Consumer Data Research Centre

An ESRC Data Investment

Data Profile: CDRC Modelled Ethnicity Proportions (LSOA Geography)

Introduction

This dataset combines historical electoral roll and consumer register data (on surnames and locations) between 1997 and 2020, with a special aggregated metric derived from ONS data[1] which lists the most frequently selected second-level ethnicity category for most surnames.

Scale and Extent

Field	Value
Data Provider	CDRC
Analytical Units	Person
Data Format	CSV
Temporal Extent	1997-2020, years
Geographical Extent	United Kingdom
Variables	Ethnicity category, year
Observations	Geography code, population proportion (1 = 100%)

Citation Information

The following statement should be included when citing the use of this dataset:

"The data for this research have been provided by the Consumer Data Research Centre, an ESRC Data Investment, under project ID CDRC [Project Number], ES/L011840/1; ES/L011891/1"

Data Classification and Access Summary

These data are classified as Safeguarded and are available only upon approved application. To make an initial application, search for "Modelled Ethnicity Proportions" at https://data.cdrc.ac.uk/

Content

CSV files, one for each of the ethnicity groups. Each row contains the LSOA/DZ/SOA ID, followed by the proportion of the population that is believed to be of that ethnicity (based on surname analysis) rounded to the nearest 0.5%.

To create the data a slight aggregation on the second-level ethnicity categories is carried out.

We then aggregate by 2011 LSOA (Data Zones for Scotland/SOAs for Northern Ireland). Category populations less than 5 are set to 0. The results are then divided by the total population and rounded to the nearest 0.005 (i.e. 0.5%).

Our aggregated ethnicity categories used are (with codes used in the data files):

- WBR - White: British (including English/Welsh/Scottish/Northern Irish)
- WIR - White: Irish
- WAO - White: Any Other
- ABD - Asian/Asian British: Bangladeshi

[1] The data was derived as part of a ESRC-funded project "Ethnicity Estimator" - Virtual Microdata Laboratory project number: 0000013. It is a diagnostic table resulting from the application of CDRC algorithms. The aggregate data was provided by ONS within the VML.

Consumer Data Research Centre

An ESRC Data Investment

Data Profile: CDRC Modelled Ethnicity Proportions (LSOA Geography)

- ACN - Asian/Asian British: Chinese
- AIN - Asian/Asian British: Indian
- APK - Asian/Asian British: Pakistani
- AAO - Asian/Asian British: Any Other
- BAF - Black/Black British: African
- BCA - Black/Black British: Caribbean
- OXX - Any Other Ethnic Group (including Mixed; Black/Black British: Any Other; Arab; All Other Ethnicities; &c.)

A value of 0 indicates there is no measurable total population for this LSOA/DZ/SOA and year combination. These values generally only occur for the first few years and in only a small number of LSOA/DZ/SOA areas.

Totals may not add up to 1.000 (100%) because of rounding, but also because of an Unknown ethnicity category which a small proportion of names are assigned to.

Details of the model method can be found in this paper: https://doi.org/10.1371/journal.pone.0201774 - the model used is EE-A6, on a deterministic (not probabilistic) basis.

Applicants should be mindful that this is a dataset of ethnicity categories and not one showing migration, citizenship, nationality or country of origin.

Novelty

The dataset is unique, as it is compiled from data which is highly detailed and covers almost the full population of the United Kingdom. However, it is modelled, not measured.

Quality

The data is modelled, based on the most common ethnicities stated for particular surnames (regardless of location) in England/Wales. It is not actual measured data for the populations. Because this particular source is only from England/Wales, we would expect marginally less accurate results from Scotland/Northern Ireland.

Representation and Bias

The full population is inferred using data processing methods at CDRC. This includes imputation of data from previous years, where this is considered reasonable, based on a household composition analysis.

Related Datasets

Ethnicity Estimator software: search for "Ethnicity Estimator" at https://data.cdrc.ac.uk/ - this software allows you to run a similar analysis, on a names list of your choosing (subject to approval).

Consumer Data Research Centre

An ESRC Data Investment

Data Profile: CDRC Modelled Ethnicity Proportions (LSOA Geography)

Data Triangulation: data sources used to establish provenance

Source	Variable	Spatial granularity of comparator	Temporal granularity of comparator	Note(s)
ONS Census	Most common ethnicities by forename, most common ethnicities by surname.	None (single composition result for all England and Wales, per name)	2011	Used for the most common forenames and surnames.
Onomap (which itself uses various sources, e.g. phonebooks)	Forename/surname pairs used in the Onomap model, from which some ethnicity based grouping are identified.	Global	Various 2000- 2012.	Used for less common forenames and surnames, to augment the main data.
Linked Consumer Registers	Names and Addresses	Individual People in Households	Every year from 1997 to 2020	Household composition imputation has been applied by CDRC.