Research update on population estimates by country of birth for England and Wales

Development and quality assurance of an admin-based country of birth dataset for 2021, using Driver and Vehicle Licensing Agency (DVLA) and NHS data. These are not official statistics.

Contact:
Population and Migration Research and Development team
pop.info@ons.gov.uk
+44 1329 444661

Release date:
10 November 2026

Next release:
To be announced

Table of contents

1. Overview

2. Assigning open-text place of birth data to countries

3. Assessing the assignment quality

4. Building an admin-based country of birth dataset

5. Coverage gaps on the admin-based country of birth dataset

6. Assessing the quality of the admin-based country of birth dataset

7. Future developments

8. Definitions

9. Related links

10. Cite this page

1 . Overview

We have been developing a way to use administrative data to create population estimates by country of birth. In this supporting methodology article, we describe how open-text place of birth data from the Driver and Vehicle Licensing Agency (DVLA) and the NHS Personal Demographic Service (PDS) can be cleaned and coded to individual countries. We discuss how the quality assurance methods used (record level and aggregate comparisons with Census 2021) show the success of the research so far, and how the final data could be used as an input for producing population estimates by country of birth.

Our first migration statistics from the Census 2021 for England and Wales, published on 2 November 2022, are our best estimate of the population of England and Wales by country of birth and passports held. A census provides the most comprehensive snapshot of society every 10 years. However, more timely and frequent statistics are needed to better understand population change.

Historically, we produced population estimates by country of birth and nationality using our Annual Population Survey (APS). However, these were discontinued in 2022 because of quality issues relating to the data used to weight the non-UK-born population. We are exploring alternative methods to produce these estimates, including using:

- administrative data (the focus of this article)
- a rolled-forward cohort component method based on Census 2021 and 2022 Scottish Census data
- our Transformed Labour Force Survey (TLFS)

We published provisional estimates that roll forward Census 2021 data to produce estimates of the non-UK-born population, split by EU+ and non-EU+, in Section 8: Estimates of the non-UK-born population of our Long-term international migration, provisional: year ending December 2025 bulletin. We are investigating extending this method to include estimates for individual countries of birth.

We are also assessing the quality of the TLFS data for both country of birth and nationality, to determine what could be published following the planned transition to the online-first TLFS, which is expected in 2027. We will provide an update on our research in spring 2027.

2 . Assigning open-text place of birth data to countries

This research uses two data sources, the Driver and Vehicle Licensing Agency (DVLA) drivers' data and the NHS Personal Demographic Service (PDS). The DVLA data comprises data on all driving licenses currently held. The Personal Demographic Service (PDS) is the central system held by NHS England that holds the demographic details of users of health and care services in England, Wales and the Isle of Man.

For this research, we focus on 2021 to allow comparisons with Census 2021 data. We currently only have place of birth information on a limited number of admin datasets; the DVLA data for 2021 and the PDS data for 2021 and 2022. We are currently working with both the DVLA and the NHS to establish on-going supplies of these data to routinely produce these statistics.

The place of birth information on both the DVLA and PDS data is open text, often without validation (for example, from paper applications). This means that country information is included in the data, but so are:

- regions
- states
- provinces
- counties
- towns
- cities
- villages
- hospitals (on the PDS)
- combinations of the above geographies
- spelling mistakes and typos
- optical character recognition errors

There are 118,000 different places of birth on the DVLA data and 619,000 on the PDS data. The open-text place of birth data on the PDS and DVLA data were similar enough that we were able to develop one approach for both datasets. We built upon the principles developed as part of our automated text coding for Census 2021 and used several "place to country" lookup data to identify the likely country for a particular place. The following lookup datasets were used:

- our National Statistics Country Classification methodology (used for Census 2021)
- city and state lookups from CountryStateCity
- the GeoNames geographical database
- a county list for the UK, including abbreviations and old names
- a custom tuning list developed during our research

Place names with their respective country were extracted from these geographical datasets to create multiple lookup lists. Both the admin data and the lookup lists were standardised. This included removing non-standard characters (brackets, hyphens, commas) and removing diacritics (accents) from letters. The latter is necessary because the admin data do not contain diacritics, while the lookup data does.

We developed an assignment strategy that was based on three main assumptions.

Our three assumptions

Prioritising place names hierarchically in the admin data

We use a geographic hierarchy, from country to state, province, or region to county to city, town, or village. This means that, for example, "Florida" would be assigned to "America" (where it is a state), rather than "Uruguay (where it is a city). Similarly, "Canterbury, New Zealand" would be assigned to "New Zealand", rather than "England", as we look for a match to a country before looking for a match to a town.

Prioritising UK places

Without further information, it is assumed that a place name will be within the UK, so while "Boston, USA" would be assigned to "America", "Boston" would be assigned to "England".

A place name is more likely to be spelt correctly than incorrectly

With this assumption, "Parys" would be assigned to "South Africa" (where it is a city), rather than "France" (as a misspelling of "Paris").

Matching methods

We used four matching methods to assign a potential country of birth to a place name on the admin data using the lookup lists. Each matching method was run for each lookup list, which can result in multiple potential assignments for each person. To resolve these and assign a single country of birth to a person, we used a priority order based on our three assumptions and our evaluation of the quality of the linkages.

Exact matches

An exact match between the place of birth string on the admin data and the lookups attributes a country to the records where the place is spelled correctly. For example, "London" would match to "England".

Partial string matches

This identifies where the place of birth string contains the place on the lookup data. For example, "Wales" is within the string "Newprt South Wales". This attributes a country to a record where at least part of the place is spelled correctly.

Levenshtein edit distance

This is a type of approximate string matching. Levenshtein edit distance is a metric for computing the similarity between two different strings. This is done by calculating the number of letter deletions, insertions or substitutions to move from one string to another.

A match with an edit distance of one has been used throughout this linkage. For example, "Alesbury" is one away from "Aylesbury" and would be identified as a match, while "Alyesbury" is two away and would not be matched.

Jaro similarity

This is also a type of approximate string matching. Jaro similarity is a calculated score for the difference between two strings from 0 (no similarity) to 1 (both strings are the same). Like the Levenshtein edit distance, the score includes deletions, insertions and substitutions. It also includes transpositions, by matching characters that appear in a different order in the two strings.

A match with a score of 0.9 or higher has been used throughout this linkage. With this method, the similarity score between "Alyesbury" and "Aylesbury" is 0.9630, so it would be identified as a match by the Jaro linkage.

3 . Assessing the assignment quality

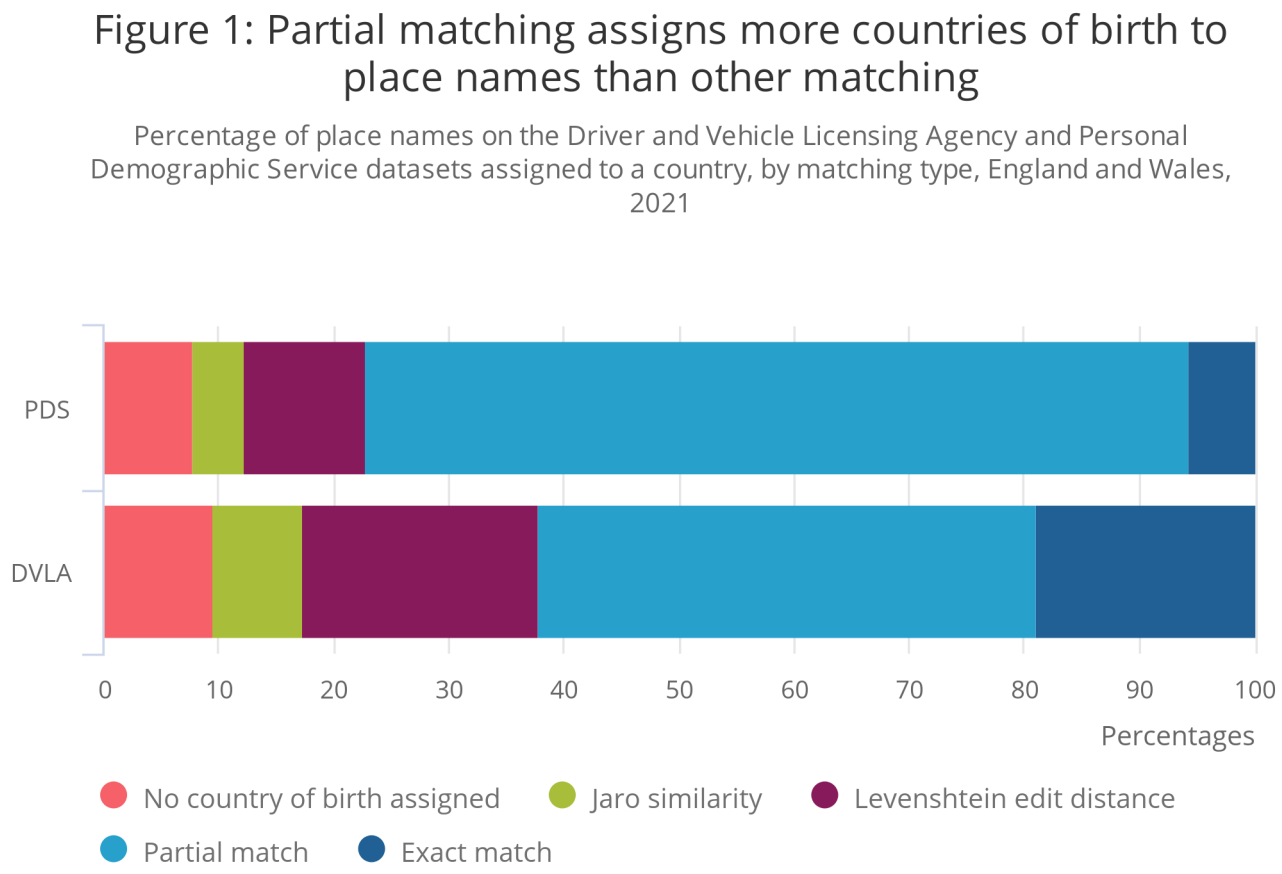
The quality of the country of birth assignment can be evaluated by assessing the proportion of assignments made and their precision.

Evaluating the accuracy of the country of birth assignment

To evaluate the accuracy of the country of birth assignment, we look at the proportion of assignments made by each matching method. A higher proportion of exact matches means a more accurate assignment and higher quality. We can look at the proportion of assignments made by each matching method by both the number of place names that are assigned to a country of birth and the number of people who are assigned to a country of birth.

Figure 1: Partial matching assigns more countries of birth to place names than other matching

Percentage of place names on the Driver and Vehicle Licensing Agency and Personal Demographic Service datasets assigned to a country, by matching type, England and Wales, 2021



Source: Place of birth analysis from the Office for National Statistics

Notes:

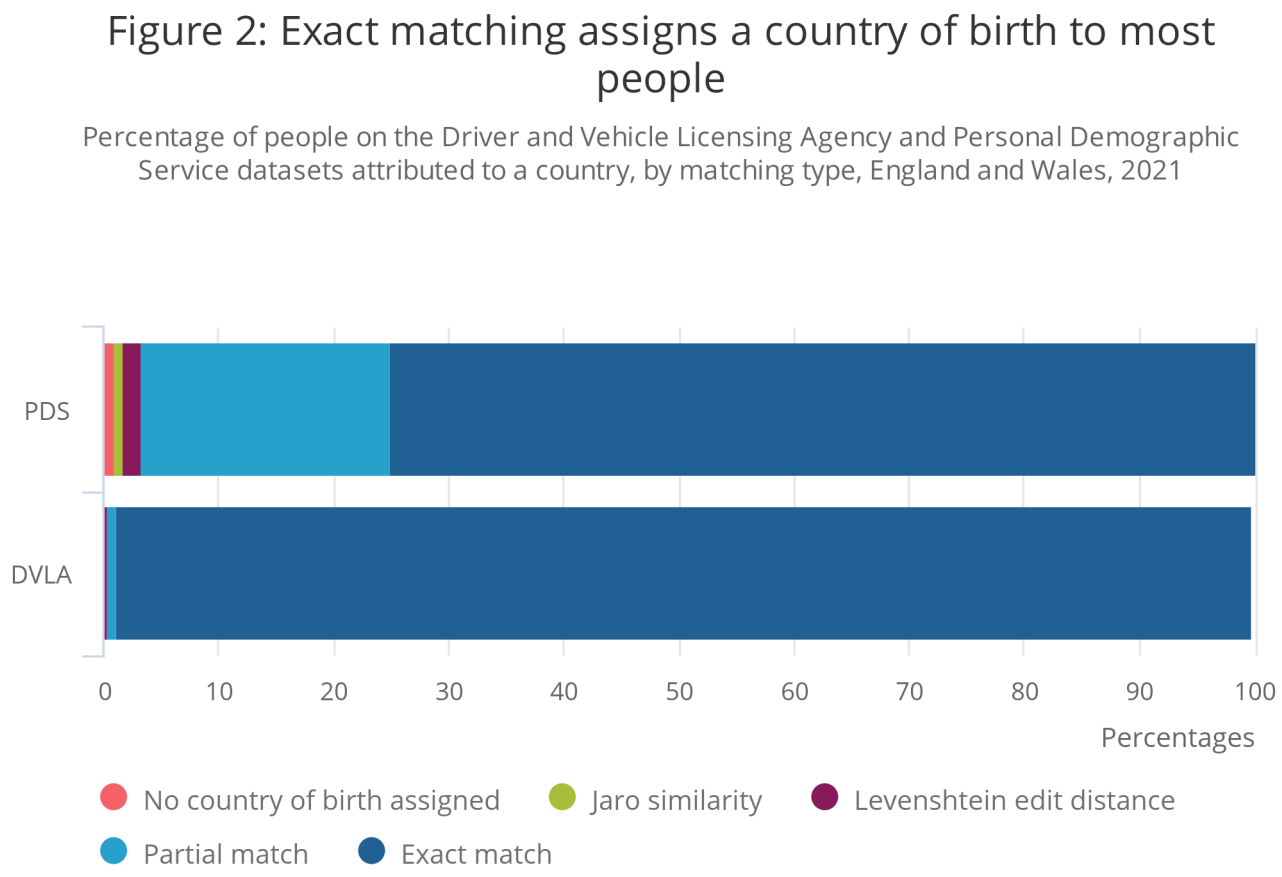
1. Records that do not contain a place of birth are not included.

Just under one-tenth of place names on both datasets were not assigned to a country of birth. Exact matches account for nearly a fifth (19.1%) of the Driver and Vehicle Licensing Agency (DVLA) linkages, but only 5.6% of those on the NHS Personal Demographic Service (PDS). On the PDS, nearly three-quarters (71.5%) of the successful matches are from partial matches, compared to under half (43.2%) of the matches on the DVLA data.

The Levenshtein and Jaro matching methods assign more countries on the DVLA data than the PDS (28.3% compared with 15.2%), indicating that miss-spellings are less prevalent in the PDS data.

Figure 2: Exact matching assigns a country of birth to most people

Percentage of people on the Driver and Vehicle Licensing Agency and Personal Demographic Service datasets attributed to a country, by matching type, England and Wales, 2021



Source: Place of birth analysis from the Office for National Statistics

Notes:

1. Records that do not contain a place of birth are not included.

At the person level, exact matches assign a country of birth to 98.9% of people on the DVLA data and to 75.0% of those on the PDS data. Partial matching assigns 21.6% of people a country of birth on the PDS data but less than 1% of those on the DVLA data. In terms of people, the Levenshtein and Jaro matches assign a country of birth for less than 2% of those on either dataset.

Less than 1% of people cannot be assigned a country of birth (0.9% for PDS and 0.1% for DVLA).

Evaluating the precision of the country of birth assignment

To evaluate the precision of the country of birth assignment, we linked the admin data at the person-level to Census 2021 data, using our Demographic Index. For this evaluation, the country of birth on the Census is assumed to be correct, and the assignment of a country of birth to a place of birth on the admin data is validated against the Census data.

Figure 3: Exact matching has the highest precision for assigning country of birth

Driver and Vehicle Licensing Agency and Personal Demographic Service assignment precision, by matching method and National Statistics Country Classification country grouping, England and Wales, 2021

Figure 3 shows the agreement rates against Census 2021 for the different matching methods for the DVLA data and the 2021 PDS data. The same analysis for PDS 2022 has been run but is not shown, as there is no appreciable difference between the match rates for 2021 and 2022. This analysis also uses the sub-group and group country groupings from our National Statistics Country Classification to test the assignment precision at higher geographic levels.

The Census 2021 agreement rates for the different linkage types are broadly consistent across both the DVLA and PDS data. For the exact, partial, and Levenshtein matches, the DVLA data marginally outperforms the PDS data, although for the Jaro matches the PDS data performs slightly better. The trend is also for the more approximate matches to have lower agreement rates with Census 2021 than the more exact matches.

For each data source and matching method, the Census 2021 agreement rate for group is higher than sub-group, which itself is higher than for country. This trend is more pronounced for the Levenshtein and Jaro linkages, where there is at least a 7.5 percentage point increase in the Census 2021 agreement rate from the country to group level.

One of the reasons that country has the lowest match rate is because of places of birth assigned to "not otherwise specified" regions. For example, at the country level, "Punjab" would be assigned to "Asia (except Middle East) Not Otherwise Specified", which would not match to Census 2021 data, where "Punjab" is not an option (it would be India or Pakistan). At the group level, where both the admin data and Census 2021 data would be "Middle East and Asia", they would match.

4 . Building an admin-based country of birth dataset

We developed an admin-based country of birth (ABCOB) dataset for England and Wales by joining the Driver and Vehicle Licensing Agency (DVLA) and NHS Personal Demographic Service (PDS) data with our Statistical Population Dataset (SPD) v5.1 as a population spine. Census 2021 data have also been included in the build, as a source for country of birth, to maximise coverage. There are 56.3 million records on the ABCOB dataset (compared with 59.6 million in the Census 2021 data); 97.2% of these (54.7 million records) have a country of birth assigned.

Where there is only one country of birth for a record on the SPD, or it is the same across the two or three possible data sources (DVLA, PDS and Census 2021), this country is attributed to the record. However, for 29.5% of the records (16.6 million) on the SPD, the country of birth conflicts across the three different datasets. This section describes the processes and methods for selecting one country of birth per record, where there are conflicting countries in the data.

As we are using two years' worth of PDS data, a person may have two different countries of birth on the PDS data. In these cases, the latest country of birth provided (that is, that for 2022) is selected.

Standardising "not otherwise specified"

There are several broad "not otherwise specified" categories in the National Statistics Country Classification that may be used when it is not possible to identify a more specific option. Many of these have been used to assign the place of birth data in the DVLA and PDS data, in particular "Great Britain not otherwise specified" and "United Kingdom not otherwise specified". Other "not otherwise specified" categories are general ones, such as for Africa, Europe, and the Middle East, while others are for countries that no longer exist, such as the Union of Soviet Socialist Republics (USSR), Yugoslavia and Czechoslovakia.

Each "not otherwise specified" area has a group of countries that could be a broad match, such as England, Scotland and Wales for "Great Britain not otherwise specified", or Slovakia and Czechia for "Czechoslovakia not otherwise specified". Where there is a consistent match for a "not otherwise specified" area in one or more of the remaining datasets, that more specific country is assigned to the record.

Most frequent country

Where a country has not yet been selected and where it is possible to identify the most frequent country on the datasets, this country is attributed to the record.

Dataset prioritisation

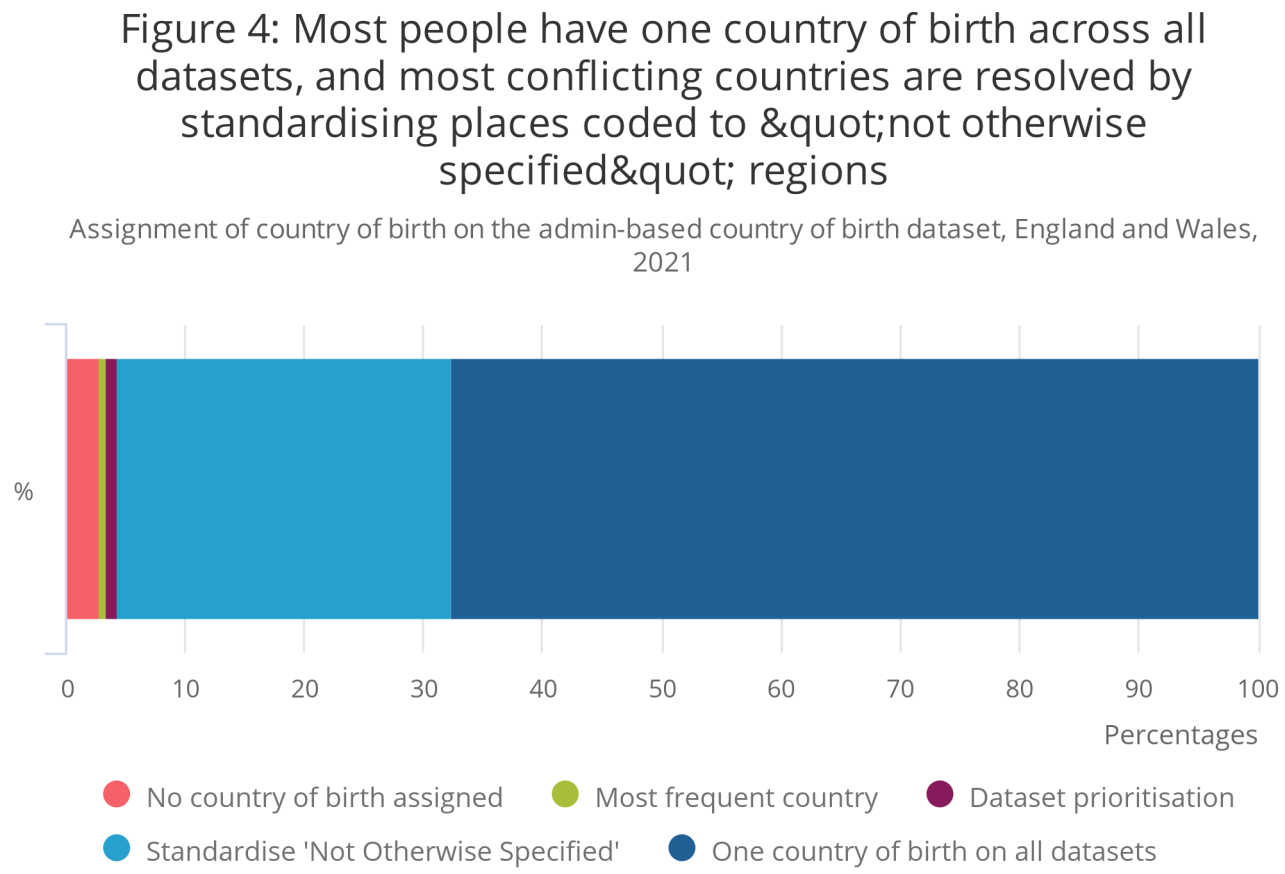
The remainder of conflicts are cases where there is a different country of birth on all the datasets that contain a country of birth for a person. To assign a country of birth for them, we used a dataset priority order:

1. Census 2021
2. PDS
3. DVLA

We selected this order after creating six separate ABCOB datasets for each of the ways that the three datasets can be ordered. The order that gave the closest agreement with Census 2021 data, after both the record and aggregate level comparison, was Census 2021, then PDS, then DVLA.

Figure 4: Most people have one country of birth across all datasets, and most conflicting countries are resolved by standardising places coded to "not otherwise specified" regions

Assignment of country of birth on the admin-based country of birth dataset, England and Wales, 2021



Source: Admin-based country of birth dataset from the Office for National Statistics

Figure 4 shows how a country of birth was assigned to a person in the ABCOB dataset. For over two-thirds (67.7%), there was just one country assigned to the person across the three component datasets. Standardising the "not otherwise specified" response assigned a country to a further 28.0% of people.

This left 1.5% of people for whom there was more variation in the datasets, meaning the most frequent country or dataset prioritisation had to be used. Just 2.8% of those on the ABCOB dataset could not be assigned a country of birth.

5 . Coverage gaps on the admin-based country of birth dataset

On the admin-based country of birth (ABCOB) dataset, 1.6 million records (2.8%) do not have a country of birth. To understand the representativeness of the dataset, we compare the profile of these residuals against the whole population.

Figure 5: Younger age bands are most over-represented in the residuals, while older ones are under-represented, and the London region is highly over-represented

Admin-based country of birth dataset residuals relative bias to Census 2021 data, by age band and region, England and Wales, 2021

Downloac

The coverage gaps are skewed towards the lower age bands. This is especially the case for those aged 0 to 16 years, who comprise nearly a third (32.6%) of the residuals; they are over-represented in the residuals by 65.7%. In contrast, those in the older age bands (apart from those aged 80 years and over) are under-represented in the residuals.

This skew towards the lower age bands will be affected by the fact that coverage for the Driver and Vehicle Licensing Agency (DVLA) drivers' data starts at 17 years of age, so data for 0- to 16-year-olds is only coming from the NHS Personal Demographic Service (PDS) and Census 2021 data.

There are also large regional differences in the residuals, with all regions being under-represented in the residuals, except for London. With a relative bias of 90.5%, the proportion of those in the residuals in London is almost double that of the Census 2021 proportion.

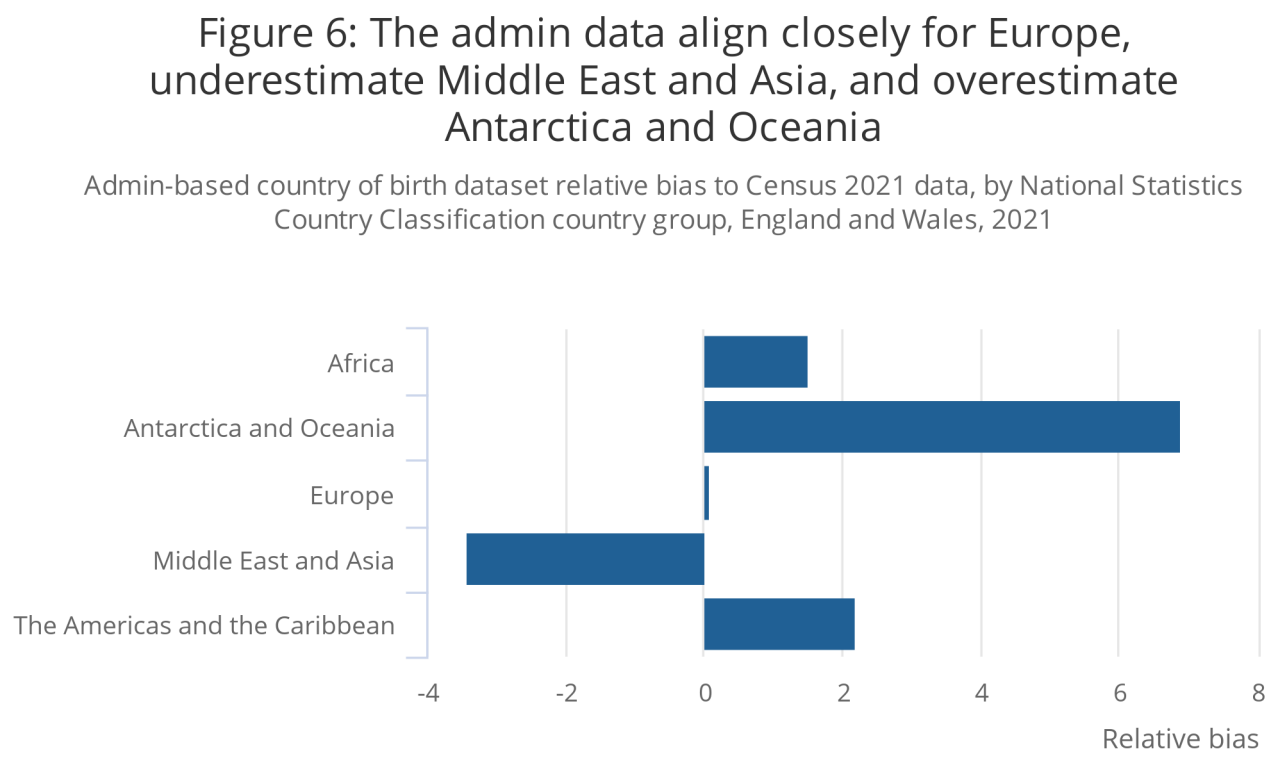
6 . Assessing the quality of the admin-based country of birth dataset

As part of developing the methods for resolving conflicting countries of birth, we made record-level comparisons between the admin-based country of birth (ABCOB) dataset and Census 2021 data. However, these comparisons can only be made where the Census 2021 data has been linked to the Statistical Population Dataset (SPD). Of the 54.7 million people in the ABCOB dataset with a country of birth, 50.5 million were successfully linked to Census; 4.2 million people could not be linked to the Census and therefore no direct comparison could be made.

To address this gap, we compared aggregate statistics from the ABCOB dataset with Census 2021 data.

Figure 6: The admin data align closely for Europe, underestimate Middle East and Asia, and overestimate Antarctica and Oceania

Admin-based country of birth dataset relative bias to Census 2021 data, by National Statistics Country Classification country group, England and Wales, 2021



Source: Admin-based country of birth dataset from the Office for National Statistics

Figures 6 and 7 show the relative bias of the ABCOB data by National Statistics Country Classification country groupings, compared with Census 2021 data, and illustrate where there is both under- and over- representation in the admin data.

The proportion of those born in Antarctica and Oceania is 6.9% higher in the ABCOB data than in Census 2021 data, while the proportion of those born in the Middle East and Asia is 3.4% lower in the admin data than in Census 2021 data.

Figure 7: The admin data for Eastern Asia have the lowest relative bias to Census 2021 data, while Eastern Europe has the highest

Admin-based country of birth dataset relative bias to Census 2021 data, by National Statistics Country Classification country sub-group, England and Wales, 2021

Downloac

At the country sub-group level, there are more pronounced variances in the relative bias between the ABCOB data and Census. The UK's ABCOB percentage is close to Census 2021 data, at negative 0.5%, but many of the other European regions are over-represented in the admin data, particularly Eastern Europe, at 10.4%. At negative 6.3%, Eastern Asia is the most under-represented region in the ABCOB data.

We cannot yet reach conclusions from our current research about why these differences exist for specific country sub-groups. However, as these data are for 2021, some of these differences may be related to the coronavirus (COVID-19) pandemic and whether a person is present in the admin data compared with being present for Census Day, 21 March 2021. Carrying out further research for more recent years of data may help us further understand these differences.

There are some differences in the relative bias across the English regions and Wales. This is particularly noticeable for the Americas and the Caribbean, which is over-represented in all regions, apart from East of England, where it is under-represented (negative 14.2%). This is because of the number of United States Air Force personnel stationed at the Royal Air Force Lakenheath and Mildenhall, in Suffolk. They are present in the Census 2021 data, but not the admin data.

7 . Future developments

This research has proven that a country of birth can be assigned to a very high proportion of people using open-text admin data. For most people this is using exact or partial matching, with high levels of agreement with Census 2021 data. However, the analysis of the residuals and relative bias, as well as the coverage gaps in our Statistical Population Dataset (SPD), shows that there are non-random coverage issues in the admin-based country of birth (ABCOB) dataset.

Therefore, to produce high-quality population estimates by country of birth, we will need to use statistical estimation methods to address these issues. The next stages of this research will focus on identifying and using appropriate estimation methods to address these coverage and bias issues.

In addition to the estimation work, we will continue to improve the development of the ABCOB dataset, where possible. This may include improvements to the country matching methodology and exploring other admin data sources that could provide information about the population by country of birth. We will also monitor developments in the ways artificial intelligence could be used to further improve the assignment of country of birth to free text data.

Alongside the admin data research, we are investigating extending the rolled-forward cohort component method to include estimates for individual countries of birth. We are also assessing the quality of Transformed Labour Force Survey (TLFS) data for both country of birth and nationality, to determine what could be published following the planned transition to the online-first TLFS, which is expected in 2027.

8 . Definitions

Administrative data

Collections of data maintained for administrative reasons. For example, registrations, transactions, or record keeping. They are used for operational purposes, and their statistical use is secondary. These sources are typically managed by other government bodies.

National Statistics Country Classification

This is the harmonised framework used by National Statistics producers. It provides a three-level hierarchy of geographies. The lowest level is country, including all countries and some further breakdowns of dependencies as countries (there are 283 in this category).

The country categories are aggregated based on geographic proximity within broadly defined continents to produce the country sub-group and country group levels. There are 31 main sub-groups and five main groups. For this research, we have not used the "Other categories" group, which covers "in the air" and "at sea".

Relative bias

Relative bias shows how much an estimate differs from the reference source, expressed as a percentage. A positive value means that the estimate is higher than the actual value (an overestimate), while a negative value means that it is lower than the actual value (an underestimate).

Precision

Precision answers the question: of all the record pairs that the linkage process identified as matches, how many were actually correct matches?

Statistical Population Dataset

A single, coherent dataset that forms the basis for estimating the size of the resident UK population. It is produced by linking records across multiple administrative data sources and applying a set of inclusion and distribution rules.

9 . Related links

International migration, England and Wales: Census 2021
Statistical bulletin | Released 2 November 2022
International migration, including country of birth, passports held and year of arrival, Census 2021 data.

Long-term international migration, provisional: year ending December 2025
Statistical bulletin | Released 21 May 2026
UK long-term international migration to year ending December 2025. Includes breakdowns by age, sex, nationality and reason. Official statistics in development.

10 . Cite this page

Office for National Statistics (ONS), released 10 November 2026, ONS website, supporting methodology article, Research update on population estimates by country of birth for England and Wales