Model selection for the coverage estimation of the 2021 Census of England and Wales

Viktor Račinskij, ONS, Titchfield, UK

21/09/2021

Version: 0.6

Commented: Paul A. Smith, Owen Abbott, Gareth Powell, Abu Hossain

Disclaimer: all results and ideas in this report are subject to further review

Requested actions from the panel

The assurance panel is asked to provide feedback and suggestions on the proposed model selection approach. Thoughts on the following would be appreciated:

- pre-selection of key estimation variables
- use of K-fold cross-validation with stepwise selection when selecting interaction terms
- use of prediction errors estimated from cross-validation for decision making, rather than significance levels
- limitations around goodness-of-fit / diagnostic checks

0. Executive summary

The 2021 Census population estimates will be derived from a combination of the Census data and the Census coverage survey, which measures the census coverage. The coverage is estimated through a statistical modelling, where the model is not specified in advance but is obtained through a process which works towards finding the best model, given some constraints. The chosen model will have an impact on the census population estimates, so it is important that the model selection process is designed to be robust, transparent and follows best practice as much as possible.

We intend to use a structured approach to model selection using a combination of standard techniques. This paper will describe this model selection strategy. Briefly, this will work as follows:

- Pre-selected key variables (those most important for outputs) will be included in all versions of the model. These include accommodation type, age-sex group, economic activity, ethnicity, household size, marital status and tenure Other variables (e.g. address one year ago, born in the UK) will be assessed as potential candidates for inclusion.
- Standard descriptive and diagnostic tests will be performed to evaluate candidate models, and numerical issues checked.
- Only variables included in the model may also be included as potential interaction terms.
- Interaction terms will be selected within K-fold cross validation (using 5 folds) with stepwise selection.
- Prediction errors estimated from cross-validation will be used for decision making, rather than significance levels.

- Selected model candidates will be assessed for goodness of fit, and model diagnostics will be performed.

1. Introduction and context

The coverage estimation of the 2021 Census of England and Wales has been developed around the mixed effects logistic regression approach [Račinskij 2018, Račinskij 2019, Račinskij and Hammond 2019, Račinskij 2020, Burke and Račinskij 2020, Hossain 2020, Račinskij 2021]. In all the above research, availability of a reasonable model of the coverage probabilities was assumed. For the live processing, such a coverage model should either be prespecified in advance based on the past modelling experience and expert knowledge or selected from the space of available models once data for modelling become available. This report discusses the model selection process and related concepts proposed for the 2021 Census coverage estimation.

Models are primarily required to estimate the census coverage probabilities given a certain combination of household and individual characteristics such as age-sex, tenure, economic activity, ethnicity, etc. Once a suitable model is fitted and the coverage probabilities are estimated, the entire census data are scored. In other words, the model is used to compute the fitted coverage probabilities for every census record. These fitted probabilities are transformed into coverage weights in the case of undercoverage or used directly in the case overcoverage to up / down-weight each census record based on the observed set of characteristics. The coverage error corrected population size for a domain of interest is obtained by summing up the weighted census records that belong to the domain.

There are several reasons motivating such a modelling approach to be used for the census coverage and error correction. The modelling approach allows reduction in the number of estimated parameters. Assuming that a model is carefully selected, producing indirect estimates of the domains of interest (say, age-sex group by local authority) results in higher accuracy, measured in terms of the mean square error, compared to the approaches where the direct ratio estimates are first obtained for some estimation domains (say, age-sex group by hard-to-count by estimation area) followed by indirect estimation of the domains of interest (age-sex group by local authority). For instance, there are approximately 11,000 such estimation domains for age-sex by hard-to-count by estimation area (35 age-sex groups, roughly 3 hard-to-counts per estimation area, 106 estimation areas), meaning that 11,000 parameters need to be estimated. The regression approach aims to reduce substantially the number of parameters that are estimated, thus providing variance reduction. Of course, direct estimation methods tend to have smaller bias compared to indirect methods. However, there exist coverage estimation specific biases, such as heterogeneity bias, that affect the direct estimation methods through the dual system estimation (dual system estimates are pooled across the ratio estimation post-strata and then used with the ratio estimator). In principle, when enough covariates are available, the regression approach tends to outperform the simple dual system estimator. Moreover, an additional estimation bias may be incurred when the direct ratio estimation is combined with some small area methods to estimate the size of domains of interest. The research cited above found that the regression approach gives a better overall balance between the variance and bias compared to the methods used in 2001 and 2011, though not necessarily uniformly.

The data being modelled are the Census coverage survey linked to census data (within the coverage survey sample areas for the undercoverage modelling and outside the sample areas for the overcoverage modelling). The target sample size of the coverage survey is around 320,000 – 335,000 households or roughly 755,000 individuals across all local authorities of England and Wales. The actual number of observations available for modelling is smaller due to household and within household non-response and because the coverage models are fitted to the margins of the two-way tables (say, when modelling the census coverage, all the observations of the census coverage survey determine the number of observations in the modelling exercise). The ultimate complexity of models supported by the data also depends on the observed coverage (as opposed to the real coverage that can differ from the observed due to the dependence between the two sources) of a source that is being modelled.

There are four modelling exercises required, as models are needed for the following stages of coverage estimation:

- undercoverage of individuals in the general population,
- undercoverage of households in the general population,
- overcoverage in the general population,
- individual undercoverage in the communal establishments.

All models under consideration are expected to reflect well the differential coverage patterns so that reliable population totals for the domains such as age-sex by local authority can be produced. On the other hand, the number of estimated parameters should be kept reasonably small ('small' can be quite large for some of the modelling exercises like undercoverage of individuals in the general population).

While all these models have many common attributes (logistic or mixed effects logistic that can produce high quality indirect estimates for multiple demographic groups) each of the modelling exercises may have its own challenges. For instance, the general population undercoverage modelling uses very large data sets which presents various issues. For instance, global goodness-of-fit tests show lack of fit unless overfitting. On the other hand, modelling the undercoverage in communal establishments may be difficult due to the small number of observations available. Nevertheless, the expectation is that good quality domain estimates will be produced. Thus, model selection procedures and systems are required to fulfil the specific demands of the census-based population size estimation and be flexible enough to accommodate the differences between modelling exercises for different coverage estimation stages.

There are also practical constraints that shape the selection procedures and system. It is expected to have a processing stage when the partial census and coverage survey data are available for some testing and tuning. Once the entire data for the coverage estimation are ready, there are six weeks available for the coverage estimation processing. This processing window includes the final stages of pre-processing, conducting all modelling exercises listed above, doing the dependence bias adjustment, waiting for the national adjustment sex ratios benchmarks to be produced, applying the national adjustment, and running variance estimation (which can in principle extend outside the six weeks window). It means that there are around three weeks for all modelling exercises. Therefore, the selection procedures must be well pre-specified and implemented in advance. In addition, a high level of automation in processing is required. Another practical consideration is software. All the research for the coverage estimation of the 2021 Census was carried out in SAS. Since the model selection system needs to be integrated with the coverage estimation systems (which is derived from the research environment written in SAS) and the fact that the model selection system itself partly re-uses existing code and largely re-uses expertise and optimization techniques, the selection system itself is implemented in SAS.

The structure of this paper is as follows. In chapter 2 an overview of potential model selection methods is presented and each is assessed for its suitability for the coverage estimation. Chapter 3 discusses general principles of the model selection process and proposes a selection system using the suitable methods from chapter 2. Goodness of fit and model diagnostics is discussed in chapter 4. Finally, chapter 5 outlines the methods for resolving issues around localized lack of fit.

2. An overview of several model selection methods and their assessment for the coverage estimation

In this chapter several model selection methods that have been considered for the census coverage estimation are discussed and assessed in terms of their suitability for the task.

Prespecified model based on expert knowledge or purpose of a model. Numerous coverage modelling exercises using the data from 2001 and 2011 Censuses were conducted over the years for research purposes and for coverage estimation simulation studies. Therefore, at least in principle, there is sufficient experience to allow us to determine in advance what a reasonable model should look like and therefore simply prespecify (or nearly prespecify) all models.

However, having a completely or nearly completely prespecified model means that an opportunity of selecting ‘near the best’ model given the observed data may be missed and the estimation could be suboptimal. In addition, our research using simulation data shows that careful selection of interaction terms is crucial in the coverage modelling. While some important interactions may be known in advance, there is always a chance that either some important interactions will be missed, or those included will be redundant. Also, a careful model selection requires special attention to potential numerical issues such as singularity covariance matrix, quasi-complete data separation, etc. and a prespecified model can have numerical failures. Therefore, the completely pre-specified approach was ruled out.

There is some room and even a need for model pre-specification, though. First, given that population size estimates should be produced by certain key census variables (for the general population these are accommodation type, age-sex group, economic activity, ethnicity, household size, marital status and tenure) it is reasonable to have all these variables as main effects in our models. Actual categorization of these variables can be altered if needed (see below). Also, the way the coverage survey was designed and the sample allocated means that the hard-to-count variable (design variable) must enter every model as a main effect. Therefore, we will design the model selection system to be flexible enough to explore any preselected model. This will allow a set of reasonable preselected models (main effects and interactions) to be passed through the system, checked for numerical issues and assessed for goodness of fit.

Full model. In some applications, it is possible to include all effects into a model. Such an approach effectively avoids any search for a model in the space of possible specifications. It includes all statistically significant and insignificant effects in the model.

It is not a suitable option for the coverage estimation task. First, since the interaction terms are very important for the coverage estimation, there is a question of what the highest order of interaction that should be regarded as an effect is. Second, choosing the full model (assuming that interactions are included) contradicts the goal of keeping the variance of the population size estimates low by having fewer parameters than the direct ratio method would have had. Essentially, the full model will lead to undesired overfitting. Third, the full model is likely to be numerically unstable and thus unreliable and unsuitable for the task.

While the full model is rejected, it is technically feasible to explore the performance of the full model (with up to n^{th} order interactions) trimmed for the numerically unstable effects. As this can be useful to benchmark against other models, we will ensure that some sort of a full model (say, the model with main effects and all second order interactions) can be passed through the system and the system will return the fullest numerically stable model.

Model averaging. In this case several plausible models are specified independently and their parameter estimates are averaged. There are several reservations regarding this approach. First, a selection procedure would still need to be designed and implemented to produce the models that are being averaged Therefore, decisions about how the individual models are selected are still required. Second, arguably, the model averaging pays off when different models give very different predictions. However, experience from the coverage estimation research and simulations shows that once several reasonably well specified and numerically stable models are selected, the results between these models are often quite similar for the majority of domains. Moreover, competing models often have similar issues with the same domains. This can be explained by localised peculiarities or data collection failures. Sometimes, such localised issues can be addressed by fitting a very complex model, say, a model with random slopes. Such complex models may solve the issue of unsatisfactory fit in a handful of domains, but their parameter estimates have very large standard errors and model averaging seems impractical in this case.

Therefore, model averaging is ruled out in general. However, our model fitting approach will need a way to deal with localised issues and problems with numerical stability, which we are calling issue resolution. We propose in chapter 5 a practical averaging-based way of dealing with localised issues called domain averaging.

Purposeful selection. Detailed discussion of this selection approach can be found in Hosmer et al., 2013. This approach stresses careful univariate analysis of each variable that can enter the model, exploring possible transformations of continuous effects. This is followed by bivariate analysis of the pairs of variables that can be considered for interactions. Based on these analyses the main effects model is specified and analysed for dropping or altering effects. Once the preliminary main effects model is selected, each plausible interaction is added and analysed. Those individually entered interactions that are significant at a certain level go into the model with multiple interaction effects. The model with interactions is analysed in a similar way to the main effects only model.

The purposeful selection is attractive because of its careful consideration of the individual effects which is useful in avoiding various modelling issues as the model gets more complicated. This is especially applicable for the bivariate analysis of the possible pairs for interactions. Even if particular main effects are planned to be included irrespective of their statistical significance, such univariate considerations are important. First, there are non-key variables that can be good candidates for modelling. Second, even the variables that are going to be in the model may need some analysis to establish possible collapsing of the categories if this is needed.

Unfortunately, when it comes to considering interactions, the purposeful selection becomes intractable in the case of coverage estimation: for the general population undercoverage models there is a large choice of plausible interactions that are individually significant. Essentially, the purposeful selection becomes cumbersome and computationally infeasible for the appropriate consideration of the interaction terms.

However, it is sensible to pursue elements of the purposeful selection method up to and including addition of each individual sensible interaction into the main effect model, but seek another approach for selection of interactions.

Best subset selection. Best subset fits all possible models and then selects the best fitting one based on a chosen criterion. It is a tempting approach, but it is rejected due to being computationally infeasible in the time available. To see why, we can consider second order interactions. It is difficult to carry out correct calculation of the number of possible models or even to calculate the number of models for a certain number of parameters. For simplicity, just consider the number of possible interactions (which is different from the number of estimated parameters). If there are roughly 120 possible second order interactions, then there are 7,140 possible models with exactly two interactions, 280,840 models with three interactions, 8,214,570 models with four interactions and so on.

LASSO (least absolute shrinkage and selection operator). LASSO is one of the regression methods that performs a shrinkage of parameter estimates. The parameter estimates are obtained by solving a constrained optimization problem which, under certain choices of the constraining or tuning parameter, results in some of the parameter estimates taking zero values. Therefore, the LASSO can be seen as a model selection method.

The LASSO was originally developed for continuous variables. Most of the variables in the coverage estimation are categorical. There are several attempts to extend the method to work with categorical variables, such as the grouped LASSO [Yuan and Lin, 2006]. The problem of the standard LASSO when categorical variables are present is that the selection results may vary depending on the coding of categorical variables. However, the issue encountered when testing the grouped LASSO method for use in the coverage estimation was running time that was too long (even with multithreading) for any practical purposes. Therefore, this approach was ruled out.

Nevertheless, the recent attempts to use the LASSO after carefully removing all the interactions leading to numerical instable results, showed feasibility of having the LASSO approach for additional reassurance.

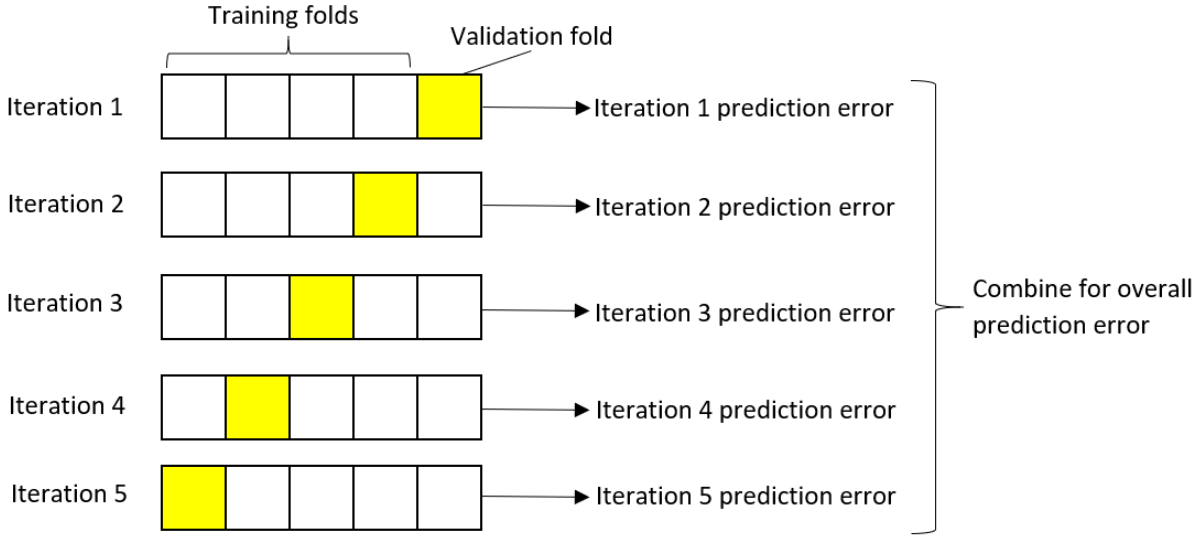
Stepwise selection. The stepwise selection is a compromise best subset selection method. Perhaps the most widespread version of the stepwise selection is when the forward selection is followed by backward elimination. Starting from an empty model (or a set of effects that is forced into the model), an iteration of the stepwise selection is as follows: on the forward selection step, an effect satisfying an entry threshold and making the biggest improvement to the fit is added from the candidate effects. On the backward elimination step, an effect not satisfying a stay threshold and making the smallest improvement to the fit is removed. This iterative procedure continues until, given the entry and stay thresholds, no effect can be added or removed.

The stepwise selection method is flawed in several ways [Harrell, 2015 and references therein]. For instance, it fails to account for multiple comparisons meaning that the selection outcome is almost always suboptimal to some extent. Nevertheless, when it comes to selecting interactions from a relatively large pool of effects, especially when the data set is relatively large, there are not many practical alternatives available. Therefore, we propose that the stepwise selection is used to select interaction terms, although this will be carefully applied to protect against some of the issues associated with the stepwise approach.

Cross-validation / bootstrap validation. Validation is a group of resampling-based methods for prediction error estimation. Validation methods are not model selection methods on their own, but they can be used to assess how a given selection approach performs in terms of predictions on a new data set. It involves repeating a chosen selection procedure on each validation resample (fold) [Hastie et al., 2009]. The validation resample (or fold) allows multiple versions of the selection procedure, thereby allowing an assessment of how robust the procedure is.

Among the most widely used validation methods are K-fold and bootstrap cross-validation. In the K-fold cross-validation the data are split into K non-overlapping parts of roughly equal size. For each part $k = 1, \ldots, K$ we run a selection procedure on the data that excludes the k^{th} part and score the k^{th} part with the selected model and then calculate the prediction error (usually based on some loss function). The process is repeated for all K folds and then the K estimates are combined to estimate the cross-validated prediction error. Figure 1 shows an illustration of five-fold cross validation.

Figure 1 – Illustration of 5-fold cross validation.


Training folds
Validation fold
Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
Iteration 1 prediction error
Iteration 2 prediction error
Iteration 3 prediction error
Iteration 4 prediction error
Iteration 5 prediction error
Combine for overall prediction error

The bootstrap validation draws bootstrap resamples from the data, performs selection with a given approach and computes the prediction error on the original data. Because the data on which the model is selected and on which prediction error is computed are different, validation methods penalize non-parametrically for increased number of parameters thus preventing overfitting.

Given that the coverage estimation is to a degree a prediction problem where the model fitted on the Census coverage survey sampled areas is used to score the entire census data, estimating the prediction error is of interest. Furthermore, we are aiming to keep the variance low, which is related to preventing overfitting. Lastly, validation methods provide some protection when using stepwise selection, as we are intending to use it for interaction selection.

As for choosing between the K-fold cross-validation and bootstrap validation, several points must be considered. First, the simple bootstrap as described above often does not provide a good estimate of the prediction error due to overlapping of the training and test data. Therefore, one or another modification of the approach, such as the ".632 estimator" [Hastie et al., 2009] may be required to obtain a good estimate. Second, the Census coverage survey has a stratified cluster sampling design. It means that resampling at the individual or household level does not reflect the way the coverage models are fitted and prediction is made. Whilst it is easy to replicate the sampling design within the bootstrap approach, it is non-trivial to implement the ".632 estimator" with the complex survey. On the other hand, it is only approximately possible to mimic the sampling design with the K-fold cross-validation, but the prediction error estimate is expected to be more accurate and the implementation easier. Therefore, we propose use of the K-fold cross-validation method to assess the prediction error for our model selection process.

3. Overview of the selection process and model selection system

In this chapter an overview of the proposed overall model selection process and the model selection system is provided.

As discussed in Chapter 2, there is no single model selection approach that can be readily applied for the coverage modelling task. Whilst we want to be able to determine the best possible model, we have a limited timeframe and therefore we require a selection process which can quickly identify the most promising effects, and rejects models which may be unstable or do not have an explainable structure. This means there will have to be some carefully chosen criteria which effectively reduces the number of possible models which can be fitted and analysed to a manageable level. Hence, a careful combination of a number of approaches is used for model selection. The high level principles are:

- We will always include key (and design) variables in the models, treating them as pre-specified.
- Standard descriptive (e.g. univariate, bivariate analysis) and diagnostic tests will be used as is good modelling practice. Numerical issues will be fully checked, and no models with issues will be accepted.
- First order interactions will be analysed as per the purposeful model selection framework
- K-fold cross validation and stepwise selection will be used to select second and third order interactions. No fourth order interactions will be considered.
- Hierarchies within the model will be enforced – that is for A*B to enter, A and B must be present.
- Prediction errors estimated from cross-validation will be used for decision making, rather than significance levels. The variance of the resulting coverage weights will also be used as a diagnostic.

- Multiple values of tuning and threshold parameters will be explored to provide a form of sensitivity analysis and ensure results are robust.

A more detailed specification of what diagnostics and set-up processes that will be adopted is detailed at Annex A. The data required for the model selection procedure are detailed in Annex B.

The proposed selection system consists of eight stages. It is semi-automated in the sense that parameters must be specified by a human, but once they are specified and the input data available, the system carries out the tasks associated with that stage. Each stage depends on the previous ones in the sense that each stage produces outputs for the succeeding stage. If input data are available for a given stage, it can be run without running the preceding or succeeding stages. It is possible to specify parameters for all stages and run all of them at once.

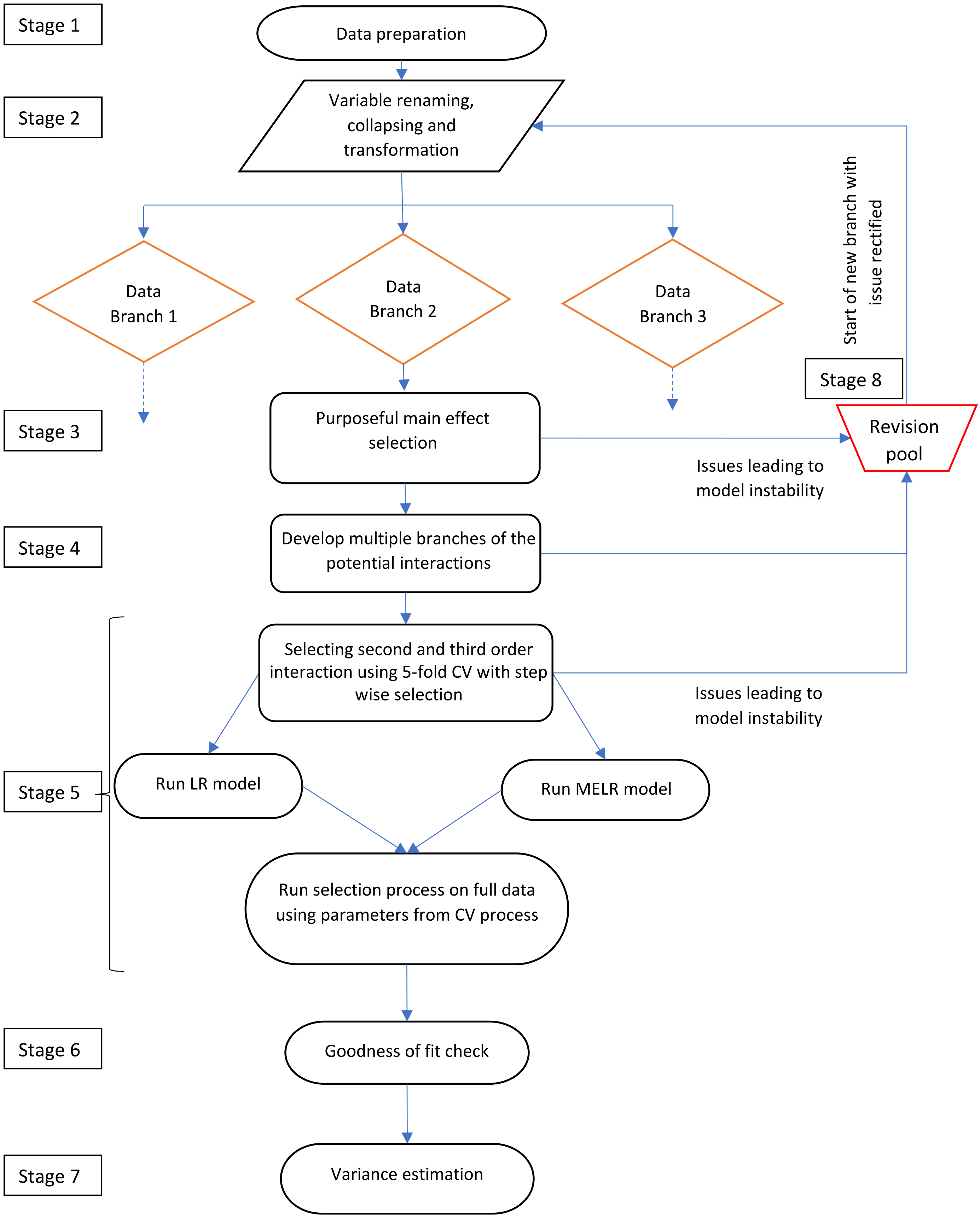
The stages and their purpose are as follows:

1) **Data preparation**: Selects the data for a specified model and does the filtering of observations (say, out of scope records).
2) **Initial data analysis**: Performs the flexible renaming, recoding, collapsing (if needed) and transformation of the variables. Produces descriptive statistics, univariate, bivariate and trivariate analysis by fitting simple logistic models, collects and stores the fit information.
3) **Purposeful main effects selection**: Automatically performs the purposeful main effects model selection under given parameter specification with the key variables forced into the model.
4) **Initial interaction analysis**: Does one-by-one analysis of the main effect model with each single interaction (or hierarchy of interactions) added. Produces the corresponding likelihood ratio tests. Checks for the numerical issues like quasi-complete separation or singularity of the covariance matrix. Collects and stores information on issues and their nature. Uses different model fitting procedures (*hplogistic* and *logistic*) and options to check whether some of them overcome issues like singularity of the covariance matrix.
5) **Cross validation for second and third order interactions**: K-fold cross-validation (using five folds) with specified parameters is run. This is the most technically involved part. The selection parameters that lead to a set of models with the smallest prediction errors for the logistic and the mixed effects logistic models are used to perform the selection on the full data. The models in this set will be assessed for the goodness of fit and model diagnostics will be performed.
6) **Goodness of fit and diagnostics**: Goodness of fit checks and diagnostics for the models in the set of 'best' performing models. This can lead to some new selection branches to be built.
7) **Variance estimation**: (Still under consideration) for the undercoverage models, use the model to produce undercoverage error corrected population size estimates and use the bootstrap approach to estimate their corresponding variance.
8) **Issue resolution**.

The selection system allows running and versioning multiple runs for each of the stages, so that different main effects models can be selected with varying selection parameters or different cross-validations run with different tuning parameters and / or forced in and out interactions.

Versioning is manual. The selection system permits a tree structure for the model selection process. Once the data for a particular type of model are ready (stage 1), the stage 2 processes allow producing multiple branches (or versions) of data based on the variable collapsing or transformations. Each of these branches will have the corresponding main effects model or models. Therefore, additional branches at stage 3 are created in the case of multiple main effects models. For a single main effects model any number of branches with selection of the interactions can be built. If any issues leading to model instability are detected along some branches, it is possible to start a new branch with issues rectified (usually at stage 2) to attempt to build a model not affected by the detected issue. In addition, stage 5 allows producing two related models (that is, allows two sub-branches): the logistic regression and mixed effects logistic regression. Each model produced by this branching is assessed using the estimated prediction error.

In practice, the plan is to start selection parameter pre-specification and test runs of the selection system once some tuning data become available. During the main estimation processing window, two persons are expected to work per modelling exercise at a time. One is responsible for running the selection system stage by stage or in batches and for deciding the branching of the selection process. Alongside the semi-automated system as outlined above, there are standalone processes. These standalone processes are usually extracts of the main selection systems that perform a single task only. As the selection process moves through the stages, explores different parameterisations and collects diagnostics, all the problematic cases go to the revision pool. The second person is responsible for investigating the cases in the revision pool. Say, if a certain interaction is detected to cause a data separation issue, this interaction is sent to the revision pool. Appropriate standalone processes are then run to establish whether the issue can be meaningfully resolved by collapsing or not. If the issue can be resolved by collapsing, the collapsed version of the variables can start a new selection branch from the stage 2. There is no need to wait until the cases in the revision pool are resolved. Models with non-problematic effects can go through the system first. Once the cases in the revision pool are revised and fixed, more models are built.


Stage 1
Data preparation
Stage 2
Variable renaming, collapsing and transformation
Start of new branch with issue rectified
Data Branch 1
Data Branch 2
Data Branch 3
Stage 8
Revision pool
Stage 3
Purposeful main effect selection
Issues leading to model instability
Stage 4
Develop multiple branches of the potential interactions
Selecting second and third order interaction using 5-fold CV with step wise selection
Issues leading to model instability
Stage 5
Run LR model
Run MELR model
Run selection process on full data using parameters from CV process
Stage 6
Goodness of fit check
Stage 7
Variance estimation

4. Goodness of fit / diagnostics

Once the tuning parameters producing the smallest prediction error are determined (it is likely that there will be several combinations of such tuning parameters), the stepwise selection of the interaction terms is performed on the entire data. The goodness of fit of the resulting models is assessed and the corresponding regression diagnostic checks are performed.

A goodness of fit measure is a single number that formally summarizes the agreement between the observed and fitted values. The purpose of diagnostic measures is to examine fit for across the covariate patterns (in our case also across the domains of interest).

The cross-validation on the full data with a fixed model specification can be also used to assess goodness of fit, but it does not provide assessment based on the distribution of some test statistic. The goal is to use the range of goodness of fit tests. For the logistic regression model the standard tests suggested by Hosmer et al. [1997] and Hosmer et al. [2013] will be used (subject to resolution of performance issues): Pearson, deviance, Hosmer-Lemeshow, Osius-Rojek, unweighted residual sum-of-squares and Stukel tests. All these tests except for the Stukel test assess the agreement between the observed and fitted values. The Stukel test assesses the validity of the assumption logit link assumption. For the mixed effects logistic model the smoothed residual based test [Sturdivant et al., 2007] is considered.

There are several limitations related to the usage of these tests. First, for a model built on the data of small or moderate size, these tests can genuinely indicate whether there is the lack of fit. However, for a models built on a large data set, these tests almost always will indicate the lack of fit, unless some trivial phenomenon is modelled or the model selected overfits. This behaviour can be explained informally by the fact that in a large data set there almost certainly are some areas of poorer fit that result in the increase of a test statistic and subsequent rejection of the null hypothesis that there is no lack of fit. Second, some of the above test, notably the Hosmer-Lemeshow, have been targets of severe criticism with various alternatives proposed (like the unweighted residual sum-of-squares). Nevertheless, such features of the Hosmer-Lemeshow test as grouping of the data based on the values of estimated probabilities is useful for a less formal assessment of goodness of fit.

For the diagnostic checks of the logistic regression the standardized Pearson residuals, the change in the values of estimated coefficients and the Pearson chi-square statistic incurred by deletion of a covariate pattern is computed. The above statistics are plotted against the estimated probabilities for each of the covariate patterns. As in the case of assessment of goodness of fit, some of these diagnostic procedures may have limited value for large models. For the mixed effects models, the quantile-quantile plots are produced to assess the assumption of normally distributed random residuals.

For both types of models, comparison of the observed and fitted probabilities for the domains of interests is useful since the models are built in attempt indirectly reflect the coverage in these domains. The difficulty with such comparison is that the correct sampling distribution of these (standardized) differences is difficult to obtain under the complex sampling design and dependence within the random blocks (like local authorities). However, even a rough approximation of the sampling distribution is more informative than some of the checks listed above.

5. Issue handling

It is possible to encounter a situation where a selected model has good performance across the majority of the key domains of interest, yet the model diagnostics detects substantial discrepancies between the observed and fitted coverage probabilities for some of the domains. For instance, a small number of age-sex by local authority domains may exhibit such localized lack of fit. Another scenario is when an entire local authority has localized lack of fit which may be exacerbated by the local authority substantially influencing the estimates for the remaining domains.

A proposed solution is to fit a local model into the data for the local authority where the problem was detected. Such model is likely to have broader categories for the key variables. Once the coverage probabilities of the local model are estimated, the composite estimator is applied to produce the weighted average of the main model and the local model probabilities. The composite estimator can simply average out the probabilities, but a more complex weighting strategy can be also used.

If a local authority simultaneously has the localized lack of fit and substantially influences the rest of estimates, the process is as follows. The local authority is taken out from the data. The selection procedure is repeated without the local authority. A local model is used to estimate the coverage probabilities for the local authority. The fitted probabilities for the local authority are obtained using the main model. The composite estimator is used to combine the main model probabilities (the main model does not use the data for the problematic local authority) with the local model probabilities.

6. References

Burke D. and Račinskij, V. (2020) The 2021 Census coverage survey: sample allocation strategy. *Report presented at the Census External Assurance Panel on 24 March, 2020.*

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. New York: Springer-Verlag.

Harrell, F. E. (2015) *Regression Modeling Strategies*. 2nd ed. New York: Springer-Verlag.

Hosmer, D.W., Hosmer, T., Le Cessie, S. and Lemeshow, S. (1997) A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, **16**, 965 – 980.

Hosmer, D.W., Lemeshow, S. and Sturdivant, R. X. (2013) *Applied logistic regression*. 3rd edition. New York: Wiley & Sons.

Hossain A. (2020) The coverage estimation strategy for small communal establishment of the 2021 Census of England & Wales. *Report presented at the Census External Assurance Panel. Available on request.*

Račinskij, V. (2018) Coverage Estimation Strategy for the 2021 Census of England and Wales. *Report presented at the Census External Assurance Panel on 16 October, 2018.*

Račinskij, V. (2019) Estimation of the household population in 2021 Census of England and Wales: initial ideas and results. *Internal ONS report. Available on request.*

Račinskij, V. and Hammond, C. (2019) Overcoverage Estimation Strategy for the 2021 Census of England and Wales. *Report presented at the Census External Assurance Panel on 17 October, 2019.*

Račinskij, V. (2020) Dealing with informative sampling in the coverage estimation of the 2021 Census of England & Wales. *Report shared by correspondence with the Census External Assurance Panel.*

Račinskij, V. (2021) Adjusting for the dependence bias in the coverage estimation of the 2021 Census of England & Wales. *Report shared by correspondence with the Census External Assurance Panel. Available on request.*

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society* Series B, **68**, 49 – 67.

Sturdivant, R.X. and Hosmer, D.W. (2007) A smoothed residual based goodness-of-fit statistics for logistic hierarchical regression models. *Computational Statistics and Data Analysis*, **51**, 3898 – 3912.

Annex A: selection principles

- The main effect model is built using the purposeful selection, but the key and design variables will be included in the model (up to suitable collapsing if needed) irrespective of their significance or effect size.
- There could be several main effects models with slightly different sets of covariates, differences in levels of categorical variables or differences in transformations of continuous variables.
- Special attention is paid to the simple descriptive, univariate and bivariate analysis in order to prevent issues like complete or quasi-complete data separation.
- All continuous variables are standardized to have zero mean (reason below).
- Individual interactions are also analysed within the framework of the purposeful selection.
- For a specified main effects model, the K-fold cross-validation with the stepwise model selection is used to perform the selection of interactions under varying tuning parameters. Significance levels at which an interaction enters the model and at which it stays in the model are the key tuning parameters.
- For a specified main effects model with certain predefined interactions forced into the selection, the K-fold cross-validation with the stepwise model selection is used to perform the selection of interactions under varying tuning parameters.
- For a specified main effects model with certain predefined interactions forced out of the selection, the K-fold cross-validation with the stepwise model selection is used to perform the selection of interactions under varying tuning parameters.
- The loss function is some linear function of the log likelihood, usually -2 * log(likelihood).
- All second order interactions except some that are meaningless in the coverage estimation context (to appear in Appendix) will be allowed in the selection.
- A subset of all possible third order interactions will be allowed in the model selection (technically, it should be possible to consider all possible third level interactions, it is being investigated).
- No interactions higher than the third order will be considered unless some exceptional

cases.

- All models have hierarchical main effects and interaction structure. That is, a higher order term is allowed only if all possible combinations of the lower terms are in the model (i.e., an interaction A*B can stay in the model only if the main effects A and B are in the model; an interaction A*B*C can stay in the model only if the interactions A*B, A*C, B*C and main effects A, B, C are in the model).
- Decisions are based on prediction error rather than significance level wherever possible.
- Checks for all sorts of real or potential numerical issues (quasi-complete data separation, singular covariance matrices, unrealistically large standard errors of the parameter estimates, unrealistically low / large parameter estimates, etc.) are part of the model selection process. During the cross-validation, application of these checks may result in removal of certain effects.
- Since most selection decisions are made based on the function of the likelihood, any numerical problems may put the existence or uniqueness of the likelihood into question. Therefore, no model with numerical issues will be accepted.
- For each selection, a logistic model is selected first, then the corresponding (i.e., the same main effects and interactions) mixed effects model is fitted. Based on the fit of the mixed effects model further removal of effects is considered both for the fixed effects only and fixed and mixed effects models. This is to account, if needed, for the inflated standard errors in the initial logistic model due to cluster sampling.
- Selecting a best performing undercoverage model based on the estimated prediction error may not result in the optimal choice of model. Estimation of the prediction error accounts for variance and bias, however, the variance is for the fitted probabilities. Estimation is conducted with the weights that are reciprocals of these probabilities and the distribution of the weights is not the same as the distribution of the probabilities. Accounting for the variance of the coverage weights should also be taken in consideration.
- Multiple values of tuning and threshold parameters are used.

Annex B: the data required for the model selection

The selection system requires the following input from the pre-processing: all the census coverage survey and census individuals in the general population within the census coverage survey sampled areas with all the relevant variables, geography information and each individual classified as being captured either in both sources, or in the survey only or in the census only. In addition, for each coverage survey individual, information on whether the link was made outside the postcode of the individual’s location in the coverage survey and geography information of where this link occurred. All the census coverage survey and census households in the general population within the census coverage survey sampled areas with all the relevant variables, geography information and the classification as above. All the census coverage survey and census individuals in the communal establishments within the census coverage survey sampled areas with all the relevant variables, geography information and the classification as above. In addition, the entire census data with all the variables coded in the way the variables of the coverage survey are coded.

Annex C: Stage by stage description of the model selection process and system

Stage 1. Data for a specified model (e.g., undercoverage of individuals, overcoverage, undercoverage in communal establishments) are loaded. The parameter file accepts a piece of SAS code with specific filtering rules which gets parsed and executed in the selection system. Therefore, no hard-coded filtering rules are used. If requested, a process of geography reconciliation is carried out provided a geography lookup file for the Coverage survey is available.

Stage 2. At this stage variables of the data filtered at Stage 2 are recoded, derived, transformed, or dropped. No hard coding is used. The parameter file accepts the rules for the mentioned manipulations. This stage also allows collapsing the levels of variables.

Descriptive statistics such as frequency of each of the binary outcome for every level of the specified variables at the national or subnational levels are produced. Corresponding tables and plots are produced and stored. Essentially, this is where the purposeful selection of the main effects begins.

Simple univariate, bivariate and trivariate models are fitted for a specified set of parameters and the corresponding fit outputs collected and stored.

Stage 3. Using the coding of variables and based on the analyses performed at Stage 2, a multivariable main effects model is specified. It is possible to specify several models (creating different branches each having unique version code) which differ in the number of effects, or the level of collapsing, or the transformations applied. It is recommended to consider a branch with no or very little collapsing to begin with. The key variables are always included (but collapsing and transformations can vary). The fit summary is produced, stored and is sent for human scrutiny. Reduced models are fitted automatically by removing one or several covariates at a time. The parameter estimates of the initial fuller model are compared with the parameter estimates of smaller models for large shifts in values. Several thresholds can be considered (again, producing different branches).

Based on the analyses performed in this stage, certain covariates may be moved into the revision pool. A person responsible for the revision pool runs procedures from stage 2 to explore possible issues or / and collapse some levels of the variables.

Stage 4. At this stage second and third level interactions are assessed in the framework of the purposeful selection. For a certain main effect model selected at Stage 3, the main effects are kept fixed and each single interaction (or hierarchy of interactions in the case of third order interactions) is added. The corresponding likelihood ratio tests are conducted and results are stored. There are several checks for the numerical issues like quasi-complete separation or singularity of the covariance matrix. Information on issues and their nature is produced and stored. Two different model fitting procedures (*hplogistic* and *logistic*) are used to check whether some of the issues can be resolved using a different procedure. Whenever the *proc logistic* fails for a certain interaction and *proc hplogistic* does not fail for the same interaction, the starting values for the *proc logistic* are obtained based on the estimates produced by the high performance procedure and the *proc logistic* is run again.

Interaction that result in issues are sent to the revision pool, but the process of model building continues with the effects available. Once an issue with a specific interaction is rectified at stage 2, a new branch starts.

Stage 5. Five-fold cross-validation is conducted with each fold being run in a parallel session. The decision to use the five-folds is a compromise between what is generally recommended as a reasonable choice [Hastie et al. and literature therein] and attempt to reflect the structure of the Coverage survey. The original structure of the sample is kept as much as possible. That is, output areas are drawn into each fold using the hard-to-count stratification. This can replicate the original sampling up to a point since the allocation is optimal for the hard-to-count index and proportional within each hard-to-count. However, in the original coverage survey sample each stratum (hard-to-count by local authority) is represented by at least two output areas. So in the fivefold cross-validation this structure cannot be preserved.

Below is the description of a single run of the five-fold cross-validation for a specified set of tuning (thresholds for effects entering and staying in the model) and threshold parameters. Cross-validation is carried out for a range of such sets of parameters. Furthermore, the range of the tuning and threshold parameters will be refined based on the first successful results of the model selection runs.

Before the actual stepwise selection is run, the training data are tested for quasi-complete separation and all the interactions that lead to separation are removed (and tracked). The following step is to run the stepwise selection for the first time to check for the interactions that result in cycling of effect addition and removal. All such effects are removed from the list of possible interactions. Such stepwise selection is performed until normal termination of the processes ics achieved and selected model specification is used for further checks. The next check is for large standard errors of the parameter estimates, first using the *hplogistic* procedure (high-performance logistic regression procedure that allows an additional level of multithreading and thus speeding up the process) then by rechecking with the *logistic* procedure. All interactions that have the standard errors larger than a specified threshold are removed. This is followed by checking for large parameter estimates. This is where the standardisation of continuous effects becomes useful. All interactions that have the parameter estimates larger than a specified threshold are removed. Then the stepwise selection is performed again (once all or most of the potential numerical issues have been removed). The selected model is fitted to the training data by two different procedures (*hplogistic* and *logistic*) and two sets of parameter estimates are used to score the test data and compute the loss function. Two values of the loss function are compared and if the difference is larger than a certain prespecified threshold the model is rejected.

If the logistic model is not rejected, then the mixed effects logistic model is fitted to the training data. Parameter estimates and their standard errors are checked, to see that they do not exceed some desired values. Since the stepwise selection is done for the logistic model fitted to the clustered survey data, the standard errors tend to be underestimated. Once the mixed effects model is fitted, some of the interactions can be removed based on the global test for significance, thus leading to the second set of pre-selected effects. Note that, in order to better account for clustered sampling, one would need to fit a mixed effects model with random output area and postcode. This is likely to be computationally infeasible and so is not being considered.

When estimating the parameters of the mixed effects model the maximum likelihood approach using the Laplace approximation method is used. The Laplace approximation allows computation of the marginal likelihood that is comparable with the likelihood obtained from the logistic regression. There is also the conditional likelihood in the mixed effects model. Test data are scored using the fitted mixed effects model. Unlike in the case of the logistic regression, there is no inbuilt implementation of computation of the value of the loss functions in the case of mixed effects models. Therefore, they should be computed from first principles. The conditional log likelihood is straightforward to compute using the outputs of the *glimmix* procedure:

$$logL\,(\hat{p} \mid \hat{u}) = \log\left(\prod_{i=1}^{N} \hat{p}\left(i\,\text{in census} \mid \hat{u}_{l,i\,\in\,l}\right)^{y_i}\left(1 - \hat{p}\left(i\,\text{in census} \mid \hat{u}_{l,i\,\in\,l}\right)\right)^{1-y_i}\right),$$

where *i* is used to index individuals (or households), *l* is used to index the subject of model's random component, $\hat{u}_{l,i\,\in\,l}$ is the estimate of the random residual and *i* belonging to *l* means that an observation *i* is nested in an area *l*, y_i is the value of the outcome variable, *N* is the number of observations in the data that are scored.

The marginal likelihood is more useful for the selection as it allows comparison of the logistic and mixed effects logistic regression-based estimates of the prediction error. However, it is more difficult to compute on the test data. The marginal log likelihood is given by

$$logL\,(\hat{p}) = log\prod_{l}\left(\int_{-\infty}^{\infty}\prod_{i}\left[\frac{exp\left(\boldsymbol{x}_{li}^{T}\widehat{\boldsymbol{\beta}} + u_l\right)}{1 + exp\left(\boldsymbol{x}_{li}^{T}\widehat{\boldsymbol{\beta}} + u_l\right)}\right]^{y_{li}}\left[\frac{1}{1 + exp\left(\boldsymbol{x}_{li}^{T}\widehat{\boldsymbol{\beta}} + u_l\right)}\right]^{1-y_{li}} f(u_l;\ 0, \hat{\sigma}^2\,)du_l\right),$$

where y_{li} is the value of the outcome variable for an observation *i* in an area *l*, $\hat{\sigma}^2$ is the estimated variance of random effect and $f(u_l;\ 0, \hat{\sigma}^2\,)$ is the normal probability density function of the random intercept.

The above expression is approximated by running a numerical integration where *M* values of $u_l^{(m)}$ are generated from the $f(u_l;\ 0, \hat{\sigma}^2\,)$:

$$logL\,(\hat{p}) \approx log\prod_{l}\left(\frac{1}{M}\sum_{m=1}^{M}\prod_{i}\left[\frac{exp\left(\boldsymbol{x}_{li}^{T}\widehat{\boldsymbol{\beta}} + u_l^{(m)}\right)}{1 + exp\left(\boldsymbol{x}_{li}^{T}\widehat{\boldsymbol{\beta}} + u_l^{(m)}\right)}\right]^{y_{li}}\left[\frac{1}{1 + exp\left(\boldsymbol{x}_{li}^{T}\widehat{\boldsymbol{\beta}} + u_l^{(m)}\right)}\right]^{1-y_{li}}\right).$$

When *M* is around 3000 this approximation is very close to the analytical value.

Once processing of each of the folds is completed, the prediction errors of the logistic and mixed effects models are estimated.

Those parameters and interactions forced into or out of the model that result in the smallest prediction errors are declared to be parameters for the selection of the preliminary final models and set to be run on the entire modelling data.

All procedures that can run on multiple CPUs are run on multiple CPUs within each fold-dedicated parallel session.