Office for National Statistics

Companies House persons of significant control to Census 2021 linkage report

Methods used to link Companies House persons of significant control to Census 2021 deterministically.

Contact:
Data Linkage and Integration Hub
Linkage.Hub@ons.gov.uk

Release date:
22 August 2024

Next release:
To be announced

Table of contents

1 . Main points

- Companies House persons of significant control (PSC) snapshot (January 2024) was linked to Census 2021 using deterministic matchkeys; a PSC is someone who owns or controls a company, sometimes called “beneficial owners”.
- This linkage was carried out to enable analysis into the characteristics of beneficial owners, with a particular focus on the intersectionality of sex and ethnicity.
- The deterministic linkage had a match rate of 50.31% of all PSC records, or 66.22% of PSC records which contained forename, surname, month and year of birth information (so were of sufficient quality to link) and had a country of residence of England, Wales, UK or Great Britain (so were reasonably likely to be included on census).
- The estimated precision for the linkage was 96.93% and estimated recall was 97.93%, indicating that the linkage was of good quality, given quality issues with the PSC data.
- Bias analysis was performed on this linkage, showing that the following groups were underrepresented in the linked data: those with a London postcode, those with nationality other than British, English or Welsh, and those born before 1930 or after 2009.
- The final linkage product provided to analysts was fully anonymised.

2 . Purpose of the linkage

This linkage was carried out following the linkage of Companies House persons of significant control (PSC) snapshot data (October 2021) to 2011 Census. This new linkage uses the most recent PSC data available at the time the project started (January 2024) and Census 2021.

This linkage allows persons of significant control business-level data to be linked to Census 2021 person-level data, enabling investigation into the characteristics of PSC (with a particular focus on sex, ethnicity and their intersection) and whether they may experience barriers to maintain active limited companies in England and Wales. Analysis will be carried out using the sex and ethnicity variables from census.

3 . Methods

Census 2021

The census, administered by the Office for National Statistics (ONS), happens every 10 years and gives us a picture of all the people and households in England and Wales. The most recent Census Day for England and Wales was on Sunday 21 March 2021.

The dataset used contained personal data identifiers in-the-clear, created for the purpose of linkage. The cut of census used for linkage was prior to any editing, imputation, estimation and statistical disclosure control. The dataset contained 58,623,712 records.

Missingness was examined and found to be low, indicating the high quality of the census data. Missingness rates for the main linkage variables are shown and compared with the corresponding rates for the persons of significant control (PSC) data in Table 1.

Companies House (PSC)

The PSC details snapshot is a company-based file comprising details for each PSC of every UK company (unless exempt), limited liability partnership (LLP) and Societas Europaea (SE) companies (European PLC) registered at Companies House.

The January 2024 version was used for linkage and contained 12,369,301 records. Each record has an associated company reference number, which allows onward joining to company level information.

After cleaning, the dataset contained 12,358,931 records. Of these, 1,551,629 records were found to have missing or null information for forename, surname, month and year of birth, meaning that they were not suitable for linkage. A further 1,417,097 records had a country of residence other than England, Wales, UK or Great Britain, so were identified as unlikely to be covered by census. These records are likely to be for limited companies, rather than individuals. Therefore, there was a total of 9,390,205 records which were of a reasonable quality and likely to be covered by census. These records were all included in the data linkage process, but provided an early indication that the total match rate was unlikely to be high.

The sex variable in the PSC dataset has been derived from the title variable and is therefore not completely comparable with the sex variable from census. Titles which are specific for males, for example, Mr, Lord and Sir, were given a derived sex value of male. Titles which are specific for females, for example, Mrs, Miss and Lady, were given a derived sex vale of female. Titles which are gender neutral, for example, Doctor and Reverend, were given a null derived sex value.

Table 1 shows that missingness is high for derived sex in PSC, and fairly high for forename, surname, month and year of birth. Missingness for postcode is lower, however, this is still problematic as postcode is necessary for linkage. Day of birth was not available in PSC, which also causes difficulty with linkage as it means full date of birth information is not present in the data.

Table 1: Missingness comparison for PSC 2024 (UK) and Census 2021 (England and Wales)

	Census 2021	PSC 2024 (all data)
Forename	0.44%	12.56%
Surname	0.41%	12.56%
Sex (Census 2021), Derived sex (PSC)	0.27%	23.36%
Month of birth	0.14%	12.56%
Year of birth	0.15%	12.57%
Postcode	0.01%	8.49%

Source: Companies House persons of significant control to Census 2021 linked data from the Office for National Statistics

Pre-processing

Both files were standardised and cleaned so that the variables used for linkage were in a consistent format, and their variables had suffixes added to represent their corresponding sources.

Both datasets underwent cleaning, steps included:

- case standardising string variables
- standardisation of nulls, missingness and blanks
- standardising date formatting
- removing non-alphabetical characters, and spaces where appropriate
- removing titles from name variables
- splitting forename and surname components into separate variables, for example, if a forename was “Joseph James” the forename one (first component) variable would be “Joseph” and the forename two (second component) variable would be “James”
- splitting postcode components into separate variables (sector, district and area) and appending region using postcode
- dropping records which were exact duplicates (PSC only)
- creating unique ID variable for each record (PSC only)

Deterministic linkage

The two datasets were deterministically linked using 55 matchkeys, shown in Table 2. Each matchkey consists of a set of rules or criteria that must be met to make a link. To account for expected errors in the data, the criteria are loosened on different linkage variables. Matchkeys are applied hierarchically, starting at the strictest matching criteria, and gradually become looser.

Where one PSC ID linked to multiple census IDs on different matchkeys the links made on the lowest (strictest) matchkey were kept. Where one PSC ID linked to multiple census IDs on the same matchkey all links were discarded, as it was not possible to know which one was most likely to be correct. These clusters likely occurred because of duplicate records in the census.

While multiple responses (where there is more than one census return for a person at an address) are removed from the census microdata file, the data still contain duplicate entries where multiple returns have been made for a person at different addresses. Examples include but are not limited to higher education students, children of divorced parents, and those staying at temporary addresses for work reasons and interpersonal relationships. Such individuals may appear more than once at different addresses. For census processing, such cases are dealt with in the estimation process.

There are many cases of one census ID linked to multiple PSC IDs in the linked data, which is to be expected as some individuals will appear numerous times in the PSC data if they are involved in more than one company.

The deterministic linkage produced 6,218,142 linked record pairs, with 6,218,142 distinct PSC IDs and 4,085,739 distinct census IDs (shown in Table 3). Therefore, the deterministic linkage has a match rate of 50.31% of the PSC data used, or 66.22% if disregarding the records which were of insufficient quality to link and unlikely to be covered by census.

Table 2: Matchkey descriptions and number of links per matchkey for the PSC 2024 (UK) and Census 2021 linkage (England and Wales)

Matchkey	Description	Number of links
1	Concordant first components of forename and surname, sex, month of birth, year of birth and postcode.	2,464,783
2	Concordant first components of forename and surname, month of birth, year of birth and postcode.	265,058
3	First components of forename and surname (PSC) concordant with first components of resident forename and surname (census). Concordant sex, month of birth, year of birth and postcode.	5,552
4	Fuzzy agreement of forename 1, 2 or 3 or full forename variable. Fuzzy agreement of surname 1 or full surname variable. Concordant month of birth, year of birth and postcode.	5,217
5	Forename contained within forename (first components). Concordant first components of surname, month of birth, year of birth and postcode.	36,658
6	Concordant first components of forename and surname, sex, month of birth and year of birth. PSC postcode concordant to census work postcode, one year ago postcode or alternative postcode.	431,809
7	First components of forename and surname (PSC) concordant with first components of resident forename and surname (census). Concordant sex, month of birth and year of birth. PSC postcode concordant to census work postcode, one year ago postcode or alternative postcode.	933
8	First components of forename and surname transposed. Concordant month of birth, year of birth and postcode.	4,072
9	First components of forename and surname (PSC) and first components of resident forename and surname (census) transposed. Concordant sex, month of birth, year of birth and postcode.	20
10	Surname contained within surname (first components). Concordant first components of forename, month of birth, year of birth and postcode.	23,428
11	Concordant nickname to first components of forename. Concordant first components of surname, month of birth and year of birth. PSC postcode concordant to census postcode, work postcode, one year ago postcode or alternative postcode.	34,535
12	Transposed first components of forename and surname. Concordant, sex, month of birth and year of birth. PSC postcode concordant to census work postcode, one year ago postcode or alternative postcode.	660
13	First components of forename concordant to first components of middle name. Concordant surname, sex, month of birth and year of birth. PSC postcode concordant to census work	2,174
14	Levenshtein distance less than 2 on forename. Concordant first components of surname, month of birth, year of birth and postcode.	13,875
15	Jaro-Winkler of less than 0.9 on forename. Concordant first components of surname, sex, month of birth, year of birth and postcode.	2,918
16	Levenshtein distance less than 2 on first components of forename. Concordant first components of surname, month of birth, year of birth and postcode.	315
17	Jaro-Winkler of less than 0.9 on first components of forename. Concordant first components of surname, sex, month of birth, year of birth and postcode.	51
18	Census first components of forename concordant to PSC first components of middle name. Concordant first components of surname, month of birth, year of birth and postcode.	9,114
19	Substrings on first components of surname (first to seventh letters must match). Concordant first components of forename, month of birth, year of birth and postcode.	1,074
20	Levenshtein distance less than 2 on surname. Concordant first components of forename, month of birth, year of birth and postcode.	18,526

21	Levenshtein distance less than 2 on forename and surname. Concordant month of birth, year of birth and postcode.	842
22	Concordant first components of first name, surname, year of birth and postcode.	26,335
23	Levenshtein distance less than 2 on year of birth. Concordant first components of first name, surname, month of birth and postcode.	19,250
24	Levenshtein distance less than 2 on forename. Concordant first components of surname, month of birth and year of birth. PSC postcode concordant to census work postcode, one year ago postcode or alternative postcode.	36,909
25	Jaro-Winkler of less than 0.9 on forename. Concordant first components of surname, sex, month of birth and year of birth. PSC postcode concordant to census work postcode, one year ago postcode or alternative postcode.	1,260
26	Levenshtein distance less than 2 on first components of forename. Concordant first components of surname, month of birth and year of birth. PSC postcode concordant to census work postcode, one year ago postcode or alternative postcode.	676
27	Jaro-Winkler of less than 0.9 on first components of first name. Concordant first components of surname, sex, month of birth and year of birth. PSC postcode concordant to census work postcode, one year ago postcode or alternative postcode.	15
28	Concordant first components of first name and surname, sex and year of birth. PSC postcode concordant to census work postcode, one year ago postcode or alternative postcode.	4,453
29	Transposed first components of forename and surname. Concordant month of birth and year of birth. PSC postcode concordant to census work postcode, one year ago postcode or alternative postcode.	77
30	First components of forename concordant to first components of middle name. Concordant first components of surname, month of birth and year of birth. PSC postcode concordant to census work postcode, one year ago postcode or alternative postcode.	61
31	Jaro-Winkler of less than 0.9 on forename. Concordant first components of surname, month of birth and year of birth. PSC postcode concordant to census work postcode, one year ago postcode or alternative postcode.	106
32	First components of forename and surname (PSC) concordant with first components of resident forename and surname (census). Concordant month of birth and year of birth. PSC postcode concordant to census work postcode, one year ago postcode or alternative postcode.	58
33	Concordant first components of forename and surname and year of birth. PSC postcode concordant to census work postcode, one year ago postcode or alternative postcode.	431
34	First components of forename and surname (PSC) concordant with first components of resident forename and surname (census). Concordant year of birth. PSC postcode concordant to census work postcode, one year ago postcode or alternative postcode.	21
35	Concordant first components of forename and surname, sex, month of birth, year of birth and postcode sector.	208,995
36	Concordant first components of forename and surname, sex, month of birth and year of birth. PSC postcode sector concordant to census work postcode sector, one year ago postcode sector or alternative postcode sector.	50,328
37	Concordant first components of forename and surname, month of birth, year of birth and postcode sector.	21,143
38	Concordant first components of forename and surname, month of birth and year of birth. PSC postcode sector concordant to census work postcode sector, one year ago postcode sector or alternative postcode sector.	4,712
39	Concordant nickname to first components of forename. Concordant middle name, first components of surname, month of birth, year of birth and postcode sector.	1,513

40	Concordant nickname to first components of forename. Concordant middle name, first components of surname, month of birth and year of birth. PSC postcode sector concordant to census work postcode sector, one year ago postcode sector or alternative postcode sector.	1,432
41	Substrings on first components of forename (first to seventh letters must match). Concordant first components of surname, month of birth, year of birth and postcode sector.	767
42	Concordant first components of forename and surname, middle name, sex, month of birth, year of birth and postcode district.	123,128
43	Concordant first components of forename and surname, middle name, sex, month of birth and year of birth. PSC postcode district concordant to census work postcode district, one year ago postcode district or alternative postcode district.	19,629
44	Concordant first components of forename and surname, month of birth, year of birth and postcode district. PSC postcode district concordant to census work postcode district, one year ago postcode district or alternative postcode district.	170,020
45	Concordant first components of forename and surname, month of birth and year of birth.	28,330
46	Substrings on first components of forename (first to sixth letters must match). Concordant middle name, first components of surname, month of birth, year of birth and postcode district.	202
47	Concordant first components of forename and surname, middle name, sex, month of birth and year of birth. PSC postcode area concordant to census work postcode area.	87,265
48	Concordant first components of forename and surname, middle name, sex, month of birth and year of birth. PSC postcode area concordant to census work postcode area, one year ago postcode area or alternative postcode area.	39,852
49	Concordant first components of forename and surname, middle name, month of birth, year of birth and postcode area.	413,619
50	Concordant first components of forename and surname, month of birth, year of birth and postcode area.	611,537
51	Concordant first components of forename and surname, middle name, month of birth, year of birth and region.	388,358
52	Concordant first components of forename and surname, month of birth, year of birth and region.	634,502
53	First component of forename (PSC) concordant to first component of resident forename (census). First component of surname (PSC) concordant to first component of resident surname (census). Concordant middle name, sex, month of birth, year of birth and region.	1,142
54	First component of forename (PSC) concordant to first component of resident forename (census). First component of surname (PSC) concordant to first component of resident surname (census). Concordant middle name, sex, month of birth and year of birth. PSC region concordant to census work region, one year ago region or alternative region.	311
55	First component of forename (PSC) concordant to first component of resident forename (census). First component of surname (PSC) concordant to first component of resident surname (census). Concordant middle name, month of birth and year of birth. PSC region concordant to census work region, one year ago region or alternative region.	91

Source: Companies House persons of significant control to Census 2021 linked data from the Office for National Statistics

Notes

1. The sex variable used for these matchkeys in census is sex, whereas in PSC it is derived sex, as explained in the Companies House (PSC) section of the report.

Table 3: Total deterministic links and ID counts for the PSC 2024 (UK) and Census 2021 linkage (England and Wales)

Total deterministic links	6,218,142
Distinct PSC IDs in the linked data	6,218,142
Distinct census IDs in the linked data	4,085,739

Source: Companies House persons of significant control to Census 2021 linked data from the Office for National Statistics

4 . Quality information

Clerical review

The standard approach to estimate error in the linked data is to perform clerical review (manual checking) on a sample of links and rejected record pairs, to estimate the number of true positives (correct links), false positives (incorrect links) and false negatives (missed matches). In linkage, there is a trade-off between two types of error: precision and recall.

Precision is a measure of the accuracy of the matches that have been made:

$$precision = \frac{number\ of\ true\ positives}{\text{number of true positives} + \text{number of false positives}}$$

Recall is a measure of the proportion of matches that have been made from all the possible matches:

$$recall = \frac{number\ of\ true\ positives}{\text{number of true positives} + \text{number of false negatives}}$$

Clerical review was carried out in two stages. Firstly, false positive analysis was carried out on a sample of the deterministically linked records to estimate precision. Secondly, false negative analysis was carried out on a sample taken from unlinked (rejected) record pairs to estimate recall.

To do this, a basic probabilistic (score-based) linkage of the residuals was implemented using Splink for the purposes of creating samples for clerical review only. Record pairs for both the false positive analysis and false negative analysis were run through the Data Linkage Hub's Clerical Review Online Widget (CROW) version two tool for review on a pair-wise basis.

For most groups, sample sizes for clerical review were derived with Statulator using a confidence level of 95.00%, an expected proportion (of false positives) of 0.05, and a relative precision of 0.4. For sample groups consisting of a small number of record pairs, all record pairs were clerically reviewed.

Precision and recall for the entire population were derived using total estimated errors. This is the sum of multiplying the error rate with the number of record pairs for each bucket and then aggregating up to the entire population.

The confidence intervals (CIs) estimated for the population were derived using the Agresti-Coull method with a confidence level of 95.00% for each bucket, and then aggregated up to the population level. Upon aggregation of the overall CI of precision or recall, a value of zero was applied for both the lower and upper bounds for each bucket where no false positives or false negatives were found and where the sample size was small. For large sample sizes with no errors, the lower bound was adjusted to zero and the upper bound was calculated by dividing 3 by the sample size. For buckets where all record pairs were reviewed the precise error rates were used.

For the false positive analysis of deterministic links, 7,718 record pairs were sampled for clerical review, stratified by matchkey and 156 true positives were identified (shown in Table 4). Precision was estimated to be 96.93% CI [95.19%, 97.88%] (shown in Table 5).

Table 4: Results of false positive analysis of the PSC 2024 (UK) to Census 2021 linkage (England and Wales)

Sample group	Matchkeys in group	Count of record pairs in group	Record pairs sampled	True positives found	False positives found	True positive estimate	False positive estimate
1	1 to 3	2,735,393	460	458	2	2,723,500	11,893
2	4 to 5	41,875	460	450	10	40,965	910
3	6 to 7	432,742	460	460	0	432,742	0
4	8 to 9	4,092	460	453	7	4,030	62
5	10 to 13	60,797	460	459	1	60,665	132
6	14 to 17	17,159	460	460	0	17,159	0
7	18 to 21	29,556	460	459	1	29,492	64
8	22 to 23	45,585	460	453	7	44,891	694
9	24 to 27	38,860	460	460	0	38,860	0
10	28 to 30	4,591	460	455	5	4,541	50
11	31 to 32	164	164	164	0	164	0
12	33 to 34	452	452	439	13	439	13
13	35 to 38	285,178	460	446	14	276,499	8,679
14	39 to 41	3,712	460	446	14	3,599	113
15	42 to 45	341,107	460	449	11	332,950	8,157
16	46	202	202	202	0	202	0
17	47 to 50	1,152,273	460	453	7	1,134,738	17,535
18	51 to 55	1,024,404	460	396	64	881,878	142,526

Source: Companies House persons of significant control to Census 2021 linked data from the Office for National Statistics

Table 5: Precision results from PSC 2024 (UK) to Census 2021 linkage (England and Wales)

Total true positive estimate	6,027,314
Total false positive estimate	190,828
Precision estimate	96.93%

Source: Companies House persons of significant control to Census 2021 linked data from the Office for National Statistics

For the false negative analysis, buckets were created based on match weight, giving a total of five buckets. A sample of 3,119 pairs were clerically reviewed to detect false negatives and 1,678.5 false negatives were identified in the sample (shown in Table 6). Recall was estimated to be 97.93% CI [97.11%, 98.23%] (shown in Table 7).

Table 6: Results of false negative analysis of the PSC 2024 (UK) to Census 2021 linkage (England and Wales)

Sample group	Match weight	Count of record pairs in group	Record pairs sampled	False negatives found	True negatives found	False negative estimate	True negative estimate
1	Less than 0	1,642,092	460	0	460	0	1,642,092
2	0 to 20	2,727,650	460	4	456	23,719	2,703,931
3	20 to 25	25,560	460	184	276	10,224	15,336
4	25 to 30	91,081	460	456	4	90,289	792
5[Footnote 1]	30 to 35	2,662	460	365.5	94.5	2,115	547
6	35 to 40	1,066	460	310	150	718	348
7	40 to 45	252	252	252	0	252	0
8	Greater than 45	107	107	107	0	107	0

Source: Companies House persons of significant control to Census 2021 linked data from the Office for National Statistics

Notes

1. One of the clerical review files from this group was reviewed twice in error, the mean was taken of the results from this file.

Table 7: Recall results from PSC 2024 (UK) to Census 2021 linkage (England and Wales)

Total false negative estimate	127,424
Total true positive estimate	6,027,314
Recall estimate	97.93%

Source: Companies House persons of significant control to Census 2021 linked data from the Office for National Statistics

Analysis of characteristic representation

Bias analysis is important for telling us about the representativeness of our linked data. It provides a measure of whether linkage failure is random or related to characteristics of those in the data. If there is bias in a linkage process, which causes a certain demographic to be more or less likely to appear in the linked data, then any conclusions drawn from the linked data could be incorrect or misleading.

Linkage failure may be caused by errors or inconsistencies between the data sources. For example, if an individual has changed addresses but their persons of significant control (PSC) address has not been updated or if their address on PSC is a work premises (rather than a home address) and has not been included in their census record.

However, it is also important to consider whether failure to link is because of differing coverage of the two data sources (meaning that the link does not exist).

Roughly 50% of PSC records were not found on the census. However, the estimated recall of 97.93% suggests that nearly all of the possible links were made successfully. Therefore, it is likely that most individuals from PSC who did not link to Census 2021, could not be found on the census.

One of the reasons for this differing coverage of the data sources is likely to be that individuals who have set up companies in the UK but live outside of England and Wales (including Scotland, Northern Ireland and abroad) would not be expected to link. Individuals who had a country of residence of the UK or Great Britain may live outside of England or Wales, and therefore would not be present on census. Some individuals may have reported a country of residence expected to be on census but could have moved or have another residence abroad, so may not have been captured on the census.

It is also possible that individuals who have died are not removed from PSC promptly. If someone had died before Census Day but were not removed from PSC then they would not have linked.

It is challenging to disentangle linkage bias from differences in coverage; but to try and differentiate the two, the bias analysis was additionally carried out on a subset of the data made up of records which were thought reasonably likely to be for individuals covered on census.

It is also important to note that a limitation of the quality analysis was that there was limited information available on PSC data to make clerical decisions and therefore may have been insufficient to correctly identify false negatives. As such, recall may be an overestimate.

Methodology

For this bias analysis, demographic characteristics for the linked data were compared with the underlying PSC data. The variables from the PSC are not validated and are likely to be of lower quality than census data. However, it was necessary to use the PSC variables for bias analysis in order to compare the original PSC data with the linked data.

Bias was analysed using proportional discrepancy. It might be expected that in a linked dataset each category (for example, age or sex) will have the same match rate as the overall match rate for the linkage, however, this is often not the case.

Proportional discrepancy measures how the match rate within a category differs from what is expected. Positive proportional discrepancies convey overrepresentation of the characteristic evaluated, while negative proportional discrepancy scores convey underrepresentation. For example, a proportional discrepancy score of negative 0.05 suggests the linked data have only matched 95% of the expected number of matches given the overall match rate. When a group is described as being underrepresented, this means that they are underrepresented in the linked data compared with the unlinked PSC data, given the match rate.

For most of the variables where proportional discrepancy is examined there are some records with null values, which represents missing data. These null groups often show as being underrepresented in the data. It is possible that records that have null values for some variables are less likely to have linked because they are of poorer quality than the more complete records. In the following analysis, null groups are shown on the graphs.

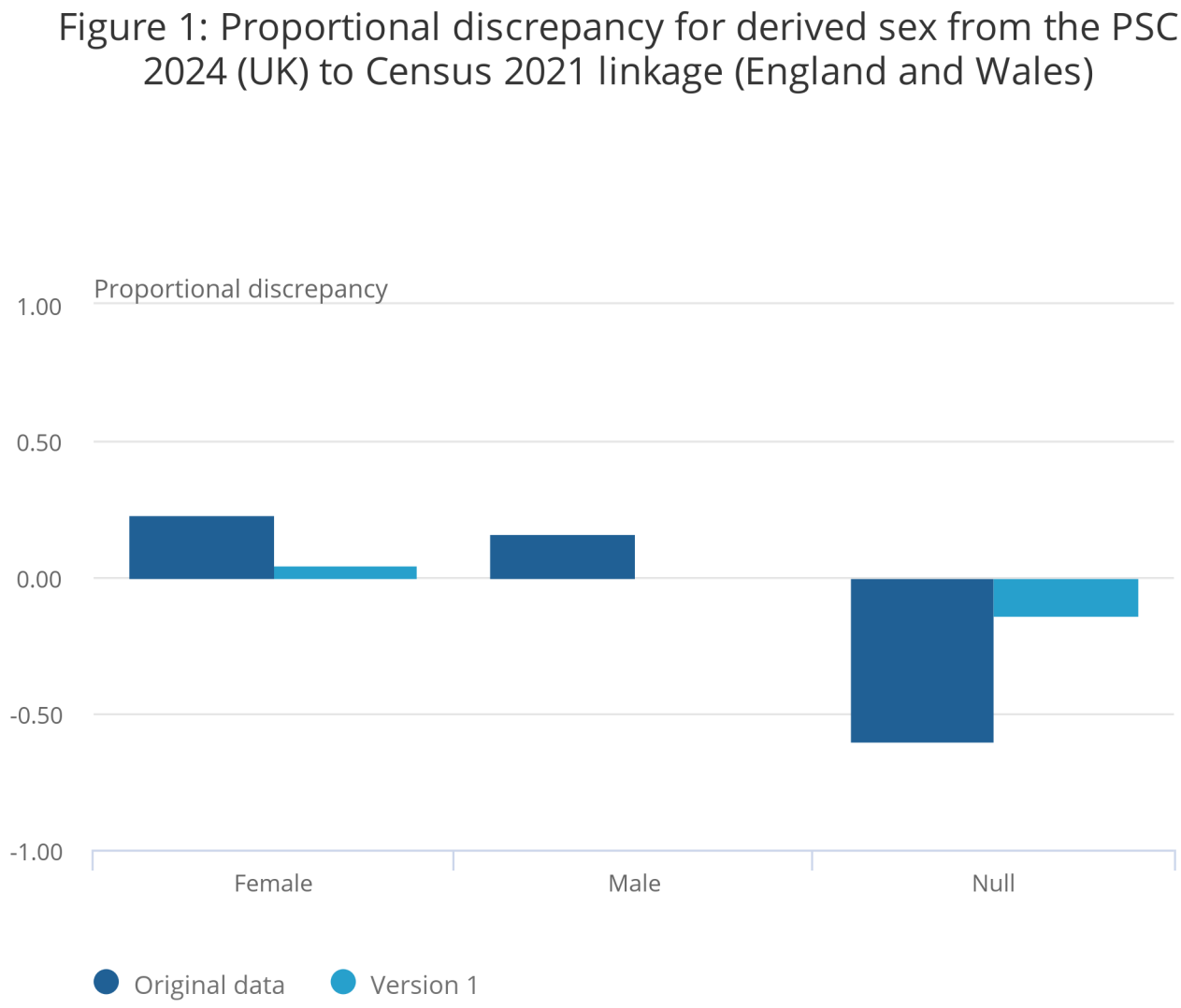
Bias analysis was carried out on multiple versions of the data, comparing PSC with the linked data. Firstly, bias analysis was carried out on the whole dataset (referred to as original data), so including records which were of too low a quality to link. Next, bias analysis was carried out on a version of the dataset where low quality records and individuals not expected to be on census based on their country of residence, region and year of birth were removed (referred to as version 1).

Derived sex

Sex was derived from the title variable of PSC, as explained in Companies House (PSC). Figure 1 shows the examination of proportional discrepancy for derived sex, showing that both males and females are well represented in the linked data in the original data and version 1 of the bias analysis.

While the trends between the bias analysis for the original data and version 1 of the data were similar, the magnitude of these trends is greater in the original data. Records with null derived sex are underrepresented, which is not unexpected as these incomplete records are likely to have been of poorer quality and so are less likely to have linked. Nulls were present here when a record had a non-gender specific title or no title.

Figure 1: Proportional discrepancy for derived sex from the PSC 2024 (UK) to Census 2021 linkage (England and Wales)



Source: Companies House persons of significant control to Census 2021 linked data from the Office for National Statistics

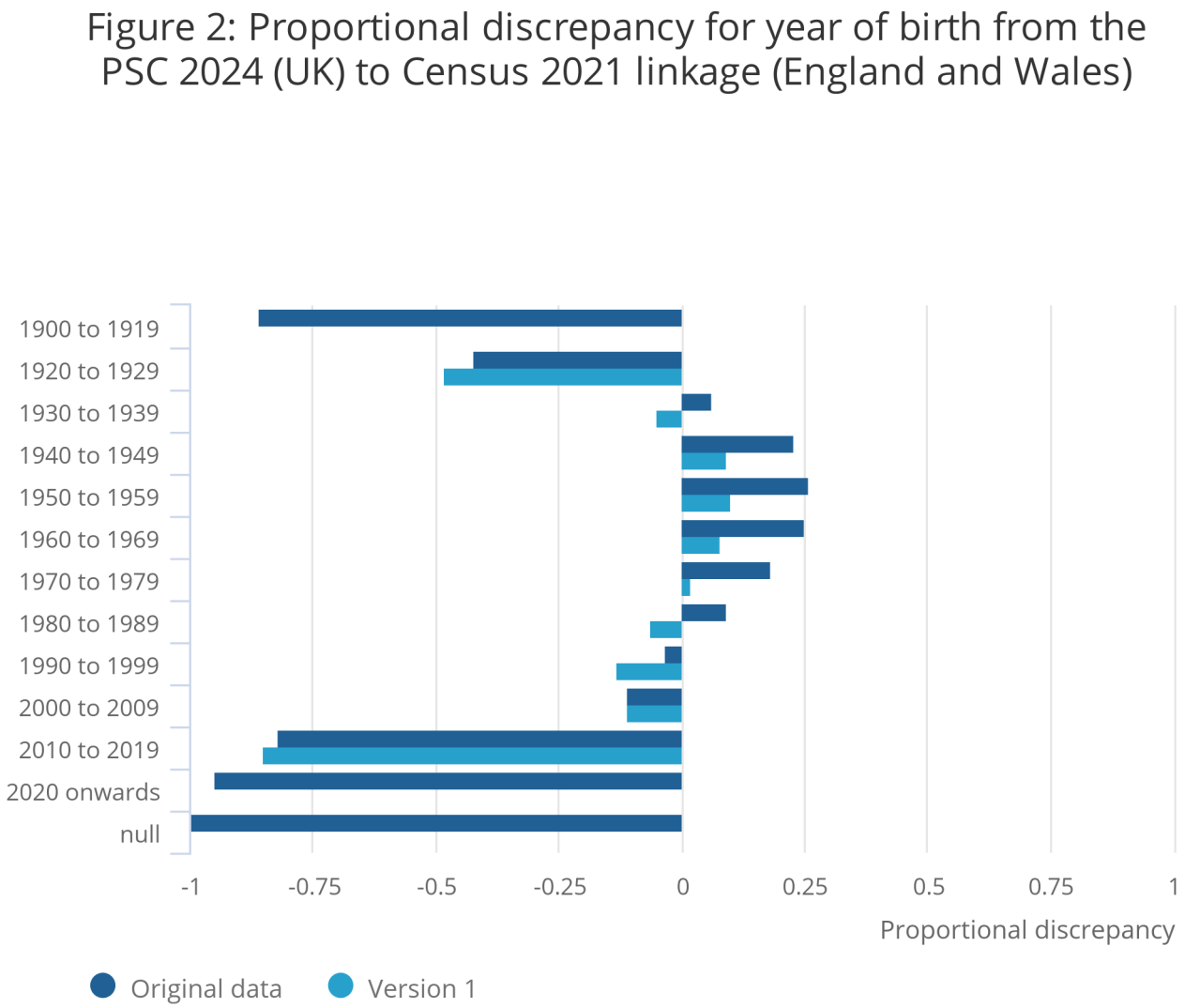
Year of birth

Figure 2 shows the examination of proportional discrepancy for year of birth, showing that for the bias analysis carried out on the original data: those born before 1930, those born after 1989, or with a null year of birth are underrepresented in the linked data.

The bias analysis carried out on version 1 of the data showed generally similar trends to those seen in the original data, but of lesser magnitude for most groups. Years of birth between 1930 to 1939 and 1980 to 1989 were well represented in the original data but were slightly underrepresented in version 1 of the data. In version 1, records with years of birth before 1920, after 2019, or with null values were removed.

It is possible that some of the individuals on PSC who were born prior to 1930 had died prior to Census 2021, but their records have not been removed from PSC, causing individuals born at this time to be underrepresented in the linked data. Individuals born between 1990 and 2009 are only very slightly underrepresented in the linked data. Individuals born from 2020 onwards would not appear on the census if born after Census Day. While it is possible that children can be on PSC it should be considered whether years of birth from 2010 onwards are erroneous.

Figure 2: Proportional discrepancy for year of birth from the PSC 2024 (UK) to Census 2021 linkage (England and Wales)



Source: Companies House persons of significant control to Census 2021 linked data from the Office for National Statistics

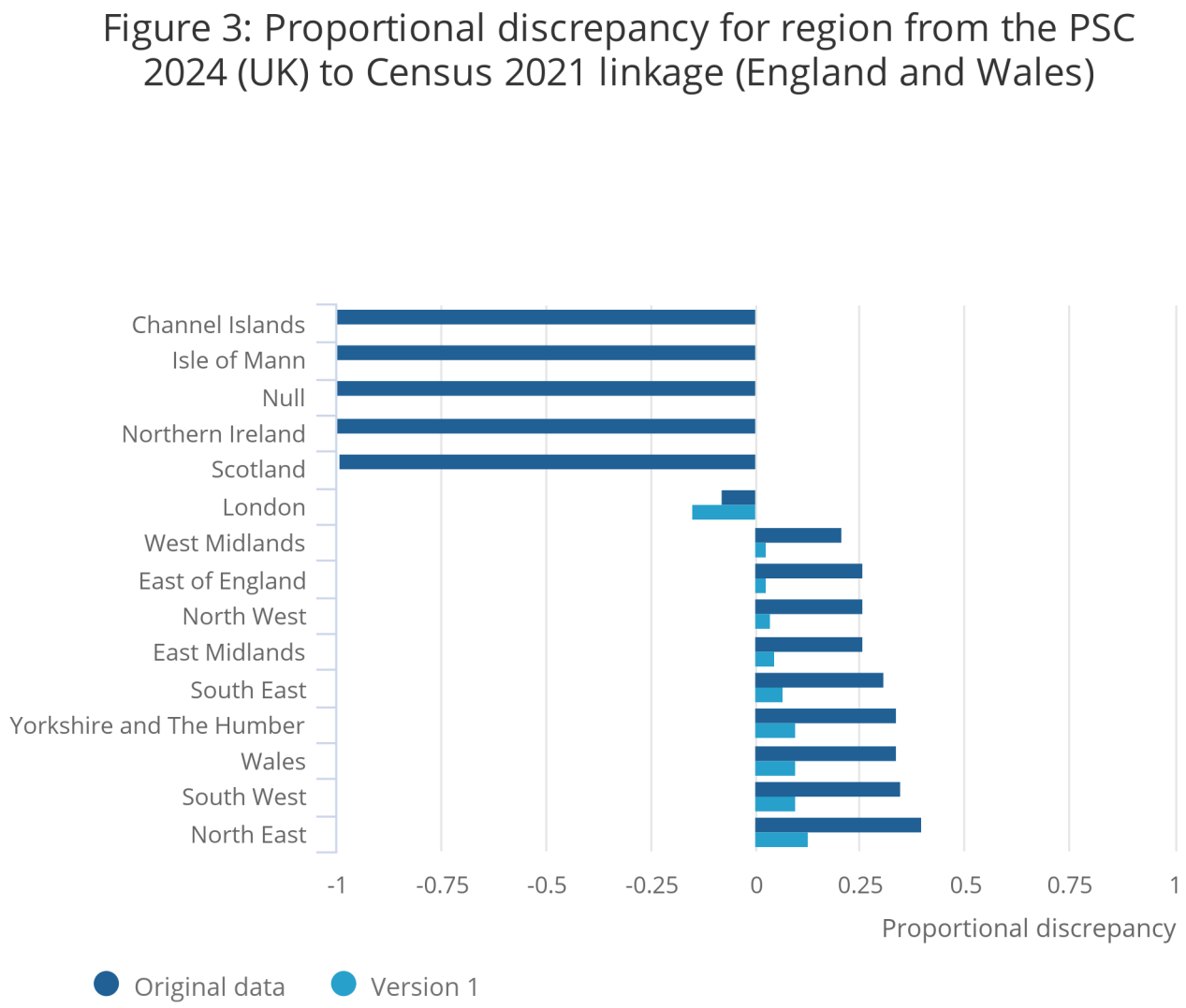
Region

Figure 3 shows the examination of proportional discrepancy for region, showing that London, Scotland, Northern Ireland, Isle of Man, Channel Islands and Nulls are underrepresented in the linked data. Nulls were present here when a postcode was invalid or missing (so failed to link to the National Statistics postcode lookup).

It is to be expected that Scotland, Northern Ireland, Isle of Man and Channel Islands are underrepresented as they are not covered by Census 2021. For London, version 1 of the data shows the same trend as seen in the original data, but with greater magnitude. Wales and all regions in England except for London are less overrepresented in version 1 than in the original data.

The proportional discrepancy score of negative 0.1549 for London in version 1 of the bias analysis suggests that the linked data have only matched 84.51% of the expected number of matches given the overall match rate. This may be caused by the population in London being more mobile and therefore harder to link.

Figure 3: Proportional discrepancy for region from the PSC 2024 (UK) to Census 2021 linkage (England and Wales)



Source: Companies House persons of significant control to Census 2021 linked data from the Office for National Statistics

Nationality and country of residence

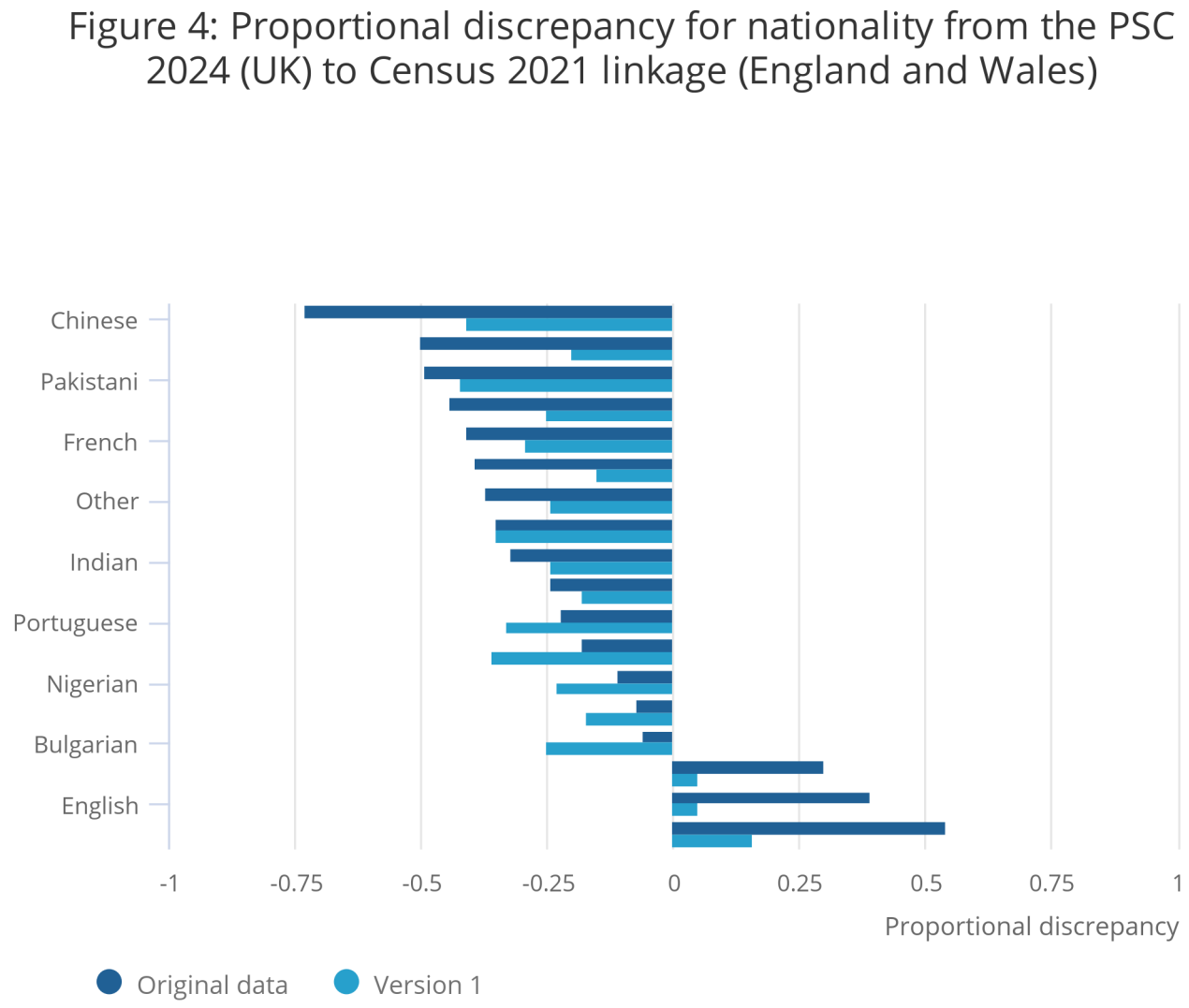
Nationality and country of residence were self-reported on PSC. The top 20 nationalities and countries of residence were examined, all other nationalities and countries were grouped into “Other” for ease of understanding as there were more than 150 distinct values for each of these variables.

Figure 4 shows the examination of proportional discrepancy for these top 20 nationalities and the Other and Null groups, showing that all nationalities are underrepresented except for British, English and Welsh. Trends are broadly similar between the bias analysis carried out on the original data and version 1 of the data, but of lower magnitude for the overrepresentation of Welsh, English and British in version 1 compared with the original data. Bulgarian, Polish, Nigerian, Romanian and Portuguese nationalities are more underrepresented in version 1 than in the original data. Overall, these differences are fairly small and bring these nationalities more in-line with the trends seen for the other underrepresented nationalities.

Linkage methods are likely better at linking Western than non-Western names, so it is possible that this underrepresentation is partly because of this. However, this underrepresentation of individuals with a nationality other than British, English or Welsh is likely in part because of individuals present on PSC not living in England or Wales (but being a person of significant control of a company registered on Companies House) and therefore not being present on Census 2021.

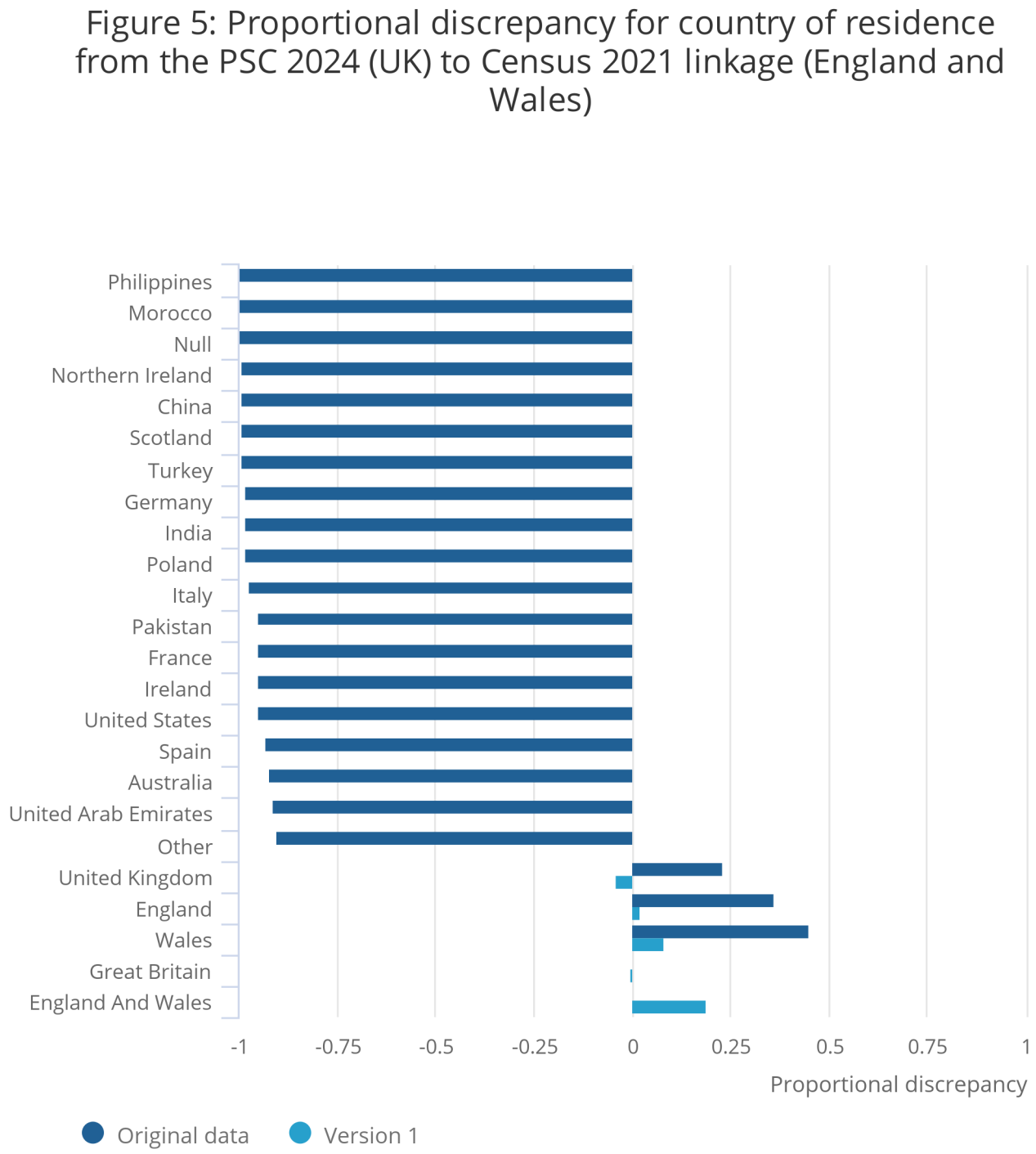
Figure 5 shows a similar trend, with all countries of residence except for the UK, England and Wales being underrepresented in the bias analysis of the original data. Version 1 of the data only included individuals with a country of residence of England and/or Wales, Great Britain or the UK (so all other countries are not included in the bias analysis of version 1 of the data). Here England, Wales, and England and Wales (grouped) were well represented in the linked data, whereas the UK and Great Britain were slightly underrepresented (these categorise were grouped within Other in the original analysis). This is not surprising as some of these records will be for people living in Scotland and Northern Ireland.

Figure 4: Proportional discrepancy for nationality from the PSC 2024 (UK) to Census 2021 linkage (England and Wales)



Source: Companies House persons of significant control to Census 2021 linked data from the Office for National Statistics

Figure 5: Proportional discrepancy for country of residence from the PSC 2024 (UK) to Census 2021 linkage (England and Wales)



Source: Companies House persons of significant control to Census 2021 linked data from the Office for National Statistics

5 . Summary, recommendations and limitations

The linkage between Companies House persons of significant control (PSC) snapshot (January 2024) and Census 2021 has been shown to be of good quality, given quality issues with the PSC data, with an estimated precision and recall of 96.93% and 97.93%, respectively. The linkage rate was lower than we would have expected at 50.31% of all PSC records, or 66.22% of PSC records which were of a reasonable quality and likely to be covered by census. Possible reasons for the linkage rate not being higher include the following.

Quality

Lack of day of birth variable and missingness for derived sex and postcode on PSC, making the data harder to link.

Coverage

PSC includes individuals living in Scotland, Northern Ireland and abroad, who are not included on Census 2021.

Time lag

It is unknown if PSC records are updated after someone moves addresses, so in the following scenarios addresses may not match:

- if someone registered on PSC, then moved address before Census 2021 and did not update PSC
- if someone registered on PSC after Census 2021, then moved address and did not update PSC
- this may have been exacerbated by the effect of the coronavirus (COVID-19) pandemic on individual's usual place of residence for Census 2021

Low data quality made it harder for clerical reviewers to make judgements about matches and non-matches. This is likely to have led to them being more conservative in what they class as a match, which means that it is possible that recall was overestimated. This means there are likely more missed matches than estimated during clerical review.

Bias analysis comparing the linked data with PSC showed underrepresentation of several groups in the linked data compared with the full PSC data. This may be because of a mix of linkage failure and coverage error between the datasets. The underrepresented groups included those born before 1930 or after 2009, those with a London postcode, and those with nationality other than British, English or Welsh.

If analysis is carried out using the linked data, caution should be taken when observing patterns for these groups as analysis outcomes could be the result of linkage bias and coverage error rather than reflecting true trends or patterns.

6 . Cite this methodology

Office for National Statistics (ONS), released 22 August 2024, ONS website, methodology, Companies House persons of significant control to Census 2021 linkage report