# SOCIAL GRADE ALLOCATION TO THE 2021 CENSUS

This document summarises how Social Grade has been mapped to the 2021 Census

Sergiy Kornyev, Barry Leventhal, Corrine Moy
August 2023

## Social Grade Approximation for the Census 2021

### Introduction

Social Grade classification is a commonly 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. These questions would be too long to include in the Census questionnaire, and instead a model can be used 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.

An approximation to the Social Grade classification was applied to respondents in the 2001 and 2011 Census data. It was modelled based on the data from the National Readership Survey[1]. To develop the 2011 model, the National Readership Survey was used which corresponded to the same year as the census. Firstly questions common to both the census and NRS survey were identified. The NRS 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. A CHAID model was eventually used to allocate respondents to their most likely category 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%.

### Approach to Social Grade model development for 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.

[1] For more information about the National readership survey please see http://www.nrs.co.uk/
[2]For more information about the PAMCo survey please see https://pamco.co.uk/

In 2016, the Publishers Audience Measurement Company (PAMCo) took over responsibility for producing the published media currency from NRS Ltd. The National Readership Survey subsequently ceased to operate and a new PAMCo[2] survey replaced it.

A dataset from PAMCo was used for the Social Grade 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.

In 2011, the census social grade model was built using Chi-square Automatic Interaction Detection (CHAID) decision tree analysis. An improved methodological approach to was adopted in 2021 using a newer, high-performance, Extreme Gradient Boosting (XGBoost) algorithm. 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.

### Identifying common questions

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. Details are in Appendix B.

## The Final Models

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

The chosen models were derived using the XGBboost algorithm, executed in R.

The weighted and un-weighted distributions of the Social Grade groups are shown in Table 1.1 for the train and test datasets (tables 1.1.1 and 1.1.2 respectively).

**Table 1.1.1** Train Data - Distribution of Social Grade for Original and Predicted (weighted and un-weighted)

| | Original | | Predicted | | Original | | Predicted | |
|---|---|---|---|---|---|---|---|---|
| **Grade** | Weighted | % | Weighted | % | Unweighted | % | Unweighted | % |
| **AB** | 6138 | 25.6 | 5454 | 22.7 | 4538 | 27.6 | 3982 | 24.2 |
| **C1** | 7207 | 30.0 | 7943 | 33.1 | 4966 | 30.2 | 5537 | 33.7 |
| **C2** | 4522 | 18.8 | 5119 | 21.3 | 2742 | 16.7 | 3117 | 19.0 |
| **DE** | 6136 | 25.6 | 5487 | 22.9 | 4188 | 25.5 | 3798 | 23.1 |
| **Total** | 24003 | 100 | 24003 | 100 | 16434 | 100 | 16434 | 100 |

**Table 1.1.2** Test Data - Distribution of Social Grade for Original and Predicted (weighted and un-weighted)

| | Original | | Predicted | | Original | | Predicted | |
|---|---|---|---|---|---|---|---|---|
| **Grade** | Weighted | % | Weighted | % | Unweighted | % | Unweighted | % |
| **AB** | 2132 | 26.7 | 1837 | 23.0 | 1559 | 28.5 | 1318 | 24.1 |
| **C1** | 2282 | 28.6 | 2605 | 32.7 | 1610 | 29.4 | 1853 | 33.8 |
| **C2** | 1531 | 19.2 | 1706 | 21.4 | 904 | 16.5 | 1049 | 19.1 |
| **DE** | 2030 | 25.5 | 1826 | 22.9 | 1405 | 25.6 | 1258 | 23.0 |
| **Total** | 7974 | 100 | 7974 | 100 | 5478 | 100 | 5478 | 100 |

The percentage of cases that were correctly allocated to the associated segment is below in Table 1.2:

**Table 1.2**. Proportion of cases correctly allocated to segment

| Social Grade | Train<br>Percentage correctly allocated to segment | Test<br>Percentage correctly allocated to segment |
|---|---|---|
| **AB** | 71.8% | 71.2% |
| **C1** | 70.5% | 65.1% |
| **C2** | 75.3% | 75.3% |
| **DE** | 69.7% | 68.3% |
| **Overall (4-way)** | 71.5% | 69.5% |

Table 1.3 shows the distribution of derived Social Grade produced by the model against actual PAMCo Social Grade. The values in bold correspond to the above table.

**Table 1.3.1** Train data **-** Percentage of actual social grade predicted by model

| Predicted Social Grade | AB | C1 | C2 | DE |
|---|---|---|---|---|
| AB | **71.8%** | 12.8% | 2.3% | 2.2% |
| C1 | 21.3% | **70.5%** | 13.6% | 13.1% |
| C2 | 5.4% | 10.4% | **75.3%** | 15.1% |
| **DE** | 1.4% | 6.3% | 8.8% | **69.7%** |
| Total | 100% | 100% | 100% | **100%** |

**Table 1.3.2** Test data **-** Percentage of actual social grade predicted by model

| Predicted Social Grade | AB | C1 | C2 | DE |
|---|---|---|---|---|
| AB | **71.2%** | 16.3% | 2.0% | 2.5% |
| C1 | 21.9% | **65.1%** | 12.7% | 14.0% |
| C2 | 5.7% | 11.9% | **75.3%** | 15.2% |
| **DE** | 1.2% | 6.8% | 10.0% | **68.3%** |
| Total | 100% | 100% | 100% | **100%** |

The above tables are shown by non-retired vs. retired in Appendix A.

## Appendix A : Modelled Social Grade by Non-retired vs. Retired

### Distribution of weighted and unweighted profiles of Social Grade, by Retired vs Non-retired status

| | | Original | | Predicted | | Original | | Predicted | |
|---|---|---|---|---|---|---|---|---|---|
| | **Grade** | Weighted | % | Weighted | % | Unweighted | % | Unweighted | % |
| **Non-retired** | | | | | | | | | |
| | **AB** | 6082 | 26.1 | 5459 | 23.4 | 4002 | 28.8 | 3566 | 25.7 |
| | **C1** | 7251 | 31.1 | 7874 | 33.7 | 4365 | 31.4 | 4766 | 34.3 |
| | **C2** | 4715 | 20.2 | 5234 | 22.4 | 2532 | 18.2 | 2792 | 20.1 |
| | **DE** | 5290 | 22.7 | 4771 | 20.4 | 2993 | 21.5 | 2768 | 19.9 |
| | **Total** | 23338 | 100 | 23338 | 100 | 13892 | 100 | 13892 | 100 |
| | | | | | | | | | |
| **Retired** | | | | | | | | | |
| | **AB** | 2187 | 25.3 | 2080 | 24.1 | 2095 | 26.1 | 1962 | 24.5 |
| | **C1** | 2238 | 25.9 | 2358 | 27.3 | 2211 | 27.6 | 2364 | 29.5 |
| | **C2** | 1338 | 15.5 | 2037 | 23.6 | 1114 | 13.9 | 1701 | 21.2 |
| | **DE** | 2876 | 33.3 | 2164 | 25.0 | 2600 | 32.4 | 1993 | 24.9 |
| | **Total** | 8639 | 100 | 8639 | 100 | 8020 | 100 | 8020 | 100 |

### Allocation to Social Grade categories split by Retired vs Non-retired status

**Distribution of cases in actual and predicted Social Grade**

| **Retired status** | **NRS Social Grade** | **Predicted Category** | | | | **% correctly allocated to Social Grade category** |
|---|---|---|---|---|---|---|
| **Non-retired** | | AB | C1 | C2 | D | |
| | AB | 73% | 23% | 3% | 1% | 73% |
| | C1 | 13% | 76% | 8% | 3% | 76% |
| | C2 | 1% | 11% | 80% | 8% | 80% |
| | DE | 0% | 8% | 13% | 79% | 79% |
| | Overall | | | | | 74% |
| | | | | | | |
| **Retired** | | | | | | |
| | AB | 67% | 18% | 12% | 4% | 67% |
| | C1 | 16% | 46% | 20% | 18% | 46% |
| | C2 | 7% | 21% | 57% | 15% | 57% |
| | DE | 6% | 23% | 20% | 52% | 52% |
| | Overall | | | | | 55% |

**Appendix B. Question definitions**

PAMCo questions were coded so that they were comparable with those in the Census.

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

Within the modeling process, the 2, 3 and 4-digit versions of SOC2020 were tested. The 2 and 3 digit versions caused noticeable decreases in the models' quality. Using the 4-digit version (unit groups) led to overly complicated models and generated considerable noise.

Through a feature engineering stage, we developed a clustering of unit groups (4-digit), based on the probability of belonging to the 6 original groups (A-E) for each code. This clustering process produced 15 clusters of unit groups. All 4-digit SOC codes were recoded to create a new variable "**soc2020_short**", which represents the 15 clusters. The final recoding scheme is detailed in an accompanying file "soc2020_short_all.csv" file.

Note: the following 10 codes did not exist in the actual PAMCo data. They were replaced with the nearest available code:

| 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)**:

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):**

Integer values : 1 to 15
*In the event of a household having more than 15 persons, recode to 15*

**GENDER (gender):**

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

Recoded to a binary variable "**gender_1_male**"

1 Male
0 Other

**HOUSEHOLD TENURE (hh_tenure):**

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

Recode to binary variables:

| | |
|---|---|
| **hh_tenure_1_own** | Owned outright [code 1] |
| **hh_tenure_2_mortage** | 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):**

1 ONE
2 TWO
3 THREE
4 NONE

Recode 4 to 0 into a new variable "**cars_recoded**"

**ETHNICITY (ethnicity):**

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):**

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 (working_status):**

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):**

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