An individual level method for improved estimation of ethnic characteristics

Tian Lan and Paul A. Longley. *International Regional Science Review*

(journals.sagepub.com/home/irx)

Abstract

This paper develops an improved method for estimating the ethnicity of individuals based on individual level pairings of given and family names. It builds upon previous research by using a global database of names from c. 1.7 billion living individuals, supplemented by individual level historical census data. In focusing upon Great Britain, these resources enable, respectively, greater precision in estimating probable global origins and better estimation of self-identification amongst long-established family groups such as the Irish Diaspora. We report on geographic issues in adjusting the weighting of groups that are systematically under- or over-predicted using other methods. Our individual level estimates are evaluated using both small area Great Britain census data for 2011 and individual level data for asylum seekers in Canada between 1995 and 2012. Our conclusions assess the value of such estimates in the conduct of social equity audits and in depicting the social mobility outcomes of residential mobility and migration across Great Britain.

Keywords: ethnicity classification; consumer data; Great Britain; social equity audit.

1. Introduction

Ethnicity is a salient characteristic of individual identity. Of relevance to regional science, it has underpinned research into residential differentiation and social segregation (e.g., Finney and Simpson 2009; Lan et al. 2020), labour market recruitment (Yemane and Fernández-Reino 2021), inter-generational social mobility (Clark and Cummins 2015), innovation processes (Wilson et al. 2018), and health outcomes (Petersen et al. 2021). It is also of policy interest to provide timely inter-census estimates of population characteristics (Office for National Statistics 2017), as demonstrated during the 2020 COVID pandemic and following Brexit. Related work has documented the correspondence between individual naming practices and ethnicity, and consequently, the ways in which given (forename) and family (sur-)names may be used to indicate ethnicity (Mateos et al. 2009; Parameshwaran and Engzell 2015). As such, names-based classification of ethnicity is of wide applicability to many issues of relevance to regional scientists in studies of migration, urban structure and regional functioning – issues that we return to in our conclusions.

Names-based ethnic classification methods typically develop algorithms to identify significant forename – surname associations and assign labels to the resulting cultural, ethnic, and linguistic groups at different levels of aggregation. A recent development of these approaches is the Ethnicity Estimator software (Kandt and Longley, 2018) that was developed in collaboration with the Office for National Statistics (ONS). A novel aspect of this latter approach is the evaluation of estimates with respect to survey respondent self-identifications: such procedures are of particular value where names span different ethnic groups (as with members of the Black Caribbean and White British UK Census groups) or where long-settled groups may no longer identify with their ancestral origins (as with some White Irish individuals in Britain). Kandt and Longley's (2018) software and the derivative small area estimates of annual changes in local ethnic group composition have been used in circa 60 research projects to date (CDRC, personal communication). The free availability of this classification software for research purposes and the peer-reviewed documentation of its predictive success marks this software as a basis for the further evaluation and improvements developed in this paper.

Kandt and Longley (2018) use the ONS Secure Research Service, previously the Virtual Microdata Laboratory (Ritchie 2008), for names classification by first using a names dictionary and queries to a secure census database to calculate the probabilities of membership of each of 11 census groups (see Table 1) used in the 2011 Census. Summed scores for each group can be calculated for every forename and surname pair that occurs in their names dictionary by summing these (equally weighted) probabilities. Using secure access to individual 2011 Census records, Kandt and Longley (2018) reweight the resulting assignments to match the pattern of self-reported assignments in the Census records. The authors demonstrate that their approach results in greater predictive success than a previous ('Onomap': Mateos et al. 2011) algorithmic approach and that their weighting factors are optimised within the constraints of secure research facility access. However, it is apparent that the White Irish group is consistently under-estimated, that there are systematic mis-assignments between individuals identifying with Indian subcontinent countries, and that there are failures in predictions of occurrences of the Other Asians, Black Caribbean and Other groups. It is also desirable for ethnicity audits to be able to disaggregate the ‘White Other’ and ‘Other Asian’ categories into constituent countries that may typically confer quite different human and social capital upon their citizens and, by extension, different migration outcomes in migration destinations such as Great Britain.

Table 1. Comparison of adult population (16+) breakdown by ethnic groups predicted by applying Kandt and Longley's (2018) estimator to the 2011 Linked Consumer Register (LCR) for Great Britain

Census groups	Abbreviation	Census	LCR	Ratio (LCR/Census)
Asian Other	AAO	663,124	200,268	30%
Bangladesh	ABD	294,505	285,860	97%
Chinese	ACN	371,521	229,645	62%
Indian	AIN	1,167,436	1,343,027	115%
Pakistani	APK	788,849	1,189,890	151%
Black African	BAF	713,257	564,556	79%
Black Caribbean	BCA	496,195	268,614	54%
Any Other	OXX	1,298,097	167,881	13%
White Other	WAO	2,286,231	2,273,858	99%
White British	WBR	41,245,227	40,915,170	99%
White Irish	WIR	551,410	311,655	57%
Total		49,875,852	47,750,424	96%

Our research objectives are to improve or refine estimates of membership of: (a) the long-established White British majority population that was actually present in the 19th century; (b) the long-established White Irish population that continues to identify with this group; (c) the Black Caribbean population that shares naming conventions with white ethnic groups; (d) groups originating in the Indian sub-continent; and (e) the 'catch all' Black African, Black Caribbean, White Other and Other Asian groups, which may be attributed to particular countries that confer quite different circumstances upon migrants from them. Details of development and SQL code used to develop the software, Onomap3, can be found on the cdrc.ac.uk website, for access for research purposes upon successful application.

2. Data Sources

Our approach is to use the near-complete Linked Consumer Register (LCR) of all adult individual names and addresses in Great Britain in 2011 (see Lansley et al. 2019; Van Dijk et al. 2021) as a frame to estimate ethnicities. The 2011 LCR provides an annual snapshot of the UK adult population created and curated by the ESRC Consumer Data Research Centre (CDRC), as part of a corpus of such data initially covering the period 1997-2016. The LCRs are individual level data compiled from the public version of the UK Electoral Register and other consumer data sources. Lansley et al. (2019) describe the data cleaning, triangulation, imputation and validation processes that are intrinsic to their creation: the 2011 LCR is documented to have similar numbers of adults compared with those recorded in the Census across a range of census geographies.

Here, we estimate the ethnicity of every individual on the 2011 LCR. By georeferencing each record we are then able to compare our estimates with Census figures for the same year at the level of the Lower layer Super Output Area (LSOA, a small area geography in England and Wales with a typical population of 1,500). We use these initial results to adjust the weights assigned to forenames and surnames for different ethnic groups. For the specific case of the White Irish population, we also refer to individual level 1881 Census records to evaluate the merit of deeming a contemporary bearer to self-identify with the ‘White Irish’ Census category. The digitised versions of the GB Censuses for 1851-1911 are curated by the I-CeM project (Higgs and Schurer 2019), and individual level records including names, addresses and birthplaces were made available to us by the UK Data Service under special licence. We use the individual level data for 1881, based on our exploratory findings that the data capture process for this year appears to have been particularly effective.

We also use the WorldNames2 (WN2) database that arises from an ongoing project to assemble a representative range of forenames and surnames for every country of the world. O’Brien and Longley (2018) detail the various sources used, including public electoral registers, telephone directories and professional or school registers. The database currently comprises *circa* 1.7 billion individuals’ names, or about one fifth of the world’s population (calculated based on 7.9 billion according to the UN estimates as of 2021), each with country attribution. Based on the sampled names in the countries and their total populations, frequencies per million (FPMs) of family name occurrences and their estimated populations sizes are derived in the WN2 database.

Aggregate 2011 Census adult population counts classified into 11 ethnicity categories (listed with their abbreviations in Table 1) provide a benchmark for evaluation of the ethnicity estimates developed using the LCRs. The ethnicity categorisations recorded in the 2011 Census questionnaires differ slightly between the different constituent countries of the UK but can be harmonised into the 11 categories. Table 1 also compares the GB population breakdown by ethnic groups estimated by applying Kandt and Longley's publicly available software to the 2011 LCR and the corresponding 2011 Census figures. Both over-estimation and under-estimation are observed amongst the LCR group assignments.

3. Methods and Outcomes of Reassignments or Enhancements.

The 2011 classifications of ethnicity used by the UK Office for National Statistics are the outcome of extensive consultation with stakeholders with regard to end uses of statistical sources so classified (Office for National Statistics 2009), which is reflected in the subtle variations among the ethnic categories adopted by Northern Ireland, Scotland, and England and Wales. The outcome is, inevitably, a snapshot of policy concerns that resonate with the governments of the constituent countries of the United Kingdom. The resultant classes also manifest a long sweep of British history that accommodates Irish and New Commonwealth migration, but not the specific consequences of successive EU enlargements during the UK's period of EU membership or refugee migration. Our dual purpose is to improve the efficacy of Kandt and Longley's assignments to the harmonised classes used in Table 1 while also extending it to differentiate between other nations, membership of which might also affect the circumstances of migrants to Britain.

As such, our aim is to extend the granularity of ethnic classification while also retaining sensitivity to the issues of self-identification developed in Kandt and Longley's (2018) work. We use their Ethnicity Estimator (EE) as a baseline model for our proposed improvements and extensions. The core process of the EE, summarised in Equation (1), is to assign each forename-surname pairing a probability of assignment to each of the Census ethnic categories E, as detailed in Table 1. For any name pairing, $p_{E,f}$ and $p_{E,s}$ denote the probabilities of assignment to each ethnic group E for the forename and surname respectively, as defined in two EE name-ethnicity lookup tables. Two weighting factors that sum to unity, w_f and w_s, are used to specify the relative contributions of forename and surname to the estimated outcome score S_E. In the original EE algorithm, these weights are each set equal to 0.5. After calculating the score S_E for every one of the 11 ethnicity categories, the name pair is assigned to the ethnic group with the highest composite score.

$$S_E = w_f * p_{E,f} + w_s * p_{E,s} \tag{1}$$

In developing and extending this approach to classify Great Britain residents, we use additional individuals' names obtained from the 1881 Great Britain Census and from WN2. We validate the results using aggregate 2011 Census small area statistics for the same year as the 2011 LCR. Ethnicity classification of the 2011 LCR follows a chronology of steps (see Table 1 for abbreviations used), for reasons set out in our discussion below:

1) The EE classifications are assigned as provisional estimates.
2) Family names classified as White British (WBR) but that are not recorded at all in the 1881 Great Britain Census are reassigned to their second highest predicted category amongst the remaining 10 census ethnic groups.
3) Individuals classified as WBR or White Irish (WIR) are then pooled. Reassignments between them are made using Bayes' Theorem and WN2 data as detailed below.

4) Individuals classified as Asian Indian (AIN), Asian Pakistani (APK), Other Asian (AAO) are pooled and reassigned using re-weightings as detailed below.
5) Individuals classified as Black Caribbean (BCA), WBR or All Other (OXX) are pooled and reassigned using rules as detailed below.
6) WN2 data are used to assign most probable countries to records assigned to the AAO, BAF, BCA and WAO groups.

3.1. The White British and White Irish groups

Kandt and Longley (2018) identify the WIR group as systematically under-estimated, attributing this to self-identification of descendants of previous generations of Irish migrants with the WBR group. We take the explicit decision to define WIR in terms of being long settled in the Irish Republic and WBR as conveying establishment in the United Kingdom. Our approach to accommodating this tendency is threefold: (a) we constrain WBR assignments by filtering out family names not present in the 1881 Great Britain Census; (b) we adjust the forename and surname relative probabilities $p_{E,f}$ and $p_{E,s}$ between WBR and WIR in the name-ethnicity lookup tables using data relating to the relative frequencies of each in the UK and Ireland as recorded in the WN2 population estimates; and (c) we tune the two weighting factors w_f and w_s in Equation (1) in order to align our estimates to compare with the total size of the WIR population in the 2011 Census (Table 1) and its geographic distribution.

3.1.1 Reassigning White British names

There are ambiguities in ascribing the label ‘White British’ to any individual whose name does not indicate ancestry beyond Great Britain within historic periods (e.g., see the genetic study of Winney et al. 2012). In refining the EE approach to reduce the over-prediction of the WBR, we choose 1881 (for which well-curated digital Census records are available) as a convenient threshold date for inclusion of any family name as long-established ‘White

British'. We begin by filtering out family names that were not present in the 1881 Census and assigning them to their second highest EE category. 1,284,829 bearers of names classified as White British by EE are thus reassigned to their second highest class. The results shown in Table 2 identify that most all such names are reclassified as White Other or White Irish.

Table 2. Reassignment of the 'White British' predicted in the previous step with family names with no bearers in the 1881 Census.

Group	No. of Individuals
AAO	1,789
ABD	93
ACN	959
AIN	2,766
APK	858
BAF	41,493
BCA	50,417
OXX	289,972
WAO	501,199
WIR	395,283
Total	1,284,829

3.1.2 Adjusting the name-ethnicity lookup tables

We next adjust the forename and surname probabilities $p_{E,f}$ and $p_{E,s}$ between WBR and WIR in the name-ethnicity lookup tables by calculating conditional probabilities of belonging to either group based upon forename – surname pairings. Estimates of the bearers of different UK and Irish Republic forenames and surnames are provided by the WN2 data. Bayes' Theorem is then used to calculate the conditional probabilities of belonging to either WBR or WIR. Table 3 illustrates the steps taken to derive the conditional posterior probabilities, taking the forename 'James' as an example. The final two rows of the Table present the conditional probability based upon the estimated populations of name bearers, independent of the total populations of the host countries. The probabilities of WBR and WIR membership for each forename or surname are thus recalculated and replaced in the look-up tables using the conditional probabilities derived in Table 3.

Table 3. Conditional probability of belonging to the WBR or WIR using the name is 'James' as an example, according to Bayes' Theorem

Variables	Notation
Population of Great Britain	G
Population of Ireland	I
Estimated population of name 'James' in the UK	g
Estimated population of name 'James' in Ireland	i
Probability of belonging to WBR	$P(A) = G/(G + I)$
Probability of belonging to WIR	$P(B) = I/(G + I)$
Probability of being named 'James' given one is British	$P(Y\|A) = g/G$
Probability of being named 'James' given one is Irish	$P(Y\|B) = i/I$
Probability of being named 'James' in the UK or Ireland	$P(Y) = P(Y\|A) * P(A) + P(Y\|B) * P(B)$
Probability of belonging to WBR given the name is 'James'	$P(A\|Y) = P(Y\|A) * \frac{P(A)}{P(Y)}$ $= \frac{\frac{g}{G} * \frac{G}{G+I}}{\frac{g}{G} * \frac{G}{G+I} + \frac{i}{I} * \frac{I}{G+I}}$ $= g/(g + i)$
Probability of belonging to the WIR given the name is 'James'	$P(B\|Y) = i/(g + i)$

3.1.3 Tuning the weighting factors

In Equation (1), the original EE adopts equally weighted contributions from a forename and a surname (w_s=w_f=0.5). We postulate, however, that members of long-established migrant Irish family groups (as identified by surnames) may be less likely to self-identify as WIR. We also postulate a lesser consideration that forename may be a useful indicator of recent migration from the Irish Republic or lingering affinity to the island amongst long-settled migrant families. Accordingly, we downweight the importance of forenames and, consistent with replicating the number of individuals identifying as WIR in the 2011 Census, experiment with a range of values for w_s from 0.76 to 0.85. We compare the numbers and spatial distributions of predicted WIR to the WIR population identified in the 2011 Census. There are tensions in this approach, since prediction success is not spatially invariant, and fine-tuning of weights may cause systematic deterioration of urban predictions at the expense of rural predictions, and vice-versa. Ethnic minorities remain concentrated in towns and cities (albeit decreasingly so), with distinctive regional patterning of different ethnic groups. There is no obvious analytical solution to this issue, particularly given that mis-assignments between some categories may have less severe implications in (some) applications than others. In what follows, we rely upon a visual comparison of observed (census) versus predicted distributions, in the context of aggregate numerical comparisons.

Figure 1 suggests that surname weight 0.84 gives the closest predictions to the Census. Table 4 presents the transition matrix of the reassignment between the WBR and WIR after the lookup table adjustments with the selected surname weight 0.84. Together with the reassignment to WIR in the previous step, we predict 546,743 White Irish at this stage, which accounts for 99% of the 2011 Census observations. Figure 2 shows the observed and estimated 2011 populations of White Irish by LSOA, where our method correctly picks up the concentration of Irish in urban areas such as London, Birmingham, Liverpool, Manchester, and Glasgow, albeit with modest underestimation. This sensitivity analysis is finely balanced, with the global solution required to balance prediction success in rural and urban areas: in particular, it is apparent from sensitivity analysis that Scottish WBR rural names bear more than passing similarities to urban WIR ones.

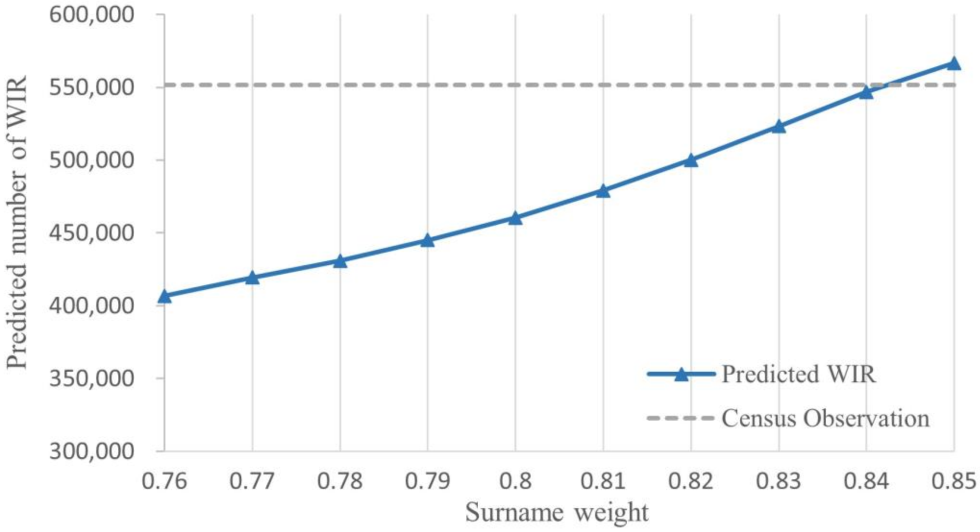

Figure 1. The predicted numbers of WIR in the 2011 LCR using different surname weights, compared to the 2011 Census observation.

Table 4. Confusion matrix of the WBR and WIR populations from EE prediction (rows) and the outcomes of reassignment between White British and White Irish (columns), using the surname weight 0.84 after the lookup table adjustments

		After reallocation between WBR and WIR		
		WBR	WIR	Total
EE prediction	WBR	39,600,396	425,228	40,025,624
	WIR	190,140	121,515	311,655
	Total	39,790,536	546,743	40,337,279

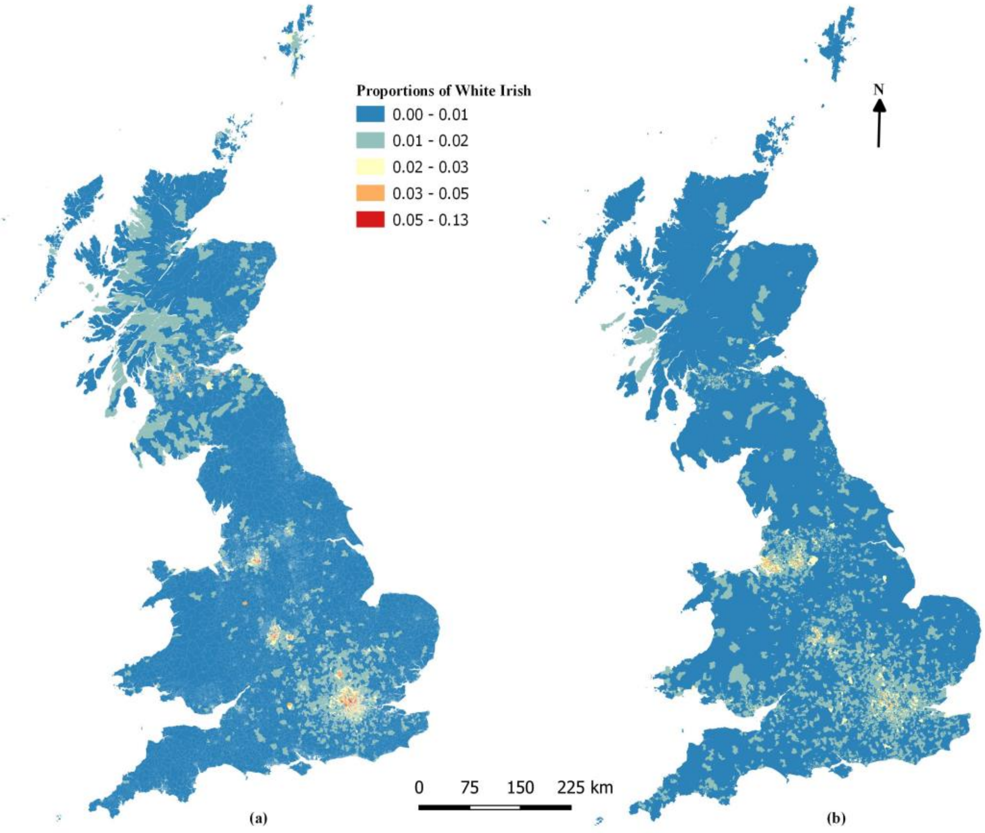

Figure 2. Distributions of White Irish by LSOAs from the (a) 2011 Census and (b) 2011 LCR.

3.2. Indian sub-continent and Other Asian groups

Among the Indian sub-continent groups shown in Table 1, the aggregate predictions of Bangladeshis (ABD) are very close to observations from the Census. However, predictions of Indians (AIN) and (especially) Pakistanis (APK) are overestimated while Any Other Asian (AAO) occurrences are substantially underestimated. The principal 'Any Other Asian' countries are listed in Table 5. We aim to improve estimation by reallocating individuals from AIN and APK to AAO. In order to address this, we first adjust the name probabilities in the name-ethnicity lookup tables relating to the three groups by using estimated populations of bearers of different names across these countries and Bayes' Theorem, as in Section 3.1.2. Additionally, since the EE predicts 136%, 152% and 37% of the observed AIN, APK and AAO Census figures, respectively, all of the adjusted name probabilities relating to the three groups are further reweighed by multiplying the corresponding reciprocal factors: 0.7 (AIN), 0.7 (APK) and 2.7 (AAO). The ABD estimates, which approximate the Census figures, are not included in this reweighting.

Table 5. Countries and codes identified as belonging to the Any Other Asians group

Country name	Country name
Afghanistan	Malaysia
Armenia	Maldives
Azerbaijan	Mongolia
Bhutan	Myanmar
Brunei	Nepal
Cambodia	North Korea
Christmas Island	Philippines
Cocos Islands	Singapore
Diego Garcia	South Korea
Georgia	Sri Lanka
Indonesia	Tajikistan
Israel	Thailand
Japan	Turkey
Kazakhstan	Turkmenistan
Kyrgyzstan	Uzbekistan
Laos	Vietnam

With the above modified name relative probabilities $p_{E,f}$ and $p_{E,s}$ for the AIN, APK and AAO groups, we explore a range of relative forename and surname weighting factors w_s. Weights for this heterogeneous group ranging from 0.25 to 0.75 are applied to names from the LCR classified by EE as Indian, Pakistani or Any Other Asian, to improve the correspondence between ethnicity estimates and 2011 Census figures (see Table 6). The closest predictions of each group to the Census observations are highlighted in bold in this Table. The comparison between predictions and census observations suggests surname weight 0.3 and forename weight 0.7 are the overall best combination, although the Indian group is still over-predicted. Future improvements could consider exploring separate surname weights for the four groups.

Table 6. Predicted populations of the four groups using different surname weights, compared with the GB Census totals.

Surname factor (w_s)	**APK**	**AIN**	**AAO**
0.25	761,302	1,362,784	**614,512**
0.30	**779,968**	1,354,202	604,428
0.35	798,158	1,348,380	592,060
0.40	808,062	1,346,253	584,283
0.45	815,676	1,344,773	578,149
0.50	830,731	1,338,812	569,055
0.55	841,415	1,331,348	565,835
0.60	852,439	1,329,348	556,811
0.65	880,009	1,312,606	545,983
0.70	885,412	1,311,719	541,467
0.75	895,948	**1,304,167**	538,483
GB Census	788,849	1,167,436	663,124

In so doing, we reallocate predictions among the AAO, AIN and APK groups from the provisional EE categories. Table 7 presents a confusion matrix of ethnic group transitions between the EE predictions and our revision following the adjustments. We compare the LSOA spatial distributions of our predictions of the APK, AAO and AIN from the 2011 LCR with 2011 Census statistics in Figures 3-5. These suggest a general alignment in the distributions of the APK, AAO and AIN groups, albeit with over-predictions of the Indian group that are particularly pronounced in South London.

Table 7. Confusion matrix between the group populations from EE (rows) and the outcomes of reassignment among AIN, APK and AAO (columns).

		After reallocation among AIN, APK and AAO			
		AIN	APK	AAO	Total
EE prediction	AIN	1,117,631	58,873	166,523	1,343,027
	APK	179,245	719,004	291,641	1,189,890
	AAO	55,985	2,068	142,215	200,268
	Total	1,352,861	779,945	600,379	2,733,185

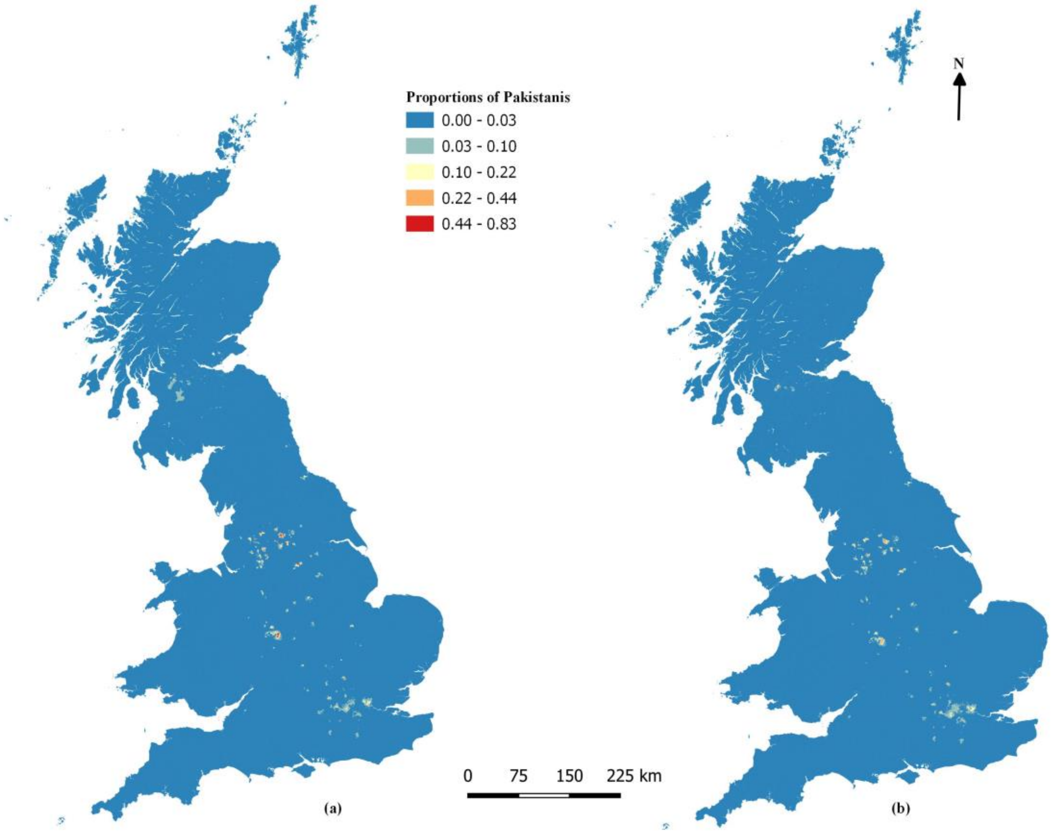

Figure 3. Distributions of the APK group by LSOAs from the (a) 2011 Census and (b) 2011 LCR.

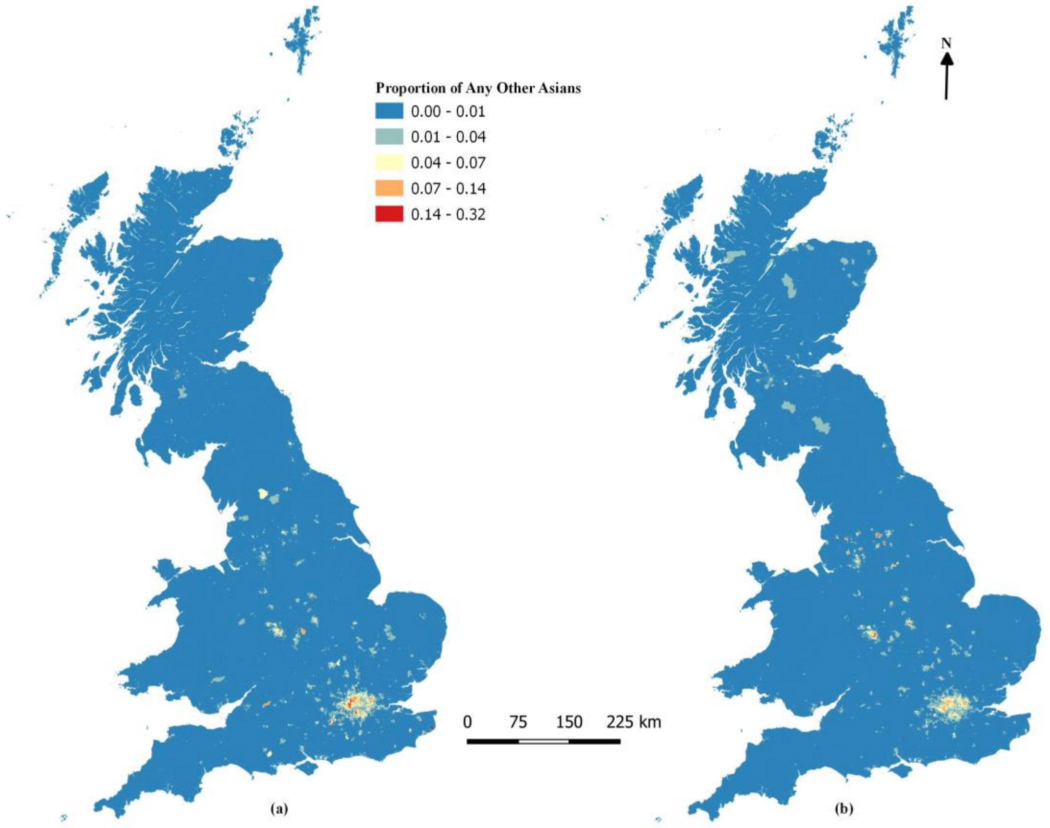

Figure 4. Distributions of the AAO group by LSOAs from the (a) 2011 Census and (b) 2011 LCR.

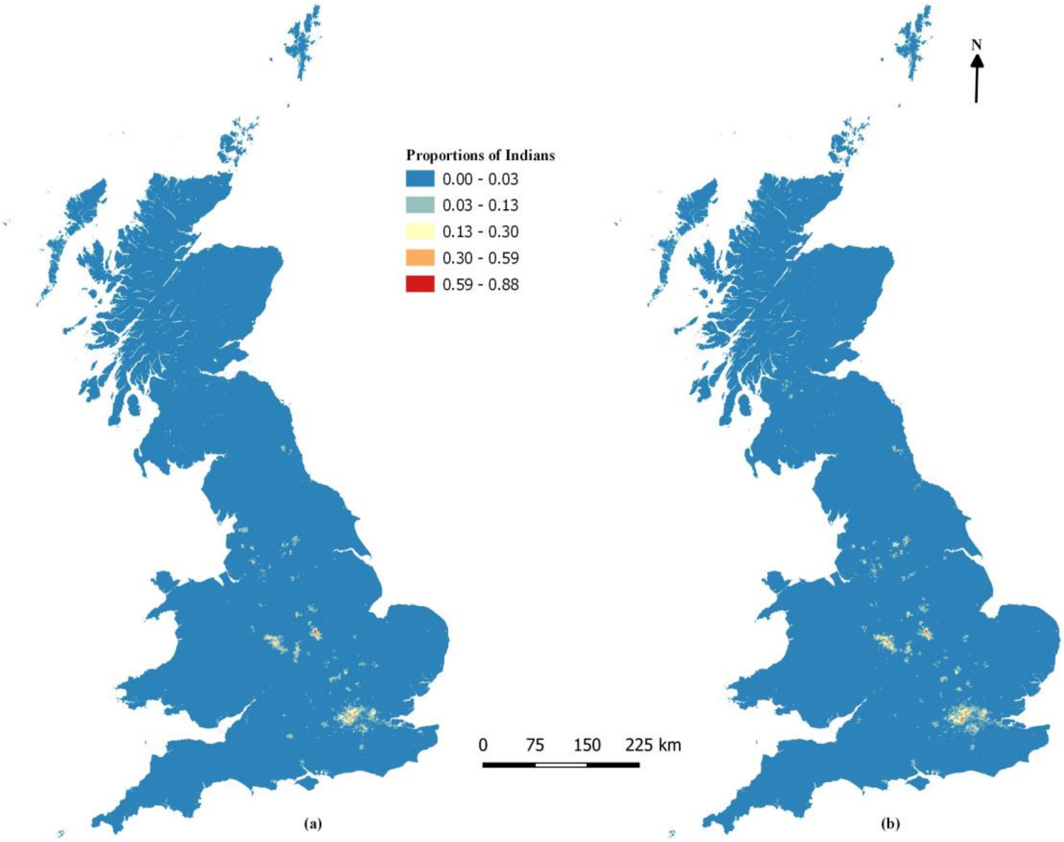

Figure 5. Distributions of the AIN group by LSOAs from the (a) 2011 Census and (b) 2011 LCR.

3.3. Black Caribbean groups

Members of the BCA group share both forenames and surnames with the White British and, as a minority population, are under-enumerated in names-based ethnicity estimators. Although the BCA group is underestimated by the EE in terms of the total population, they are nevertheless overestimated by the EE in some parts of Great Britain, where they are possibly confounded with members of the Any Other (OXX) group. We seek to accommodate this by comparing the frequencies per million (FPM) of forenames and surnames in the UK with those for Caribbean countries with British colonial history. The FPMs of forenames and sur names in available relevant Caribbean countries (Table 8) are extracted from the WN2 database and the highest FPM of a name in any single Caribbean jurisdiction is retained as the FPM of that name in the Caribbean.

Table 8. Caribbean countries with British colonial histories (including current British overseas territories) used in the analysis

Country Name	Relevance
Anguilla	British overseas territory
Antigua and Barbuda	British colonial history
Bahamas	British colonial history
British Virgin Islands	British overseas territory
Cayman Islands	British overseas territory
Grenada	British colonial history
Jamaica	British colonial history
Saint Kitts and Nevis	British colonial history
Saint Lucia	British colonial history
Saint Vincent and the Grenadines	British colonial history
Trinidad and Tobago	British colonial history
Turks and Caicos Islands	British overseas territory

After experimentation and sensitivity analysis, we alight upon a multiplicative index to measure the likelihood of a name being assigned to the BCA group (Equation (2)). The first component of the index records how many times more popular a forename is in the Caribbean than in the UK. The second component records the corresponding multiplier for a surname. The product of the two terms is used as an indicator of the likelihood of belonging to the Black Caribbean group. Making use of the index, Figure 6 illustrates the logic of assigning possible 'WBR' and 'OXX' to 'BCA'. For those who are classified as WBR, BCA and OXX, their multiplicative indices are calculated and compared with different empirical thresholds: 1.5 for 'BCA', 4.9 for 'WBR' and 15 for 'OXX'. The outcomes determine whether the original classifications are retained or they are reassigned to another group among BCA, WBR and OXX.

$$\text{Index} = (\text{Caribbean forename FPM} / \text{UK forename FPM}) * (\text{Caribbean surname FPM} / \text{UK surname FPM}) \quad (2)$$

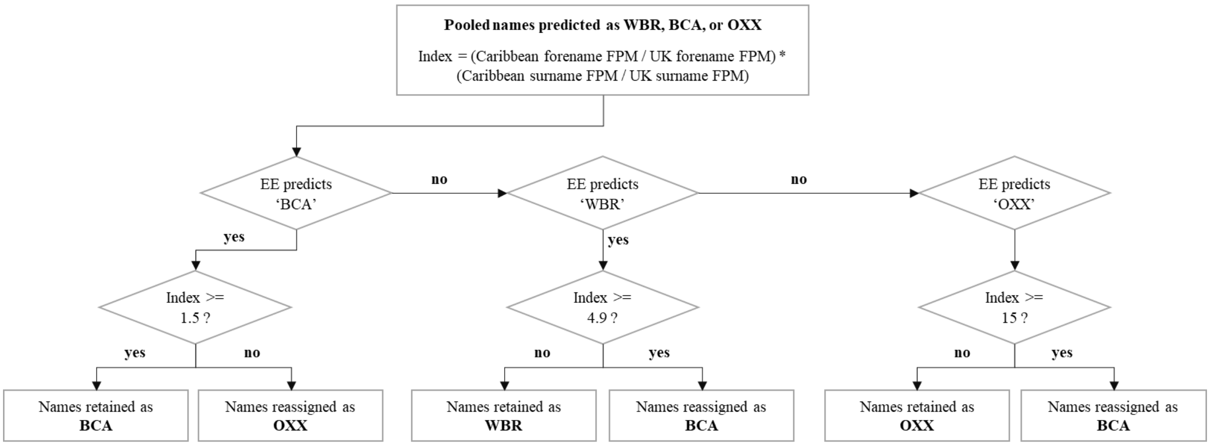

Figure 6. The workflow of assigning possible 'WBR' and 'OXX' to 'BCA' based on forename and surname index scores.

Table 9 shows the confusion matrix of reassignments for the LCR following the adjustments to allocations between the WBR, OXX and BCA groups. The 377,245 adult BCA assignments after all of the previous adjustments compare with 496,195 recorded in the Census, and the estimated 133,803 Caribbean Londoners compare with a Census figure of 268,014. It should be noted that there are 1,793 WIR estimated by EE that are reassigned to WBR in the previous steps but are returned to BCA in this step. Figure 7 illustrates the general geographic correspondence between our estimates and the observed incidence in the Census. There is a dilemma posed by adjusting classification thresholds since under-prediction in London and in Birmingham is partially offset by over-prediction elsewhere in predominantly rural areas. There is scope, however, for further improving estimates for urban areas for applications in which rural areas are not of primary concern.

Table 9. Confusion matrix between the group populations from EE (rows) and the outcomes of reassignment among BCA, WBR and OXX (columns).

		After reallocation among BCA, WBR and OXX			
		BCA	WBR	OXX	Total
EE prediction	BCA	153,664	0	114,950	268,614
	WBR	213,569	39,393,510	0	39,607,079
	OXX	8,219	0	159,662	167,881
	WIR*	1,793	0	0	1,793
	Total	377,245	39,393,510	274,612	40,045,367

* Note: The 1,793 WIR estimated by EE are reassigned to WBR in the previous steps but are returned to the BCA group in this step.

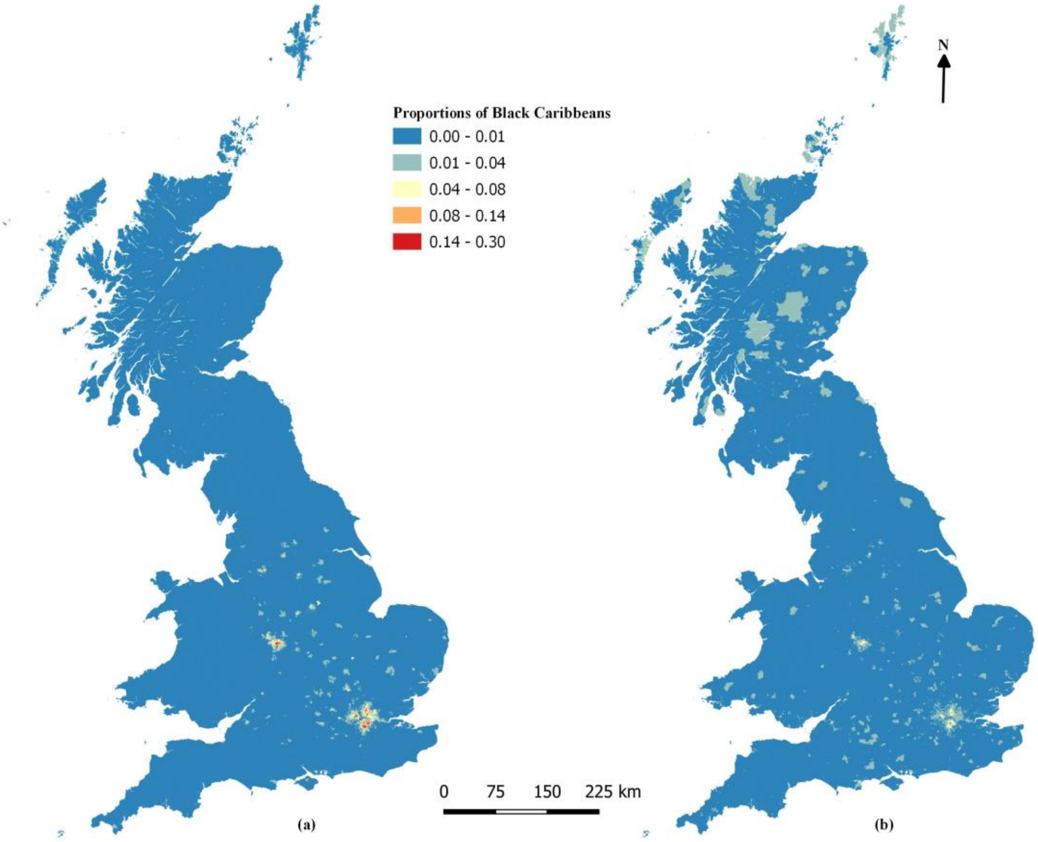

Figure 7. Distributions of the BCA group by LSOAs from the (a) 2011 Census and (b) 2011 LCR.

3.4. Summary of reassignments

Table 10 presents the combined reallocation effects of the adjustments proposed for the WIR, WBR, AIN, APK, BCA and OXX groups in this paper compared to the original EE results.

Table 11 extends Table 1 with the improvements in over- and under-predictions relative to 2011 Census figures. These figures are, of course, based upon Census aggregations and, unlike the original EE predictions, cannot be verified at the individual level. The comparison is also slightly strained by the requirement that individuals recorded in the Census are 16+ compared to 17+ in the LCRs.

However, the flows of individuals from over-represented to under-represented groups are very encouraging, as shown in Table 11. The 235,088 increase in the size of the White Irish group improves capture of WIR estimates from 54% to 99% of the recorded Census total, achieved by transfers from the over-represented White British majority group. For the Black Caribbean group, the corresponding ratio increases from 54% to 76%, with most transfers (213,569) from the White British group. Changes in the predictions of the Indian sub-continent groups are more mixed. The underestimated AAO group is improved from 30% to 91%. The overestimation of the Pakistani group is reduced from 151% to 99%, while the overestimation of the Indian group is slightly increased from 115% to 116%. Referring to Table 7, the biggest outflows from APK (291,641) and AIN (166,523) are transferred to the under-estimated Other Asian group – the size of which increases substantially. The improvement of the catch all Other (OXX) is a by-product of other reassignments. Apart from the BCA group, OXX has no outflows but increases in size following other reassignments such as the requirement that WBR names appear in the 1881 Census.

Table 10. Confusion matrix between the sizes of the GB population from the original EE (rows) and the adjusted estimates with all adjustments of the WIR, AIN, ABD, APK, BCA, WBR and OXX (columns)

	AAO	**ABD**	**ACN**	**AIN**	**APK**	**BAF**	**BCA**	**OXX**	**WAO**	**WBR**	**WIR**	**Total**	**% of GB population**
AAO	142,215	-	-	55,985	2,068	-	-	-	-	-	-	200,268	0.4%
ABD	-	285,860	-	-	-	-	-	-	-	-	-	285,860	0.6%
ACN	-	-	229,645	-	-	-	-	-	-	-	-	229,645	0.5%
AIN	166,523	-	-	1,117,631	58,873	-	-	-	-	-	-	1,343,027	2.8%
APK	291,641	-	-	179,245	719,004	-	-	-	-	-	-	1,189,890	2.5%
BAF	-	-	-	-	-	564,556	-	-	-	-	-	564,556	1.2%
BCA	-	-	-	-	-	-	153,664	114,950	-	-	-	268,614	0.6%
OXX	-	-	-	-	-	-	8,219	159,662	-	-	-	167,881	0.4%
WAO	-	-	-	-	-	-	-	-	2,273,858	-	-	2,273,858	4.8%
WBR	4,049	93	959	1,341	23	41,493	213,569	333,706	501,199	39,393,510	425,228	40,915,170	85.7%
WIR	-	-	-	-	-	-	1,793	-	-	188,347	121,515	311,655	0.7%
Total	**604,428**	**285,953**	**230,604**	**1,354,202**	**779,968**	**606,049**	**377,245**	**608,318**	**2,775,057**	**39,581,857**	**546,743**	**47,750,424**	100%
% of GB population	1.3%	0.6%	0.5%	2.8%	1.6%	1.3%	0.8%	1.3%	5.8%	82.9%	1.1%	100%	
GB Census	663,124	294,505	371,521	1,167,436	788,849	713,257	496,195	1,298,097	2,286,231	41,245,227	551,410	49,875,852	
% of GB population	1.3%	0.6%	0.7%	2.3%	1.6%	1.4%	1.0%	2.6%	4.6%	82.7%	1.1%	100.0%	
Differences between % in the LCR estimates and Census	0.0%	0.0%	-0.2%	0.5%	0.0%	-0.1%	-0.2%	-1.3%	1.2%	0.2%	0.0%	0.0%	

Table 11. Comparison of the predicted population sizes between the EE and adjusted estimates, retaining GB 2011 Census figures for comparison

	Census	**Before adjustments**		**After adjustments**	
		LCR	Ratio	LCR	Ratio
AAO	663,124	200,268	30%	604,428	91%
ABD	294,505	285,860	97%	285,953	97%
ACN	371,521	229,645	62%	230,604	62%
AIN	1,167,436	1,343,027	115%	1,354,202	116%
APK	788,849	1,189,890	151%	779,968	99%
BAF	713,257	564,556	79%	606,049	85%
BCA	496,195	268,614	54%	377,245	76%
OXX	1,298,097	167,881	13%	608,318	47%
WAO	2,286,231	2,273,858	99%	2,775,057	121%
WBR	41,245,227	40,915,170	99%	39,581,857	96%
WIR	551,410	311,655	57%	546,743	99%
Total	49,875,852	47,750,424	96%	47,750,424	96%

3.5. Enhanced estimation of countries of origin

Census categories such as the White Other Group (WAO) have been agreed by the ONS over time through consultation for policy purposes and they inevitably cannot include all groups. Blanket categorisation masks within group variation, potentially straining any assumption of within group homogeneity in research applications: for example, study of UK residential segregation (e.g., Lan et al. 2021) would likely benefit were it possible to differentiate between different groups within the ONS ‘catch all’ categories. We therefore use the WN2 data to apportion the WAO, AAO, BAF and BCA categories to probable countries of ancestral origins.

We evaluate each name pair’s relative probabilities of assignment to a specific country using similar procedures to those underpinning Equation (1). We replace the name-ethnicity lookup probabilities $p_{E,f}$ and $p_{E,s}$ with the normalised frequencies per million (FPMs) for each individual's forename and surname in the assignment process. Following extensive sensitivity analysis, we adopt 0.65 and 0.35 as the surname and forename weighting factors w_f and w_s. We retain the three most probable countries of origin: in deference to subjective self-assignments in Britain, where the most probable country estimate is inconsistent with the EE classification, we defer to the second highest country and, if necessary, the third highest. If no consistent estimate can be found the observation is assigned to the 'Any Other' (OXX) category.

Following these rules, we further disaggregate the blanket groups including AAO, BAF, BCA and WAO into countries of origin. Table 12 lists the largest populations in the 2011 LCR by country of origin within each of the four groups. We take the largest WAO group in London, the Polish community, as an example and map their geographic distribution across Greater London in 2011 in Figure 8. They were mainly concentrated in West and North London, particularly in Ealing, Brent and Waltham Forest.

Table 12. Examples of the largest populations in the 2011 LCR by country of origin within each of the AAO, BAF, BCA and WAO Census groups

Census Group	**Country**	**Population in LCR**
AAO	Sri Lanka	26,062
	Nepal	23,533
	Afghanistan	22,225
	Malaysia	21,566
	Vietnam	15,790
BAF	Nigeria	88,038
	Ghana	66,190
	Somalia	40,856
	Zimbabwe	35,171
	Uganda	17,191
BCA	Jamaica	153,513
	Trinidad and Tobago	55,886
	Haiti	12,459
	-	-
	-	-
WAO	Poland	427,545

	Italy	211,110
	Germany	96,427
	Australia	93,211
	Turkey	75,421

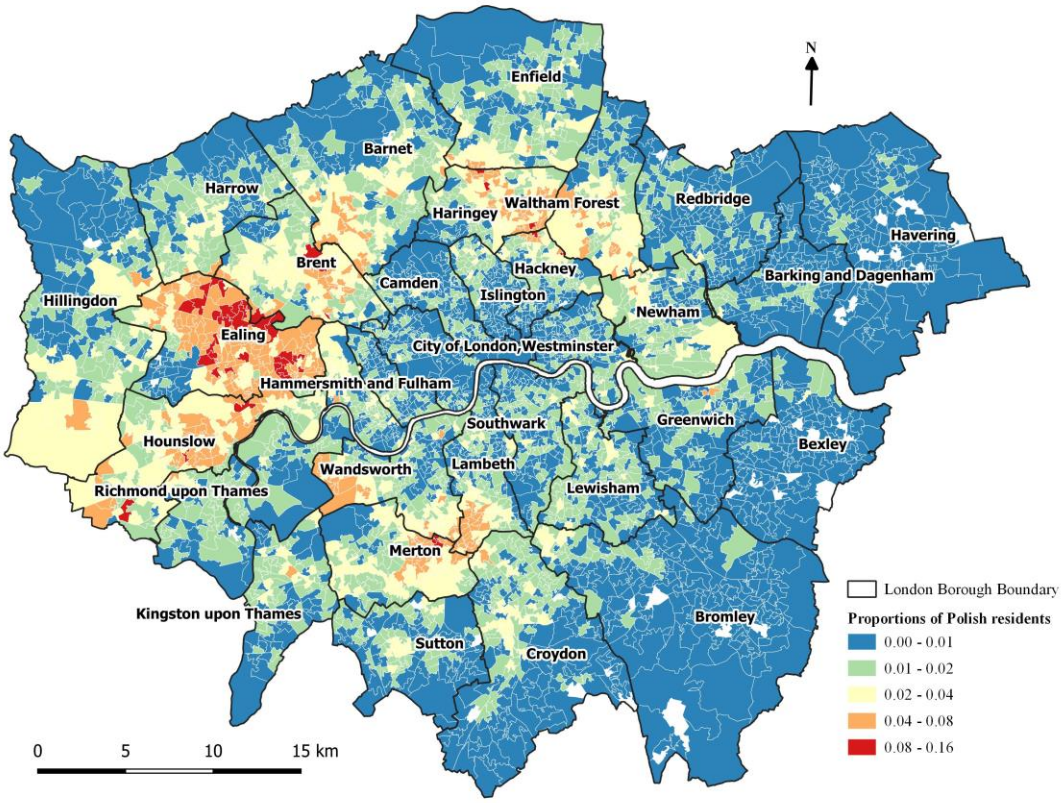

Figure 8. Distribution of Polish residents in London estimated from the 2011 LCR.

4. Validation and Discussion

It is usually difficult to obtain ground truth ethnicity data at individual level to validate the results of ethnic classification. Here we use data obtained under Freedom of Information requests pertaining to 47,979 seekers of asylum in Canada (Norris 2019), which records individuals’ names and self-reported countries of origin. Reported countries of asylum seeker origins may be imprecise (e.g., sub-continent rather than specific country) or inaccurate, particularly in instances of chain migration. Such data are thus inherently ambiguous, and also do not pertain to the UI, where strictures of General Data Protection Regulation (GDPR) make it particularly difficult to obtain names data classified by ethnicity – a Sensitive Personal characteristic under GDPR. With these caveats, we assign stated countries in the Canadian data to the 11 ONS Census groups used in EE and use our procedures to estimate group and most probable country of origin. Table 13 compares the estimates with the stated countries: the last column of the table shows the percentages of 'true positives' with an overall success of 73%, derived from the highlighted elements of the principal diagonal. The results suggest greatest success in predicting groups such as ACN, AIN, WAO, and APK: other predictions are less successful, with only about one in three of the Black Caribbean group successfully predicted. The majority of misclassifications of the BCA are assigned to the WBR group.

Table 13. Confusion matrix between our predictions for Canadian asylum seeker data and manually coded ethnicity groups based on stated country of origin.

		Estimated group												
		AAO	ABD	ACN	AIN	APK	BAF	BCA	OXX	WAO	WBR	WIR	**Total**	**% of true positives**
Coded group	AAO	**955**	26	210	526	73	99	13	191	291	101	-	2,485	38%
	ABD	44	**227**	2	51	51	6	-	2	15	3	-	401	57%
	ACN	19	-	**3,246**	5	6	9	1	-	74	16	-	3,376	96%
	AIN	81	15	6	**1,270**	84	15	2	7	31	8	-	1,519	84%
	APK	176	36	5	111	**900**	4	1	3	26	6	-	1,268	71%
	BAF	126	24	27	78	79	**2,267**	61	413	877	377	4	4,333	52%
	BCA	146	16	26	245	44	154	**1,294**	277	767	1,124	10	4,103	32%
	Arabic (OXX)	293	41	1	64	126	185	5	**451**	91	12	1	1,270	36%
	WAO	2,098	76	110	124	98	428	264	955	**24,188**	851	6	29,198	83%
	WBR	5	-	-	1	1	3	2	-	2	**11**	-	25	44%
	WIR	-	-	-	-	-	-	-	-	-	1	-	1	-
	Total	3,943	461	3,633	2,475	1,462	3,170	1,643	2,299	26,362	2,510	21	47,979	73%

We have mixed reflections on these results. Migrating and asylum-seeking are heavily selective, and the phenomenon of chain migration likely renders the dataset very noisy. Asylum seekers may be more likely to be of mixed heritage (best represented by the OXX category), something that names-based classification finds very difficult to discern. Asylum seekers may perceive their chances of success to be increased with identification with white groups – with our predictions of many ‘Other Asian’ group members to be ‘White Other’ providing a prominent example. There are also ambiguities in the assignment of countries to EE groups, such as classifying South African asylum seekers uniformly as ‘Black African’.

In some respects, data pertaining to Canadian asylum seekers present an unreasonable challenge: the ONS ethnicity classification is designed to fulfil UK needs and the prominence of the White British and White Irish groups is an irrelevant distraction in this context. In the global context, our enhancements to predictions of origins within the Indian sub-continent appear to be robust. But in other instances, the results confirm global challenges to names classification, with the inherent ambiguity of Black Caribbean names presenting a prominent example. Our own analysis of geographic variation in prediction success within Great Britain also testifies that this problem occurs across different geographic scales, and it may also be affected by changing fashions for particular forenames.

5. Conclusion

Issues of ethnicity underpin our understanding of population diversity and the regional patterning of population characteristics in the wake of recent and historic waves of migration. Elsewhere (Longley et al. 2021) we have argued that regional origins in ‘Old World’ countries have enduring inter-generational consequences for social mobility outcomes, and one of our motivations for improving the efficacy of names-based classification is to describe and evaluate the relative social circumstances of citizens who can trace their origins through any of a succession of waves of migration to the UK. As such, the creation of Onomap3 has several methodological and substantive touchpoints with research previously reported in this journal, as well as for regional science investigations more generally. Most fundamentally, the work is consistent with the view that data pertaining to human individuals, rather than aggregations of them, provide the most secure foundations to regional analysis. The advent of new sources of georeferenced data at highly disaggregate scales (Longley et al. 2018) enables new methods of conducting migration research that go far beyond early aggregate formulations in regional analysis (Greenwood and Hunt 2003). It also has potential implications for the conduct of input – output analysis (Miller and Blair 1981). Such detail and flexibility enable a much more robust and transparent definition of the urban structures that are arranged in urban hierarchies (Broitman et al. 2020), while names-based classifications enable the variegated social mixing of established populations and more recent migrants to be described and analysed (Lan et al. 2021). Our use of asylum seekers to validate the research is integral to the case for using names to identify and appraise migrant characteristics in regional analysis more generally (e.g., Lozano-Gracia et al. 2010).

In other respects, names-based classification is of strategic importance in synthesising data that are not routinely collected. Ethnicity is a sensitive personal characteristic under the General Data Protection Regulation (GDPR), and our experience is that names classifications become essential when data collection about ethnicity has not been considered proportionate in service delivery, but subsequently becomes essential in unforeseen social equity audits or health care studies. Our own involvement in auditing the rehousing decisions made post the Grenfell Tower disaster and evaluating hospitalisation outcomes during the COVID-19 pandemic (Thomas et al. 2021) provide prominent examples. In future, the development of trusted research environments (TREs, see Chalstrey 2021) may provide data linkage solutions, but in the meantime names-based classification provides the only expedient solution, particularly in emergency situations.

In methodological terms, the research reported here provides several lessons to guide this quest. It is widely understood that the heterogeneity of ethnic groups varies geographically, and our work highlights that names-based classification should be cognisant of context: our prediction success is better for Great Britain – the territory for which it was intended – than Canada, yet this focus allows issues of self-assignment in particular cultural contexts to be incorporated, analysed and evaluated. The WN2 data present global evidence of the need to reweight the relative importance of forenames and surnames for some origin jurisdictions and we acknowledge that there is scope for further empirical refinement of the procedures developed here. Our sensitivity analysis and evaluation of results rely upon visual interpretation of mapped results alongside aggregate numerical comparisons. This approach might be supplemented in future research by the use of optimisation criteria and weightings to prioritise assignments (or 'near misses') to particular groups of interest. Future research might also address issues arising from transliteration of names (O'Brien and Longley 2018), homonymic family names, the mutation of family names over time and following migration over space, and cultural practices in assembling unique forenames or surnames.

Our approach is guided by the virtue of retaining self-assignments of census respondents in England and Wales while expanding and future-proofing the dictionary of names to include current popular forenames as well as new names imported into Britain from abroad. The classification is thus data led but also guided by GB cultural conventions. Issues of self-assignment may reinforce apparent inequalities of outcome or (as in COVID-19) set researchers on a search for physiological sources to societal problems. Yet our own view is that these issues are best addressed through classifications that are robust, transparent and open to scrutiny and that evaluations such as ours are instructive to minimise risks of misuse or misinterpretation.

Our own motivation for this work is to develop tools to understand the processes that underpin inter-generational inequalities of social mobility outcomes in Great Britain, at geographical and ethnic granularities that range from the effects of local ancestral origins of long-established populations through the inter-generational outcomes experienced by Irish migrants through to the outcomes of global migration in the 20th and 21st centuries. We intend this paper as a contribution to justify the approaches we are taking in this endeavour but hope that it stimulates wider debate about the value and veracity of names-based classification in the widest range of investigations into issues of social equity.

6. References:

Broitman, D., Benenson, I. and Czamanski, D., 2020. 'The impact of migration and innovations on the life cycles and size distribution of cities.' *International Regional Science Review* 43(5): 531-549. https://doi.org/10.1177/0160017620914061.

Chalstrey, Ed. 2021. *Developing and Publishing Code for Trusted Research Environments: Best Practices and Ways of Working.* The Alan Turing Institute (The Alan Turing Institute). https://www.turing.ac.uk/research/publications/developing-and-publishing-code-trusted-research-environments.

Clark, Gregory, and Neil Cummins. 2015. 'Intergenerational wealth mobility in England, 1858–2012: surnames and social mobility.' *The Economic Journal* 125 (582): 61-85. https://doi.org/10.1111/ecoj.12165.

Finney, Nissa, and Ludi Simpson. 2009. *'Sleepwalking to segregation'? Challenging myths about race and migration*. Policy Press at the University of Bristol.

Greenwood, M.J. and Hunt, G.L., 2003. 'The early history of migration research'. *International Regional Science Review*, 26(1): 3-37. https://doi.org/10.1177/0160017602238983.

Higgs, E., and K. Schurer. 2019. Integrated Census Microdata (I-CeM), 1851-1911. edited by UK Data Service. UK: UK Data Service.

Kandt, Jens, and Paul A. Longley. 2018. 'Ethnicity estimation using family naming practices.' *PLOS ONE* 13 (8): e0201774. https://doi.org/10.1371/journal.pone.0201774.

Lan, Tian, Jens Kandt, and Paul Longley. 2020. 'Geographic scales of residential segregation in English cities.' *Urban Geography* 41 (1): 103-123. https://doi.org/10.1080/02723638.2019.1645554.

---. 2021. 'Measuring the Changing Pattern of Ethnic Segregation in England and Wales with Consumer Registers.' *Environment and Planning B: Urban Analytics and City Science.* 48 (6): 1591-1608. https://doi.org/10.1177/2399808320919774.

Lan, T., van Dijk, J. and Longley, P. 2021. 'Family names, city size distributions and residential differentiation in Great Britain, 1881–1901.' *Urban Studies*, Online First. https://doi.org/10.1177/00420980211025721.

Lansley, Guy, Wen Li, and Paul A. Longley. 2019. 'Creating a linked consumer register for granular demographic analysis.' *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182 (4): 1587-1605. https://doi.org/10.1111/rssa.12476.

Longley P, Cheshire, J. and Singleton, A. 2018. Consumer Data Research. London: UCL Press.

Longley, P. A., van Dijk, J., & Lan, T. 2021. ‘The geography of intergenerational social mobility in Britain.’ *Nature communications* 12(1): 1-8. https://doi.org/10.1038/s41467-021-26185-z

Lozano-Gracia, N., Piras, G., Ibáñez, A.M. and Hewings, G.J. 2010. ‘The journey to safety: conflict-driven migration flows in Colombia.’ *International Regional Science Review* 33(2): 157-180. https://doi.org/10.1177/0160017609336998.

Mateos, Pablo, Paul A. Longley, and David O'Sullivan. 2011. ‘Ethnicity and Population Structure in Personal Naming Networks.’ *PLOS ONE* 6 (9): e22943. https://doi.org/10.1371/journal.pone.0022943.

Mateos, Pablo, Alex Singleton, and Paul Longley. 2009. ‘Uncertainty in the Analysis of Ethnicity Classifications: Issues of Extent and Aggregation of Ethnic Groups.’ *Journal of Ethnic and Migration Studies* 35 (9): 1437-1460. https://doi.org/10.1080/13691830903125919.

Miller, R.E. and Blair, P. 1981. ‘Spatial aggregation in interregional input-output models.’ *Papers of the Regional Science Association* 48(1): 149-164. https://doi.org/10.1111/j.1435-5597.1981.tb01152.x.

Norris, Samuel. 2019. ‘Examiner inconsistency: Evidence from refugee appeals.’ *University of Chicago, Becker Friedman Institute for Economics Working Paper* (2018-75). https://doi.org/10.2139/ssrn.3267611. https://ssrn.com/abstract=3267611.

O'Brien, Oliver, and Paul Longley. 2018. ‘Given and Family Names as Global Spatial Data Infrastructure.’ In *Consumer Data Research*, edited by Paul Longley, James Cheshire and Alex Singleton, 53-67. London: UCL Press.

Office for National Statistics. 2009. Final recommended questions for the 2011 Census in England and Wales: Ethnic group. edited by Office for National Statistics.

---. 2017. Population estimates by ethnic group.

Parameshwaran, Meenakshi, and Per Engzell. 2015. ‘Ethnicity in England: What parents' country of birth can and can't tell us about their children's ethnic identification.’ *Journal of Ethnic and Migration Studies* 41 (3): 399-424. https://doi.org/10.1080/1369183X.2014.920690.

Petersen, Jakob, Jens Kandt, and Paul A. Longley. 2021. ‘Ethnic inequalities in hospital admissions in England: an observational study.’ *BMC Public Health* 21 (1): 862. https://doi.org/10.1186/s12889-021-10923-5.

Ritchie, Felix. 2008. ‘Secure access to confidential microdata: four years of the Virtual Microdata Laboratory.’ *Economic & Labour Market Review* 2 (5): 29-34. https://doi.org/10.1057/elmr.2008.73.

Thomas, Daniel Rh, Oghogho Orife, Amy Plimmer, Christopher Williams, George Karani, Meirion R Evans, Paul Longley, Janusz Janiec, Roiyah Saltus, and Ananda Giri Shankar. 2021. ‘Ethnic variation in outcome of people hospitalised during the first COVID-19 epidemic wave in Wales (UK): an analysis of national surveillance data using Onomap, a name-based ethnicity classification tool.’ *BMJ Open* 11 (8): e048335. https://doi.org/10.1136/bmjopen-2020-048335.

Van Dijk, Justin, Guy Lansley, and Paul A. Longley. 2021. ‘Using linked consumer registers to estimate residential moves in the United Kingdom.’ *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. https://doi.org/https://doi.org/10.1111/rssa.12713.

Wilson, Nick, Mike Wright, and Marek Kacer. 2018. ‘The equity gap and knowledge-based firms.’ *Journal of Corporate Finance* 50: 626-649. https://doi.org/https://doi.org/10.1016/j.jcorpfin.2017.12.008.

Winney, Bruce, Abdelhamid Boumertit, Tammy Day, Dan Davison, Chikodi Echeta, Irina Evseeva, Katarzyna Hutnik, Stephen Leslie, Kristin Nicodemus, Ellen C. Royrvik, Susan Tonks, Xiaofeng Yang, James Cheshire, Paul Longley, Pablo Mateos, Alexandra Groom, Caroline Relton, D. Tim Bishop, Kathryn Black, Emma Northwood, Louise Parkinson, Timothy M. Frayling, Anna Steele, Julian R. Sampson, Turi King, Ron Dixon, Derek Middleton, Barbara Jennings, Rory Bowden, Peter Donnelly, and Walter Bodmer. 2012. ‘People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population.’ *European Journal of Human Genetics* 20 (2): 203-210. https://doi.org/10.1038/ejhg.2011.127.

Yemane, Ruta, and Mariña Fernández-Reino. 2021. ‘Latinos in the United States and in Spain: the impact of ethnic group stereotypes on labour market outcomes.’ *Journal of Ethnic and Migration Studies* 47 (6): 1240-1260. https://doi.org/10.1080/1369183X.2019.1622806.