Model selection for coverage estimation for Census 2021 in England and Wales

The model selection process and chosen models for coverage estimation of Census 2021 in England and Wales.

Contact:
Census Customer Services
census.customerservices@ons.gov.uk
+44 1329 444972

Release date:
9 November 2022

Next release:
To be announced

Table of contents

1. Overview of methods

2. Methodology and steps

3. The selected models for census coverage estimation

4. Variables before collapsing levels

5. Additional information

6. Cite this methodology

1 . Overview of methods

To adjust Census 2021 in England and Wales for coverage errors and produce coverage-adjusted estimates, we used statistical modelling techniques. To do this, the Census Coverage Survey (CCS) was linked to the census, and models fitted to the resulting linked data. For more information, please see our Coverage estimation for Census 2021 in England and Wales methodology.

The models used were logistic regression, which are an extension of the dual-system estimation as used in the 2011 Census. Five models were used at various stages of processing to estimate the population of England and Wales. For each of these, the model selection process was designed to work towards finding the best fitting model, given some constraints. As each fitted model gives different answers, it was important that the model selection process for estimating coverage was built to be robust, transparent, and follow best practice using a combination of standard techniques. This process is designed and outlined in Model selection for coverage estimation (PDF, 338KB).

The models were built using a combination of data from the census and the CCS. The CCS is a 1.45% sample of postcodes across England and Wales, which means roughly 340,000 households are included in the sample. However, the population used for modelling is smaller because of non-response to the CCS. The CCS data were matched to the full census dataset.

Models are used to estimate census coverage error probabilities given a combination of household and individual characteristics, such as age-sex, tenure, ethnicity, and so on. As a result, every census record gets estimated (or predicted) undercoverage and overcoverage probabilities assigned to it based on the observed characteristics of that record. For undercoverage estimation, these probabilities are then transformed into weights, which up-weight the census data. Overcoverage probabilities are used to down-weight the census records. To estimate the population for different domains of interest, these weights are then summed within those domains.

Outline of the model selection strategy

1. Univariate and bivariate analysis of all available explanatory variables and their interactions.
2. Certain pre-selected variables were the basis for census outputs, and so were included in all versions of the model.
3. Main effects models were selected using the purposeful selection approach
4. Only variables included as main effects in the model could be used creating interaction terms.
5. Interaction terms were selected within K-fold cross validation (using five folds) with stepwise selection. Prediction errors estimated from cross-validation were the main metric used to make decisions about which model to choose.
6. Model candidates were checked for numerical issues.
7. Selected model candidates were assessed for goodness of fit and model diagnostics were analysed.

Five modelling exercises were required for coverage estimation, and this process was applied to each of them. The modelling exercises were:

- household population undercoverage of households
- household population undercoverage of individuals
- household population overcoverage of individuals
- household population wrong location of overcoverage individuals
- small communal establishment undercoverage of individuals

All models need to reflect the differing coverage patterns across the populations of interest and should therefore produce reliable population totals for the domain of interest (age-sex by local authority). Although the models have many common attributes, each modelling exercise had its own challenges. For example, for undercoverage, individual population uses a very large dataset, which presents issues such as global goodness of fit tests showing a lack of fit unless overfitting. In contrast, small modelling population sizes, such as for the small communal individual undercoverage model, meant it was difficult to select a model because of the small number of observations available.

2 . Methodology and steps

Several model selection methods were considered for the census coverage estimation, which were discussed and assessed in Model selection for coverage estimation (PDF, 338KB).

The aim of the model selection approach was to select the best possible model, within a limited timeframe. It was important this enabled quick identification of effects for potential inclusion and rejection of unstable or unexplainable models. This chosen approach followed high level principles, including:

- main variables such as age-sex were always included in the models and treated as pre-specified
- standard descriptive statistics, such as univariate and bivariate analysis and diagnostic tests were used to check for numerical issues, such as a small number of observations for a category within a variable and to follow good modelling practice
- first order interactions were analysed by including each interaction one-by-one to see how they performed in the model with the chosen main effects
- K-fold (fivefold) cross validation and stepwise selection were used to select second and third order interactions; no fourth order interactions were considered
- hierarchical structure in the models were enforced, so for an interaction between variables A and B to be included in the model, the variables A and B must also be included in the model as main effects
- the smallest prediction error was used to select the chosen models; the variance of coverage weights was not used as a diagnostic
- to ensure robust results and sensitivity analysis, we explored multiple tuning and threshold parameters

The implemented model selection approach consisted of eight stages. Each stage depended on the previous one as the outputs from a previous stage were required for the next stage.

Stage 1: data preparation

Selects and filters the data for the specified model.

Stage 2: initial data analysis

Flexible renaming, recoding, collapsing and transformation of the variables. Univariate, bivariate and trivariate descriptive statistics were created by fitting simple logistic models and collecting and storing fit information.

Stage 3: purposeful main effects selection

Main variables were forced into the model and purposeful main effects model selection, where parameter specification was selected, such as looking at the significance levels of these effects in the models.

Stage 4: initial interaction analysis

One-by-one analysis of the selected main effects model with each single interaction (or hierarchy of interactions, where interactions can only be included if the variables are main effects) added.

This stage produced likelihood ratio tests, which assesses the goodness of fit between two models. It also checked for numerical issues because of quasi-complete separation where the dependent variable separates one or more independent variables or singularity of the covariance matrix. This, for example, can be caused when the number of variables in the model is greater or equal to the number of observations.

Stage 5: cross validation for second and third order interactions

K-fold cross-validation, outlined in Model selection for coverage estimation (PDF, 338KB), using five folds. This was used to calculate the prediction error for candidate models. From these, those with the smallest prediction errors for logistic and mixed effects logistic models were used on the full modelling dataset.

Stage 6: goodness of fit and diagnostics

The results from the best performing models were then checked for any issues.

Stage 7: variance estimation

For undercoverage models, the bootstrap approach was used to estimate the variance for the undercoverage error-corrected population size estimates. This variance estimation is different from the “main” variance estimation, it only allows to estimate variance for the estimates produced by a single model, while the “main” variance estimation combines all models and adjustments.

Stage 8: issue resolution

Any issues from the model selection process were resolved here and any necessary stages were re-run.

These stages were iterated through to explore different approaches and ideas.

3 . The selected models for census coverage estimation

Household population models

Household population undercoverage of households

For undercoverage estimation of households, a model was selected from the subpopulation of the matched census-to-census coverage survey data. The model was selected to estimate the probability of a match between a census household and Census Coverage Survey (CCS) household.

Following the model selection approach outlined in Section 2: Methodology and steps, the final model is as follows, where the main effects in the model include:

- accommodation type
- household ethnicity
- household structure
- household size
- hard to count index
- region
- self-contained tenure
- initial contact

Two factor interactions include:

- accommodation type by region
- accommodation type by tenure
- household size by region
- hard to count index by region
- self-contained by tenure

Throughout this selection approach, collapsing of the main variables was done to enable numerical issues to be resolved. This allowed for interactions to be included in the model that are representative of the characteristics of those who are census and CCS matched individuals.

Collapsing of main variables

The collapsed levels for household size are:

- one
- two
- three
- four
- five plus

The collapsed levels for tenure are:

- owns outright
- owns with mortgage or part owns or part rents (private)
- rents – council
- rents – housing association
- rents – private
- rents – relative or employer
- rents – other or free

The collapsed levels for region are:

- East of England
- East Midlands and West Midlands
- Inner London
- North East and North West
- Outer London
- South East and South West
- Wales
- Yorkshire and The Humber

The collapsed levels for accommodation type are:

- detached
- semi-detached
- terraced
- purpose-built flat block or tenement
- converted or shared house or caravan
- commercial

The 62 levels for household structure were collapsed into:

- single male, aged 16 to 34 years or other, no children
- single female, aged 16 34 years, no children
- single male, aged 35 to 49 years, no children
- single female, aged 35 to 49 years, no children
- single male, aged 50 to 64 years, no children

- single female, aged 50 to 64 years, no children
- single male, aged 65 years and over, no children
- single female, aged 65 years and over, no children
- single male, aged 16 years and over, or female, aged 16 to 24 years, or other, with children
- single female, aged 25 to 34 years, with children
- single female, aged 35 years and over, with children
- two adults, related, average age 16 to 24 years, no children
- two adults, related, average age 25 to 34 years, no children
- two adults, related, average age 35 to 49 years, no children
- two adults, related, average age 50 to 64 years, no children
- two adults, related, average age 65 years and over, no children
- two adults, related, other, no children
- two adults, unrelated, no children, or 3 plus adults, average age 25 years and over, no children, or 2 plus adults, unrelated, with children, or any other household
- two adults, related, average age 16 to 34 years, with children
- two adults, related, average age 35 to 49 years, with children
- two adults, related, average age 50 years and over, with children
- two adults, related, other, with children
- 3 plus adults, related, average age 16 to 24 years or other, no children
- 3 plus adults, related, average age 25 to 34 years, no children
- 3 plus adults, related, average age 35 to 49 years, no children
- 3 plus adults, related, average age 50 to 64 years, no children
- 3 plus adults, related, average age 65 years and over, no children
- 3 plus adults, unrelated, average age 16 to 24 years or other, no children
- 3 plus adults, related, other, with children
- 3 plus adults, related, average age 16 to 34 years, with children
- 3 plus adults, related, average age 35 to 49 years, with children
- 3 plus adults, related, average age 50 years and over, with children

Household population undercoverage of individuals

For undercoverage estimation of individuals, a model was selected from the subpopulation of the matched census-to-census coverage survey data. The model was selected to estimate the probability of a match between a census individual and CCS individual.

Following the model selection approach outlined in Section 2: Methodology and steps, the final model is as follows, where the main effects in the model include:

- age groups
- accommodation type
- activity last week
- address one year ago
- legal partnership status
- born in the UK
- household size
- initial contact
- hard to count index
- response rate
- individual ethnicity
- region
- relationship
- self-completion
- sex
- student
- tenure

Two factor interactions include:

- age group by sex
- household size by relationship
- legal partnership status by relationship
- address one year ago by tenure
- age group by address one year ago
- born UK by relationship
- activity last week by tenure
- address a year ago by student
- hard to count index by student
- accommodation by born UK
- hard to count index by relationship
- person ethnicity by self-contained
- self-contained by tenure
- age group by student
- household size by hard to count index
- household size by person ethnicity
- accommodation by region
- region by tenure
- accommodation by age group

Throughout this selection approach, collapsing of the main variables was done to enable numerical issues to be resolved. This allowed for interactions to be included in the model that are representative of the characteristics of those who are census and CCS matched individuals.

Collapsing of main variables

The collapsed levels for age group are:

- 0 to 2 years
- 3 to 7 years
- 8 to 17 years
- 18 to 21 years
- 22 to 24 years
- 25 to 29 years
- 30 to 34 years
- 35 to 39 years
- 40 to 44 years
- 45 to 49 years
- 50 to 54 years
- 55 to 59 years
- 60 to 64 years
- 65 to 69 years
- 70 to 74 years
- 75 to 79 years
- 80 years and over

The collapsed levels for household size are:

- one
- two
- three
- four
- five plus

The collapsed levels for tenure are:

- owns outright
- owns with mortgage or partial ownership
- rents – council
- rents – housing association
- rents – private
- rents – employer or relative
- rents – other or free

The collapsed levels for person ethnicity are:

- English
- White other
- Mixed
- Indian, Pakistani, Bangladeshi
- Other Asian
- African, Caribbean, Black other
- Other

The collapsed levels for legal partnership status are:

- never married
- married
- divorced, separated, or widowed
- civil partnerships

Household population overcoverage of individuals

For overcoverage estimation, two models were used. The first model was selected to estimate the probability of a census individual having an observed response correctly enumerated in the census, where the overcoverage population would be referred to as incorrectly enumerated.

As an important assumption of the CCS is that it enumerates the “correct” location of these individuals, we were able to determine if an individual was observed response correctly enumerated (duplicate or enumerated in the wrong location) in the census, by comparing the enumerated locations. The second model was selected to estimate the probability of a census individual being enumerated in the wrong location from correct census enumerations.

Household population correct enumeration model

Following the model selection approach outlined in Section 2: Methodology and steps, the final model is as follows, where the main effects in the model include:

- age groups
- accommodation type
- activity last week
- address one year ago
- legal partnership status
- born in the UK
- household size
- hard to count index
- observed return rate
- individual ethnicity
- region
- relationship
- self-completion
- sex
- student
- tenure

Two factor interactions include:

- age group by sex
- accommodation type by region
- age group by address one year ago
- age group by individual ethnicity
- age group by tenure
- address one year ago by tenure
- household size by tenure
- household size by student
- legal partnership status by relationship
- observed return rate by region
- region by tenure

Throughout this selection approach, collapsing of the main variables was done to enable numerical issues to be resolved. This allowed for interactions to be included in the model that are representative of the characteristics of those who are correctly enumerated individuals.

Collapsing of main variables

The collapsed levels for legal partnership status are:

- never married
- married and civil partnerships
- divorced, separated, or widowed

The collapsed levels for ethnicity are:

- English
- White other
- Mixed and Other
- Indian, Pakistani, Bangladeshi, and Other Asian
- African, Caribbean, and Black other

The collapsed levels for household size are:

- one
- two
- three
- four
- five plus

The collapsed levels for tenure are:

- owns outright
- owns with mortgage and owns with mortgage or partial ownership
- rents – council
- rents – housing association
- rents – private
- rents – employer or relative or other or free

The collapsed levels for age groups are:

- 0 to 17 years
- 18 to 21 years
- 22 to 24 years
- 25 to 29 years
- 30 to 34 years
- 35 to 39 years
- 40 to 44 years
- 45 to 49 years
- 50 to 54 years
- 55 to 59 years
- 60 to 64 years
- 65 to 69 years
- 70 years and over

Household population wrong location model of individuals

Following the model selection approach outlined in Section 2: Methodology and steps, the final model is as follows, where the main effects in the model include:

- age groups
- accommodation type
- activity last week
- address one year ago
- legal partnership status
- born in the UK
- household size
- hard to count index
- observed return rate
- individual ethnicity
- region
- relationship
- self-completion
- sex
- student
- tenure

Two factor interactions include:

- age group by sex
- activity last week by relationship

As per the previous model selection, collapsing of the main variables was needed. Many combinations of collapsed variables were tried to reduce the prediction error of the chosen model, using knowledge from the correct enumeration modelling exercise.

Collapsing of main variables

The collapsed levels for legal partnership status are:

- never married
- married and civil partnerships
- divorced, separated or widowed

The collapsed levels for person ethnicity:

- English
- White other
- Mixed
- Indian, Pakistani, Bangladeshi
- Other Asian
- African, Caribbean, Black other
- Other

Small communal establishment undercoverage of individuals

To estimate census undercoverage of individuals in small communal establishments, individuals enumerated within the CCS area are considered for modelling. Because of the small size of the small communal population, we assume there is no overcoverage, as it would be very difficult to estimate given the population size. Selecting a robust model to estimate undercoverage of individual in small communal establishments was challenging because of the small number of observations in the coverage survey.

The final model is as follows, where the main effects included in the model are:

- collapsed age-sex groups
- collapsed regions
- hard to count index

Decisions to establish possible collapsing of the categories are made based on the information generated through univariate and bivariate analysis of the variables.

The collapsed levels for female age-sex group are:

- 0 to 69 years
- 70 to 79 years
- 80 years and over

The collapsed levels for male age-sex group are:

- 0 to 79 years
- 80 years and over

The collapsed levels for region are:

- North East, North West, and Yorkshire and The Humber
- East Midlands, West Midlands, and East England
- London
- South East and South West
- Wales

4 . Variables before collapsing levels

The individual main variables before collapsing with original levels are as follows.

Accommodation

- Detached
- Semi detached
- Terraced
- Purpose built flat block or tenement
- Converted or shared house
- Another converted building
- Commercial building
- Caravans or other mobile temporary

Age group

- 0 to 2 years
- 3 to 7 years
- 8 to 17 years
- 18 to 21 years
- 22 to 24 years
- 25 to 29 years
- 30 to 34 years
- 35 to 39 years
- 40 to 44 years
- 45 to 49 years
- 50 to 54 years
- 55 to 59 years
- 60 to 64 years
- 65 to 69 years
- 70 to 74 years
- 75 to 79 years
- 80 to 84 years
- 85 to 89 years
- 90 years and over

Activity last week

- Working
- Unemployed or economically inactive
- Student
- Retired
- Long-term sick or disabled
- Home or family carer
- Not required

Address one year ago

- Same address one year ago
- Different address one year ago

Born in the UK

- Born in the UK
- Born elsewhere

Household size

- One
- Two
- Three
- Four
- Five
- Six plus

Hard to count index

- One
- Two
- Three
- Four
- Five

Legal partnership status

- Never married
- Married
- Separated
- Divorced
- Widowed
- In a registered civil partnership
- Separated but in a legal civil partnership
- Ex-civil partnership now legally resolved
- Surviving partner of civil partnership
- Not required

Individual ethnicity

- English
- Irish
- Gypsy
- Roma
- White other
- White and Black Caribbean
- White and Black African
- White and Asian
- Other Mixed
- Indian
- Pakistani
- Bangladeshi
- Chinese
- Asian other
- Caribbean
- African
- Black other
- Arab
- Other

Region

- East of England
- East Midlands
- Inner London
- North East
- North West
- Outer London
- South East
- South West
- Wales
- West Midlands
- Yorkshire and The Humber

Any relationship

- Some related
- None related

Self-contained

- Self-contained
- Not self-contained

In full time education

- In full time education
- Not in full time education

Tenure

- Owns outright
- Owns with mortgage
- Part owns or part rents (shared ownership)
- Rents – council
- Rents – private
- Rents – employer
- Rents – relative
- Rents – other
- Rent – free

Initial contact

- Online questionnaire
- Paper questionnaire

Hard to count score

Continuous variable.

Observed return rates

Continuous variable.

The household main variables before collapsing are as follows.

Accommodation

- Detached
- Semi detached
- Terraced
- Purpose built flat block or tenement
- Converted or shared house
- Another converted building
- Commercial building
- Caravans or other mobile temporary

Household size

- One
- Two
- Three
- Four
- Five
- Six plus

Hard to count index

- One
- Two
- Three
- Four
- Five

Household ethnicity

The household ethnicity variable is defined as the ethnicity of the household reference person.

- White
- Mixed
- Asian
- Black
- Other

Household structure

The household structure variable is defined by the following broad categories.

- Single adult only
- Single adult with children
- 2 adults no children (related)
- 2 adults no children (unrelated)
- 2 adults with children (related)
- 3 plus adults no children (related)
- 3 plus adults no children (unrelated)
- 3 plus adults with children (related)
- Children only
- Other

Within these broad categories, sub-categories are created where appropriate, stratified by the age and sex of the adults in the household, for a total of 62 categories.

Region

- East of England
- East Midlands
- Inner London
- North East
- North West
- Outer London
- South East
- South West
- Wales
- West Midlands
- Yorkshire and The Humber

Any relationship

- Some related
- None related

Self-contained

- Self-contained
- Not self-contained

Tenure

- Owns outright
- Owns with mortgage
- Part owns or part rents (shared ownership)
- Rents –council
- Rents – housing association
- Rents – private
- Rents – employer
- Rents – relative
- Rents – other
- Rent – free

Hard to count score

Continuous variable.

Observed return rates

Continuous variable.

5 . Additional information

Parameter estimates from each model can be accessed on request. Please email census.customerservices@ons.gov.uk.

6 . Cite this methodology

Office for National Statistics (ONS), released 8 November 2022, ONS website, methodology article, Model selection for coverage estimation for Census 2021 in England and Wales