

SOCIAL GRADE APPROXIMATION FOR THE 2021 CENSUS

This document summarises how the 2021 Census Social Grade Approximation can be applied to other data sources

Corrine Moy, Sergiy Korniyev, Barry Leventhal

December 2023

1. Introduction and Key Points

The purpose of this document is to explain how the Social Grade approximation developed for the 2021 Census can be applied to other external datasets such as market research surveys.

The approximation employs an advanced predictive modelling technique that was developed on a large-scale media research survey. The following section provides an overview of this development, while a model development report is available at: <https://www.mrs.org.uk/resources/approximated-social-grade-on-the-2021-census>

The models were built using the R software system and the resultant algorithms have been published together with this document. All the R files and datasets required to run the models are included in a companion folder, 'Social Grade Approximation - 2021 Census.zip' – see Section 8. For comprehensive instructions on loading and running R software/scripts, see: <https://www.r-project.org/>

After the models were applied to the 2021 Census data, a small number of final processing steps were applied, such as recodes for grading special categories. These processing steps are defined in Section 4, for users who wish to create the Census Approximated Social Grade (ASG) variable on their own data, as exactly as possible.

2. Development of the Social Grade Approximation for the 2021 Census

The Social Grade classification is a widely used measure among marketing and market research practitioners. It is a powerful classification that broadly differentiates groups of people with regards to some attitudes and behaviours, as well as discriminating well on the types of goods and services consumed. The classification groups people into 6 categories A, B, C1, C2, D and E. Social Grade is often recorded about survey respondents by interviewers using a series of questions.

Since these questions are not included in the Census questionnaire, the Market Research Society (MRS) develops models to assign respondents to a Social Grade category, based on a restricted set of questions common to both the data on which the model is built and the Census. Such models were first developed for the 2001 Census, and have been similarly developed for the 2011 and 2021 Censuses.

The models have been developed using the PAMCo (Publishers Audience Measurement Company) UK media currency survey in 2021. In 2001 & 2011, its antecedent, the National Readership Survey (NRS) was used.

To develop the models, the PAMCo/NRS survey data was used which corresponded to the same year as the relevant Census. Firstly questions common to both the census and PAMCo/NRS survey were identified. The survey data was then reduced to include only those who were assigned their own Social Grade (rather than that of the household). A number of models were tested to determine which best predicted the respondents Social Grade.

In 2011, a CHAID model was eventually used to allocate respondents to their most likely category. The NRS survey data was then reduced to include only those who were assigned their own Social Grade (rather than that of the household). The final model was based on the Standard Occupational Code 2010 code, Employment status, qualification, Tenure and whether they work Full time, Part time or not working. Modelling the data achieved an overall allocation rate of 73% (for the 5-way classification AB, C1, C2, D, E), with individual categories being correctly allocated at a rate of between 66% and 77%.

In 2021, a different approach was taken for modelling Social Grade onto the 2021 Census. The agreed aim was to predict the Social Grade of the Household Reference Person (HRP), the Census equivalent to the Chief Income Earner (CIE) in market research. Other members of the household are then assigned the Social Grade of the CIE, in accordance with MRS social grading guidelines (see: "Occupation Groupings: A Job Dictionary" from The Market Research Society).

Two separate models were developed for application to 2021 Census data:

Model 1 - for HRPs who are not retired, built using occupation details and other variables.

Model 2 – for retired HRPs, built using household attributes and individual characteristics such as qualifications, but not using occupation details.

The reason for creating a separate model for retired HRPs is that in market research a retired person is graded on their previous occupation prior to retirement (as long as they receive a pension from it). On the other hand, the Census captures the current or most recent job that the person is/was doing, which may lead to a different social grade being assigned.

A dataset from PAMCo was used for the analysis and modelling; this data contained records for all CIEs who took part in the PAMCo Survey during 2019, which was the last year the survey ran prior to the Covid Pandemic.

To develop the models, we used an improved methodological approach to the CHAID models in 2011 – an Extreme Gradient Boosting (XGBoost) algorithm. Like CHAID, XGBoost is also based on decision tree learning but can better cope with more complex decisions for smaller subgroups of Census respondents. Other advanced analytics techniques were first used for selecting the best predictors of social grade and the most important variables were then included in the XGBoost model.

In total, 21,912 PAMCo survey CIEs were included in the analysis for the 2021 census model development. The dataset was restricted to respondents from England & Wales, as Census data was not available for Scotland, due to delays in processing. This dataset was split into subsamples for training (75%) and testing (25%).

Modelling the data achieved an overall train/test allocation rate of **64%/60%** - for the 6-way classification A, B, C1, C2, D, E. The individual Social Grade categories were correctly allocated at a rate of between 61%/57% and 69%/69%. For the 4-way classification (AB, C1, C2, DE) the correct allocation rate was **72%/70%**.

The models for the 4-way classification (AB, C1, C2, DE) achieved acceptable levels of accuracy. Hence it was agreed to code the 4-way classification onto the Census.

3. Questions included in the models

There are a set of questions that feed in to deriving the Social grade approximation.

For the 2001 census, the key questions used in the model to predict social grade included Employment Status, working status the Standard Occupational Code of the respondent's job (or previous job under certain circumstances), size of establishment, gender, qualification and tenure.

For the 2011 census, the size of establishment was not asked in the census, which affected what information was available to use. Those questions which were found to be common to both the Census and the NRS data (on which the models are based) in 2011 were Marital Status, Working status, Employment Status, Qualifications, Ethnicity, Gender, Number of cars in household, Tenure and Number of adults in household.

For the 2021 the common questions between the PAMCo survey and the Census are: Standard Occupational Code 2020 code, Age, Gender, Number of adults in Household (16+), Number of people in Household, Number of cars in household, Ethnic group, Tenure, Highest qualification, Working status, Employment Status and Marital Status.

These questions were coded up into categories that were also common across the PAMCo survey and the Census.

The final models used the following questions:

Model 1 (non-retired HRPs) : Standard Occupational Code 2020, Employment status, Qualification, Tenure, Working status, Ethnicity, Car ownership, Household size, Gender

Model 2 (retired HRPs) : Qualification, Tenure, Ethnicity, Car ownership, Household size, Gender

4. Additional Census processing steps to create Approximated Social Grade (ASG)

After attaching the predicted social grade from the model to each HRP, a quality assurance process was carried out. This resulted in the following additional processing steps and recodes:

Step 1: Limit ASG population to Household Reference Persons aged 16-64

For HRPs aged 65 or more, the predicted Social Grade was found to be less accurate:

If the HRP is aged 15 and under, or aged 65 and over:

APPROX_SOCIAL_GRADE = 8 ('does not apply')

Step 2: Recode ASG for HRPs who are full-time students

In line with MRS social grading guidelines:

Else if ECONOMIC_ACTIVITY = 'Full-time Student'

APPROX_SOCIAL_GRADE = 2 (C1)

Step 3: Recode ASG for HRPs who are unemployed or have never worked

In line with MRS social grading guidelines as closely as possible in the Census:

Else if Has_Ever_Worked = 'Yes but not in last 12 months' or 'Never Worked'

and ECONOMIC_ACTIVITY = 'Seeking Work' or 'Economically Inactive'

APPROX_SOCIAL_GRADE = 4 (DE)

5. Application of the Models to external datasets

We have created a package of resources to allow users to allocate Social Grade codes to external datasets, that contain records/details for CIEs e.g. market research surveys

There are seven steps to create this coding in your dataset:

1. Prepare an input CSV file according to the prescribed data map shown below
 - "Social_Grade_Test_10.csv" is an example file
2. Place this file into a folder containing the R code 'SocGrade Modelling.R'
3. Load R library for XGBOOST
4. Edit the input file name in the R code accordingly [**line 31 in code**]
5. Run the entire R code. The Social Grade variable is called "SocGrade4"
6. Merge results from the output file "predicted_your_file_name.csv" to your original database by ID
7. Apply additional processing steps to replicate Census ASG variable (defined in Section 4).

6. Data map

Data must be coded to match the exact codes shown below. [Any deviations will result in incorrect allocation of Social Grade]. Names shown in brackets are the variable names used in the input CSV file – these are required for the R code to run correctly.

The first column must contain an ID for each record, so that the output Social Grade value can be matched back to your data source.

SOC2020 (soc2020) – this is the Standard Occupational Code for 2020.

4-digit SOC2020 codes (unit groups) must be recoded into 15 codes to create a new variable "**soc2020_short**". The recoding scheme is detailed in an accompanying file "soc2020_short_all.csv" file.

The R code will do this recoding automatically. You simply need to take the following steps:

- replace missing SOC codes with 0
- the following 10 codes did not exist in the actual PAMCo data. If they occur within your dataset, they should be replaced with the relevant code shown below:

Original code	Replacement code
1233	1232
2317	2316
2435	2434
2464	2463
5412	5411
5419	5422
5421	5422
5433	5432
6117	6116
8144	8143

NUMBER OF ADULTS IN HOUSEHOLD (numb_adults_hh):

Coding should be as follows:

- | | |
|----|-------------|
| 1 | None |
| 2 | One |
| 3 | Two |
| 4 | Three |
| 5 | Four |
| 6 | Five |
| 7 | Six |
| 8 | Seven |
| 9 | Eight |
| 10 | Nine |
| 11 | Ten or more |

TOTAL NUMBER OF PERSONS IN HOUSEHOLD (numb_in_hh):

Coding should be as Integer values : 1 to 15

[In the event of a household having more than 15 persons, recode to 15]

GENDER (gender_1_male):

Categories generally collected in surveys would include the following categories:

- | | |
|---|-------------------|
| 1 | Male |
| 2 | Female |
| 3 | In another way |
| 4 | Prefer not to say |

These should be recoded to a binary variable "**gender_1_male**", where:

- | | |
|---|-------|
| 1 | Male |
| 0 | Other |

HOUSEHOLD TENURE:

Create the following codings:

- | | |
|---|----------------|
| 1 | Owned outright |
|---|----------------|

- 2 Owned with mortgage/loan
- 3 Rented from council
- 4 Rented from someone else
- 5 Rent free
- 6 Don't know/ Refused

Then recode to 4 binary variables:

hh_tenure_1_own	Owned outright [code 1]
hh_tenure_2_mortgage	Owned with mortgage [code 2]
hh_tenure_3_4_rented	Rented from council or someone else [codes 3,4]
hh_tenure_5_rent_free	Rent free [code 5]

where

1 = falls into this group

0 = does not fall into this group

Number of cars in household (cars_recoded):

Coding should be as follows:

- 1 1 car
- 2 2 cars
- 3 3 or more cars
- 0 no cars

ETHNICITY (ethnicity):

Coding should be as follows:

- 1 White
- 2 Mixed Race
- 3 Black - Caribbean
- 4 Black - African
- 5 Black - Other
- 6 Indian
- 7 Pakistani
- 8 Bangladeshi
- 9 Chinese
- 10 Asian - other
- 11 Any other ethnic group
- 12 Refused

HIGHEST QUALIFICATION OBTAINED (highest_qual):

Coding should be as follows:

- 1 Postgraduate qualification (e.g. PHD, MSc, MBA, PGCE)
- 2 Professional qualifications of degree status
- 3 First degree (e.g. BA, BSc)
- 4 Other qualification requiring A levels (or equivalent) for entry (e.g. NVQ Level 4 or 5, HNC, HND, BTEC Higher Level)
- 5 One or more A levels (or equivalent - including Scottish Highers)
- 6 Other Level 3 qualification (e.g. BTEC, NVQ, Advanced GNVQ, ONC, OND etc.)
- 7 5 + GCSE grades A*-C or 4-9 OR 5+ O level (passes)/CSEs at grade 1
- 8 1-4 Levels/CSE/GCSEs (any grades) NVQ Level 1, Foundation GNVQ, Basic Skills or Other Level 1 Qualification
- 9 Trade apprenticeship
- 10 None of these/not applicable
- 11 Refused
- 12 Don't Know

WORKING STATUS:

Create the following codings:

- 1 Full-time (30+ hours per week)
- 2 Part-time (8-29 hours per week)
- 3 Part-time (under 8 hours per week)
- 4 Unemployed
- 5 Retired from full-time job
- 6 Not employed
- 7 Student
- 8 Refused

Recoded into binary variables:

working_status_4_unemployed	Unemployed [code 4]
working_status_5_retired	Retired [code 5]
working_status_6_not_employed	Not employed [code 6]
working_status_7_student	Student [code 7]

where

1=falls into this group

0 = does not fall into this group

EMPLOYMENT STATUS (employment_status):

Create the following codings:

- 1 Self-employed - with employees - large establishments (25+)
- 2 Self-employed - with employees - small establishments (1-24)
- 3 Self-employed - without employees
- 4 Employees - Managers - Large establishments (25+)
- 5 Employees - Managers - Small establishments (1-24)
- 6 Employees - Foremen and supervisors - Manual
- 7 Employees - Foremen and supervisors - Non-manual
- 9 Employees - Employees not elsewhere classified, including all armed forces
- 10 Not employed
- 11 Don't know/refused

Recoded to the new binary variables:

employment_status_1_2_se_w_e	Self-employed - with employees [codes 1,2]
employment_status_3_se_wo_e	Self-employed - without employees [code 3]
employment_status_4_5_6_7_e_supervisor	Employed – Supervisor [codes 4,5,6,7]

where

1=falls into this group

0 = does not fall into this group

The final list of input variables for modelling is:

soc2020
numb_adults_hh
numb_in_hh
gender_1_male
hh_tenure_1_own
hh_tenure_2_mortgage
hh_tenure_3_4_rented
hh_tenure_5_rent_free
cars_recoded

ethnicity
 highest_qual
 working_status_4_unemployed
 working_status_5_retired
 working_status_6_not_employed
 working_status_7_student
 employment_status_1_2_se_w_e
 employment_status_3_se_wo_e
 employment_status_4_5_6_7_e_supervisor

7. Variables in the Output File

Column 1: ID (carried forward from input file)

Column 2: Predicted 4-way Social Grade, coded as:

- 1 - AB Higher and intermediate managerial/administrative/professional occupations
- 2- C1 Supervisory, clerical and junior managerial/administrative/professional occupations
- 3 - C2 Skilled manual occupations
- 4 - DE Semi-skilled and unskilled manual occupations; unemployed and lowest grade occupations 5 - -
- 8 - Does not apply

8. Files included in 'Social Grade Approximation - 2021 Census.zip'

SocGrade_Modelling.R	R code for creating Social Grade approximation
modeling_SG_2021_4_short_NotRetired_3_S	XGBoost model for Non-retired CIEs
modeling_SG_2021_4_short_Retired_3_S	XGBoost model for Retired CIEs
soc2020_short_all	Look-up table to recode SOC2020 into 15 groups
Social_Grade_Test_10	Test input file
predicted_SG4_Social_Grade_Test_10	Test output file