 2021 Census by Response Profile cluster derived from a time series model.

# Appendix

Appendix 1. Model summary from the two-step cluster analysis performed on SPSS version 23.



Appendix 2. Main and interaction effects from the regression model with a sample of 34,753 LSOAs.

Appendix 3. Main and interaction effects from the regression model with a random sample of 3,488 LSOAs.

Appendix 4. Estimates from interaction between HtC and Age for five time periods.



#  References

Cai, K., & Wilson, J. R. (2015). How to Use SAS® for GMM Logistic Regression Models for Longitudinal Data with Time-Dependent Covariates. In *Proceedings of the SAS Global Forum 2015 Conference*.

Dietrich, T., Schuster, L., & Connor, J. (2014). Segmenting Australian high school students: Two-step cluster analysis preliminary. *European Journal of Public Health*, *24*(suppl\_2).

Dietrich, T., Rundle-Thiele, S., Schuster, L., Drennan, J., Russell-Bennett, R., Leo, C., & Connor, J. (2016). Segmenting Australian High School Students Utilising a Two-Step Cluster Analysis: Differential Effects Following the Game on Know Alcohol Program. In *Rediscovering the Essentiality of Marketing* (pp. 413-414). Springer, Cham.

Knapp, T. R. (2017). Significance Test, Confidence Interval, Both or Neither?. *Clinical Nursing Research*, *26*(3), 259-265.

Lin, M., Lucas Jr, H. C., & Shmueli, G. (2013). Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research*, *24*(4), 906-917.

Meyer, B. D., Mok, W. K., & Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, *29*(4), 199-226.

Tkaczynski, A. (2017). Segmentation using two-step cluster analysis. In *Segmentation in Social Marketing* (pp. 109-125). Springer, Singapore.

Racinskij, V. (2015). Precision of population estimates for different levels and variability of response. Internal ONS paper

Royall, R. M. (1986). The effect of sample size on the meaning of significance tests. *The American Statistician*, *40*(4), 313-315.

Salganik, M. J. (2017). *Bit by bit: social research in the digital age*. Princeton University Press.

Sarstedt, M., & Mooi, E. (2014). Cluster analysis. In *A concise guide to market research* (pp. 273-324). Springer, Berlin, Heidelberg.

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

1. The 2016 test was a small-scale test conducted by the ONS where 11,463 participants were contacted and asked, “If tomorrow was Census Day, how would you prefer to complete your form?". Using a universe of 1,789 online respondents this test revealed that all age groups between the ranges of 18 to 74 had a preference to respond online (above 75% preference) and only the above 75 preferred to respond on paper (55%). [↑](#footnote-ref-1)