UCL CENTRE FOR ADVANCED SPATIAL ANALYSIS

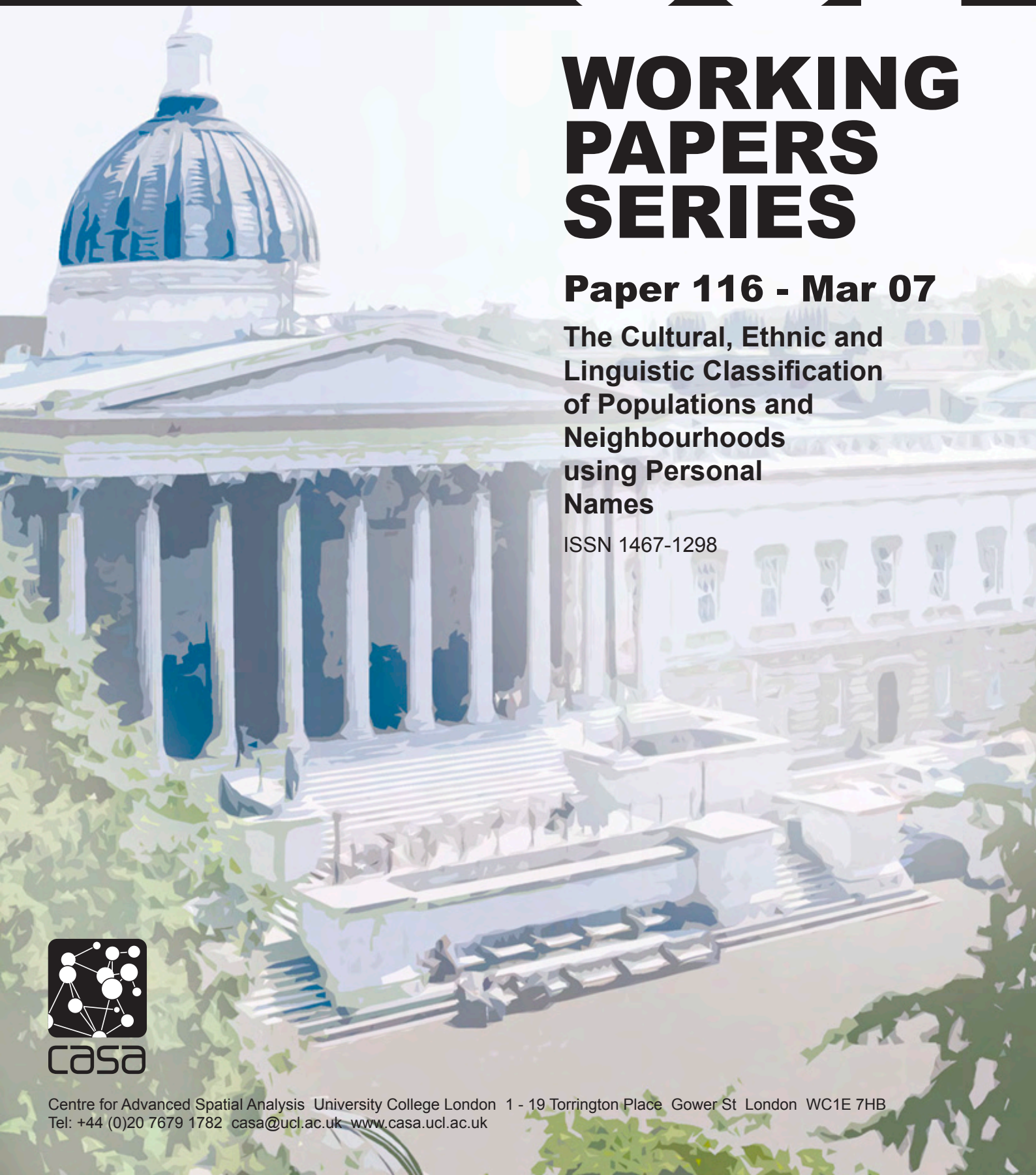
UCL

WORKING PAPERS SERIES

Paper 116 - Mar 07

The Cultural, Ethnic and Linguistic Classification of Populations and Neighbourhoods using Personal Names

ISSN 1467-1298



casa

Centre for Advanced Spatial Analysis University College London 1 - 19 Torrington Place Gower St London WC1E 7HB
Tel: +44 (0)20 7679 1782 casa@ucl.ac.uk www.casa.ucl.ac.uk

The Cultural, Ethnic and Linguistic Classification of Populations and Neighbourhoods using Personal Names

Pablo Mateos *, Richard Webber and Paul Longley

Department of Geography and Centre for Advanced Spatial Analysis
University College London
Gower Street, London WC1E 6BT
* Corresponding author: p.mateos@ucl.ac.uk

CASA Working Paper 116

March 2007

Available at:
http://www.casa.ucl.ac.uk/working_papers/paper116.pdf

Abstract

There are growing needs to understand the nature and detailed composition of ethnic groups in today's increasingly multicultural societies. Ethnicity classifications are often hotly contested, but still greater problems arise from the quality and availability of classifications, with knock on consequences for our ability meaningfully to subdivide populations. Name analysis and classification has been proposed as one efficient method of achieving such subdivisions in the absence of ethnicity data, and may be especially pertinent to public health and demographic applications. However, previous approaches to name analysis have been designed to identify one or a small number of ethnic minorities, and not complete populations.

This working paper presents a new methodology to classify the UK population and neighbourhoods into groups of common origin using surnames and forenames. It proposes a new ontology of ethnicity that combines some of its multidimensional facets; language, religion, geographical region, and culture. It uses data collected at very fine temporal and spatial scales, and made available, subject to safeguards, at the level of the individual. Such individuals are classified into 185 independently assigned categories of Cultural Ethnic and Linguistic (CEL) groups, based on the probable origins of names. We include a justification for the need of classifying ethnicity, a proposed CEL taxonomy, a description of how the CEL classification was built and applied, a preliminary external validation, and some examples of current and potential applications.

Table of Contents

List of Tables

List of Figures

1 Introduction

This working paper presents a new methodology to classify the UK population and neighbourhoods into groups of common origin using surnames and forenames, termed the 'Cultural, Ethnic and Linguistic'(CEL) Taxonomy. It proposes a new ontology of ethnicity that is multidimensional in nature, assimilating aspects of language, religion, geographical region and culture through the shared characteristics of names. Names are currently classified into 185 independently assigned CEL categories. The paper includes an exhaustive explanation of the tools and techniques used to build this classification, using data collected at very fine temporal and spatial resolutions. The research presented here includes current work in progress at University College London to optimise the automatic classification of individuals into CEL categories, and should be not taken as in any sense complete; rather it documents a series of heuristics developed in an essentially ad hoc manner, that we believe help us to understand and capture the diversity in worldwide naming practices.

The paper is structured in seven sections. Section 1 includes a statement of motivation for classifying ethnicity and a brief review of name-based methods. Section 2 describes the 'Cultural, Ethnic and Linguistic' (CEL) Taxonomy and discusses the data sources used in the research. Section 3 introduces the seven techniques used to classify names into CEL categories, while Section 4 provides a detailed explanation of the heuristics that underpin the CEL classification, in three distinct stages. Sections 3 and 4 consolidate the core methodology presented in the paper. Section 5 provides a preliminary validation using Census data. Section 6 outlines some current and potential applications for this methodology, and Section 7 offers some concluding remarks. A full list of the taxonomy of 185 CELs is included in an Appendix at the end of the paper.

1.1 The need for ethnicity classifications

Two major events in 2005 reopened a long-standing debate about the model of multicultural societies in Europe; the London bombings of July 7th, and the urban riots in France later in November of that year. These events triggered a heated public debate that focused diverse issues upon an apparent failure of European society to assimilate immigrant communities (Leppard, 2005, The Economist, 2005). Furthermore, rare goes the

day without headlines in the European media about issues related to immigration, ethnic minorities or religion, portrayed as somehow 'problematic issues' either in policy debates or in the streets, even resulting in a change of government in the case of the Netherlands (The Economist, 2006).

Behind this intense debate, in a context of a rapidly changing multicultural Europe, it is likely that there lie too many prejudices and too little evidence. One of the main causes of the dearth of evidence about immigration, ethnicity and religious observance is the difficulty of defining members of such groups, in ways that are robust and defensible to scrutiny. This is the much contested arena of 'identity politics'(Brubaker, 2004), where groups often lobby for official recognition as a precursor to claiming collective rights (Skerry, 2000). In some countries, such as France, the State refuses to acknowledge different identities within an otherwise equal society, in the interest of promulgating an egalitarian republic (Haut Conseil à l'Integration, 1991).

Ethnicity is a multi-dimensional concept that encompasses different aspects of group identity, in relation with kinship, religion, language, shared territory, nationality, and physical appearance (Bulmer, 1996). Measuring ethnicity is problematic because of the subjective, multi-faceted and changing nature of ethnic identification, and because there is no clear consensus on what constitutes an 'ethnic group' (Coleman and Salt, 1996, ONS, 2003). Despite these evident difficulties, ethnicity is today measured for a wide range of purposes in many countries, and governmental statisticians try to respond to surges of interest in collective identity formation and the struggle of States to monitor and sometimes help to shape these processes (Kertzer and Arel, 2002).

Ethnicity is usually measured as a single variable, an 'ethnic group' into which the individual self-assigns his or herself from a narrow typology of discrete classes, with scant regard to the richness and multi-faceted nature of the underlying phenomenon. Ethnic classifications are used, rather than open questions, in order to arrange data according to common features, and to facilitate the comparative consistency of the resulting statistics over time and between different sources (ONS, 2003). To the inevitable simplifications that arise from measuring ethnicity as a single variable must be added the highly contested issue of assignment to discrete categories – an issue that is highly contested and that involves decisions in the arena of identity politics (Kertzer and Arel, 2002). Bhopal et al

(2004) state that, however carefully or elaborately defined, ethnic classifications bear no direct correspondence with cultural, linguistic, dietary or religious preferences, of key interest for epidemiological research. Aspinall (2000) contends that most ethnic groupings hide massive within group heterogeneity, diminishing the value of ethnic categorisation as a way of delivering culturally appropriate health care, and in understanding the causes of ethnic variations in disease. A third problem comes with the method of self-assessment of ethnicity (as opposed to it being assigned by a third person or a computer), because perceptions of identity change over time (Aspinall, 2000) and according to the type of ethnicity question asked, the definitions of categories offered (Olson, 2002), and the method of data collection.

Despite all these issues, there is a general consensus that measuring ethnicity is vitally important for the provision of equitable public services for an increasing multicultural population (Mason, 2003), the eradication of discrimination (Parsons et al, 2004), and to build accurate demographic forecasts for the whole population (Coleman, 2006). Furthermore, the *de facto* 'gold standard' for such measurement usually emanates from the ethnic categories created by the national population censuses (Kertzer and Arel, 2002). The UK Office for National Statistics recognises that this measurement should be done in a way that is sound, sensitive, relevant, useful, and consistent over some period of time (ONS, 2003).

The ethnic classification currently used by most UK public bodies and many private institutions is that of the 2001 Census of Population, which included a question on ethnicity for the second time in history, along with religion (asked for the first time after over a century in 2001) and country of birth. Despite the census classification having become the standard for ethnic information collection, ethnic group is still not recorded in most routine basic population registers, such as birth, death, electoral and general practice registrations (London Health Observatory, 2003, Nanchahal et al, 2001). In the health arena, collection of this information has been mandatory in hospital admissions since 1995 (NHS Executive, 1994), yet it still is recorded for only 74% of events (London Health Observatory, 2005) and to only a low quality when compared with other research sources (Bhopal et al, 2004).

Table 1 shows the results of a recent study by the Association of Public Health Observatories, analysing the percentage of records with incomplete ethnicity coding in eight different datasets. The study concludes that a substantial proportion of events are not being assigned to an ethnic group, and that this failure is attributable to organisational issues, rather than the size of ethnic minority groups at the local level (APHO, 2005, Association of Public Health Observatories, 2005). However, these datasets are the exception rather than the norm, and in the majority of datasets available to social science as well as health researchers, ethnicity data are simply not recorded at all (Bhopal et al, 2004, Harding et al, 1999).

'Population' Dataset	***England***	***London***
Pupil Level Annual School census, 2004		
Primary schools	2.3	1.6
Secondary schools	3.4	2.5
Educational attainment/PLASC 2003	5.7	3.9
Children in need 2003	8	8
Enhanced TB Surveillance 2000-02	6.6	5
AIDS/HIV: SOPHID data 2003	3	4
Drug misuse: NDTMS data	15.6	9.5
Social Services Workforce 2004	8.9	7.1
Non-Medical Workforce 2004	11.7	16.8
Medical & Dental workforce 2004	2	1.9
Hospital Episode Statistics, 2003/04	36	34

Table 1: Percentage of records with incomplete ethnicity coding in different datasets
Source: (Association of Public Health Observatories, 2005, 12)

In the absence of ethnicity data, other proxies, such us country of birth, have been used to ascribe a person's ethnicity (Marmot et al, 1984, Wild and McKeigue, 1997). Despite its utility to classify migrant origins, the reliability of this indicator is eroding (Harding et al, 1999) with growing numbers of second generation migrants, an increasing proportion of 'white British' people born abroad, and migrants being born in 'intermediate' countries (i.e. East African Indians). In the 2001 Census only half of the minority ethnic population was born outside the UK. Many health studies use death certificate data on country of birth: such data are reliant upon an informant and may be less accurate Census measures, where the person is still alive to provide the information (Gill et al, 2005) – albeit possibly not consulted by the householder who completes the questionnaire.

Another method employed as a proxy for ethnicity is the analysis of name origins, which in particular has been used to identify South Asian, Chinese and Hispanic populations, with different degrees of accuracy. This research seeks to contribute to this approach, and this will be the theme of the rest of this paper.

The different dimensions that define ethnicity are usually summarized as kinship, religion, language, shared territory, nationality, and physical appearance (Bulmer, 1996). In principle one could accurately ascribe a person to an ethnic group if these six dimensions were to be measured separately. This conclusion has been reached by several studies of ethnic inequalities in health (Bhopal, 2004, Gerrish, 2000, McAuley et al, 1996) that lead investigators to use a range of variables in the measurement of ethnicity as a multi-dimensional phenomena, instead of just one, measuring separately; language, religion, country of birth, family origins, and length of residence. Physical appearance seems to be a much more sensitive aspect to ask about, and even more so to classify. Four of these dimensions – language, religion, country of birth, family origins – are manifest to some extent in the forenames and surnames that we all carry, and hence may be deemed to be a useful proxy for them. In fact, this was the approach taken in a study commissioned by the US Senate in the 1930's. It estimated the ethnic composition of the "original national stock" of the population of the United States, through the origin of surnames in the 1790 Census, upon which the US government based their new immigration quota restrictions from 1932 (American Council of Learned Societies, 1932, US Senate, 1928). Since these studies in the first third of the 1930's there have been different successful attempts to provide such ethnicity classifications based on names.

1.2 The need for alternative ethnicity classifications; name-based methods

A thorough review of the literature of the measurement of ethnicity and of the name origin techniques used in demography, epidemiology and genetic studies, is presented in Mateos (2007a). It concludes that name-based ethnicity classification methods present a valid technique that relates individuals to ethnic groups through the classification of their name origins. Some of the methods provide a high degree of reliability in the assignment of an ethnic group to individual names, while others offer the probable religion and language associated with each group of names. However, none of them was designed for the task of classifying entire populations into ethnic groups, instead focusing on the identification of

one or just a few ethnic minority groups, rather than discriminating between all of the potential groups present in a given population. Amongst the most studied groups in some of the main immigration countries (US, Canada, Australia, UK Netherlands and Germany) are: South Asian (Indian, Pakistani, Bangladeshi, Sri Lankan), Chinese, other East and South-east Asian (Vietnamese, Japanese, Korean, and Filipino), Hispanics, Turks, and Jews. However, as stated, each individual classification attempts only to focus upon one of these groups, and not all of them (and more) at the same time. In order to create a true population name classification system, the name reference list upon which it is to be built needs to be sourced using a large number of names covering a entire society, and such classification has to seek to accommodate all the potential ethnic groups present in a society.

This is the task that this research has investigated for the entire population resident in the UK, through a methodology that will be described and discussed in this paper. This research develops a new name-based ethnicity classification for the most common surnames and forenames present in Britain, which have been assigned to a large number of cultural, ethnic and linguistic categories. This paper describes in detail the methods employed to build a prototype *Cultural, Ethnic and Linguistic* classification (CEL), and also presents a validation of the classification using internal and external datasets, before describing some representative applications and overall conclusions.

Our basic hypothesis is that the classification of surnames and forenames into ancestral groups creates valuable insights when ethnicity, linguistic or religious data are not available at appropriate temporal, spatial or nominal (number of categories) resolutions. Related to this, we contend that this method is suited to subdivision of populations and classification of neighbourhoods into groups of common origin. Furthermore, we contend that this methodology offers an advantage over traditional information sources such as the UK Census of Population, since it: develops a more detailed and meaningful classification of people's origins categories; offers improved updating (annually through electoral or patient registers); better accommodates changing perceptions of identity than self-classification of ethnicity (through independent assignment of ethnicity and or cultural origins according to name); and is made available at the individual or the UK postcode unit level (average of 30 people) rather than the Output Area (150 people).

2 The CEL Taxonomy and Data Sources

This section explains the concepts used to formalise a new classification of names in cultural, ethnic or linguistic groupings, termed 'CELs', including the development of a taxonomy of CELs and the data sources utilised.

Hereinafter two types of people's names will be distinguished and denoted as follows; *surnames* (also known as family names or last names), which normally correspond to the components of a person's name inherited from his or her family, and *forenames* (also known as first names, given names, or Christian names), which refer to the proper name given to a person usually at birth.

2.1 The Concept of CEL and its Taxonomy

The term 'CEL', as used in this paper, is used as shorthand for a 'Cultural, Ethnic or Linguistic' groupings, a concept first introduced in Hanks' (2003) Dictionary of American Family Names (DAFN) as a basis for the analysis and classification of surnames (Tucker, 2003). The principal purpose of the development of the CEL concept by the compliers of the DAFN was to divide each of the 70,000 surnames in the dictionary into 23 general groups of origin defined by any of these three general dimensions (Culture, Ethnicity or Language). Each of these 23 CEL groups corresponds to each of the etymology specialists to whom the names were referred for the purpose of writing the description of the etymological origins of each name and assigning them to 74 subgroups or finer CELs (for a list of the 74 CELs see Hanks and Tucker, 2000).

As a result DAFN comprises 70,000 entries that follow the pattern of the following example:

> **Abadi** (147) **1.** Arabic: denoting someone whose ancestors belonged to the 'Abbad tribe (see Abad). **2.** Jewish (Sephardic): adoption of the Arabic surname.
>
> Given Names: Arabic 27%; Jewish 11%.

The first number in brackets (147) is the frequency of the surname in the U.S. telephone directory, and the percentages listed as Given Names are the proportions of those 147 people whose forenames are deemed to belong to the top CELs (those with a value equal or above 4%), in the example given; Arabic and Jewish.

In this research the CEL concept is used as a basis for classifying both forenames and surnames currently present in the UK, defined as those names of UK residents with 3 or more occurrences. Each CEL is used to define a human group whose names share a common origin in terms of their culture, ethnicity or language, and is judged to be distinct enough from other CELs along one or several of these dimensions. The CEL concept summarizes four dimensions of a person's identity: a religious tradition, a geographic origin, an ethnic background - usually reflected by a common ancestry (genealogical or anthropological links) - and a language (or common linguistic heritage). These four dimensions define a CEL; religion, geography, ethnicity and language, the "trail" of which can today be discerned from the characteristics of the forenames or surnames that belong to each CEL. These characteristics can be a name's morphology (elements, letters patterns, endings, stems, etc), its etymology (meaning and origin), and its historic or current geographic distribution (other more subtle characteristics such as phonetic or calligraphic differences are not considered here). These characteristics are the 'raw materials' used in the field of onomastics, a division of linguistics which deals with the study of the origins and forms of proper names.

The criterion used to create the CEL taxonomy, both in DAFN and in the research presented here, is primarily an onomastic one, that is, a list of human groups based on name origins. The CEL taxonomy created in this research is based on the empirical analysis of name characteristics, grouping them in a way that maximises each group's homogeneity along the four dimensions of human origins (geography, religion, ethnicity and language) identified above. A subset of the four dimensions may be allowed to dominate in the classification of a particular name. This approach produces a taxonomy of CELs that is hierarchical and varies in scope of detail from very fine categories (e.g. Cornish, Romania Transylvania or Sephardic Jew) to very broad ones that overarch others (e.g. Muslim or European), as to best represent the common aspects shared by homogeneous groups of names present in Western Societies.

The taxonomy is exhaustive but not fixed, in that new CELs can be created through the classification process as a sufficient number of names with distinct commonalities are either newly gathered or spun off from a pre-existing CEL category. The CEL taxonomy presented here is optimised for the names present in the contemporary UK population, and

currently includes 185 CEL categories of which 7 describe different aspects of 'void or unclassified names' and 178 'true' CELs (see Table 2 for the complete list). The resulting CEL taxonomy is thus comprised of a series of homogenous categories of various resolutions (in terms of size and scope) that primarily follow an onomastic criterion to classify names according to their common origins. The individual CELs form the building blocks of a multidimensional system, in which they can be aggregated into higher level groups not only following onomastic criteria, as applied here, but also using alternative combinations according to religious, geographic, ethnic or linguistic criteria. These different aggregations of CELs can then be applied to classify a population according to the criterion that best fits the purpose of each application (see Table 7 in the Appendix for the correspondence between CELs and the different aggregations proposed).

The process by which the CEL Taxonomy was created is therefore a heuristic one, and has been developed in parallel with the overall classification of names, since the original very coarse groupings of languages, religions or continents (e.g. Hispanic, Muslim, or African categories) have been subdivided into finer categories during the process by which the classification rules explained in Section 3 and section 4 shed new light upon the homogeneous characteristics of subgroups of names. As a result of this process, a categorization of 185 CELs has been created, termed here 'CEL Types', which are grouped into 15 coarser categories according to onomastic criteria and termed here 'CEL Group'. A list of these CEL Types, ordered by CEL Group, is presented in Table 2, while the full details by CEL Type are described in Table 7 in the Appendix.

After defining the CEL concept and generating a taxonomy of CELs, the next step in the research was to classify the most common forenames and surnames present in the UK into CELs in order to create a 'Name-to-CEL dictionary' that could then act as a reference list to classify target populations.

CEL GROUP	CEL TYPE
AFRICAN	AFRICA, BENIN, BLACK SOUTHERN AFRICA, BOTSWANA, BURUNDI, CAMEROON, CONGO, ETHIOPIA, GAMBIA, GHANA, GUINEA, IVORY COAST, KENYAN AFRICAN, LIBERIA, MADAGASCAR, MALAWI, MOZAMBIQUE, NAMIBIA, NIGERIA, OTHER AFRICAN, RWANDA, SENEGAL, SIERRA LEONE, SWAZILAND, TANZANIA, UGANDA, ZAIRE, ZAMBIA, ZIMBABWE
CELTIC	CELTIC, IRELAND, NORTHERN IRELAND, SCOTLAND, WALES
ENGLISH	BLACK CARIBBEAN, BRITISH SOUTH AFRICA, CHANNEL ISLANDS, CORNWALL, ENGLAND
EUROPEAN	AFRIKAANS, ALBANIA, AZERBAIJAN, BALKAN, BELARUS, BELGIUM, BELGIUM (FLEMISH), BELGIUM (WALLOON), BOSNIA AND HERZEGOVINA, BRETON, BULGARIA, CANADA, CROATIA, CZECH REPUBLIC, ESTONIA, EUROPEAN, FRANCE, FRENCH CARIBBEAN, GEORGIA, GERMANY, HUNGARY, ITALY, LATVIA, LITHUANIA, MACEDONIA, MALTA, MONTENEGRO, NETHERLANDS, POLAND, ROMANIA, ROMANIA BANAT, ROMANIA DOBREGA, ROMANIA MANAMURESCRIANA, ROMANIA MOLDOVA, ROMANIA MUNTENIA, ROMANIA TRANSILVANIA, RUSSIA, SERBIA, SLOVAKIA, SLOVENIA, SWITZERLAND, UKRAINE, YUGOSLAVIA
NORDIC	DENMARK, FINLAND, ICELAND, NORDIC, NORWAY, SWEDEN
GREEK	GREECE, GREEK CYPRUS
HISPANIC	ANGOLA, BASQUE, BELIZE, BRAZIL, CASTILLIAN, CATALAN, COLOMBIA, CUBA, GALICIAN, GOA, HISPANIC, LATIN AMERICA, PHILIPPINES, PORTUGAL, SPAIN
JEWISH OR ARMENIAN	ARMENIAN, JEWISH, SEPHARDIC JEWISH
MUSLIM	AFGHANISTAN, ALGERIA, BALKAN MUSLIM, BANGLADESH MUSLIM, EGYPT, ERITREA, IRAN, IRAQ, JORDAN, KAZAKHSTAN, KUWAIT, KYRGYZSTAN, LEBANON, LIBYA, MALAYSIAN MUSLIM, MIDDLE EAST, MOROCCO, MUSLIM, MUSLIM INDIAN, MUSLIM INDIAN, MUSLIM OTHER, OMAN, PAKISTAN, PAKISTANI KASHMIR, SAUDI ARABIA, SOMALIA, SUDAN, SYRIA, TUNISIA, TURKEY, TURKISH CYPRUS, TURKMENISTAN, UNITED ARAB EMIRATES, UZBEKISTAN, WEST AFRICAN, WEST AFRICAN MUSLIM, YEMEN
SIKH	INDIA SIKH
SOUTH ASIAN	ASIAN CARIBBEAN, BANGLADESH HINDU, BHUTAN, GUYANA, HINDU NOT INDIA, INDIA HINDI, INDIA NORTH, INDIA SOUTH, KENYAN ASIAN, MAURITIUS, NEPAL, SEYCHELLES, SOUTH ASIAN, SRI LANKA
JAPANESE	JAPAN
EAST ASIAN	CHINA, EAST ASIA, EAST ASIAN CARIBBEAN, FIJI, HONG KONG, INDONESIA, MALAY, MALAYSIAN CHINESE, MYANMAR, POLYNESIA, SINGAPORE, SOLOMON ISLANDS, SOUTH KOREA, THAILAND, TIBET, VIETNAM
INTERNATIONAL	INTERNATIONAL
VOID AND UNCLASSIFIED	UNCLASSIFIED, VOID, VOID - SURNAME, VOID INITIAL, VOID OTHER, VOID PERSONAL NAME, VOID TITLE

Table 2 The CEL Type taxonomy and its groupings into CEL Groups

See Table 7 in the Appendix for a full lookup table between CEL Types and their various groupings

2.2 Reference and Target Populations

The literature review conducted by Mateos (2007a) has suggested that two main types of population datasets are required in order to build a new name classification: a reference and a target population. The *reference population* is a list of names of individuals with their ethnicity, or a proxy for it (e.g. country of birth), that is used to build a unique Name-to-ethnicity *reference list*. On the contrary, the *target population* is just used for validation purposes, to evaluate the accuracy of the reference list. The target population has to be independently sourced from the reference population, and it must also contain names of individuals and their ethnicity or a proxy for it, but always obtained via non-name methods (self-reported, country of birth, nationality, third-person reported, etc). Therefore, the *target population* is classified into ethnic groups according to the name categories in the *reference list* and compared with the 'true ethnicity'.

In the 13 studies described in Mateos (2007a), such 'true ethnicity' (or a proxy for it) in both the reference and target populations had to be previously known using an independent method (i.e. not name-based). This research aims to classify the whole population of the UK into ethnic groups based on names, and there is only one dataset that covers the whole population and collects ethnicity at the individual, the decennial Census of Population. However, for reasons of privacy protection, individual ethnicity and name data are not available until 100 years after the Census is carried out. Therefore, in this research the objective of creating a classification that guarantees near total population coverage is intrinsically at odds with the possibility of accessing a total population dataset of names that also includes the individuals' ethnicity. Therefore, in this paper a reference population with total population coverage of individual names but without any 'true ethnicity' information will be used. The names in such reference population will be classified following an onomastic approach, in other words, names will be classified according to their intrinsic characteristics (morphology, etymology, geographic distribution, etc) rather than the ethnicity reported by their bearers.

2.3 Sources of Data: Reference Population

The data sources used to build the name reference lists used for this research are comprised of name frequency datasets with high reference population coverage and at various

temporal and spatial resolutions for different English speaking countries (derived from the Electoral Roll or Telephone Directories). These data sources are listed in Table 3, which also includes other characteristics such as the number of names included, and their temporal and geographic coverage. These datasets were obtained under a variety of use conditions from the data providers which restricted the level of disaggregation, as described in Table 3 ('resolution' columns), or the locations and methods of data manipulation under different data sharing protocols.

The major source of data amongst those listed in Table 3 has been the Electoral Register for Great Britain, both in its 1998 and 2004 editions. The purpose of these registers is to record the names and addresses of British and foreign citizens entitled to vote in local or national elections in Great Britain (British, EU and Commonwealth citizens aged 18 or over, plus those that will attain age 18 during the year of the Roll's currency). Since 2000, UK residents have had the right to remove their records from of the public version of the Electoral Register, an option known as 'opt-out' which is now exercised by an estimated 30% of electors. To compensate for 'opt-outs', Experian and other private sector providers (such as CACI or 192.com) supplement the Register with other data sources, such as public registers (company directors and shareholders registers) as well as commercial surveys or third party customer data, in order to compile population databases. In the case of Experian (Nottingham, UK), this is now commercialised as a 'Consumer Dynamics' file that in 2004 contained 46,336,087 adults, a higher number than those in the full version of the Electoral Register (Sparks, 2005). Two versions of this dataset for the UK were kindly made available by Experian to University College London, one from 1998 including all surnames held by 100 people or more and their frequencies by postal area, and a second one for 2004 including all surnames and forenames at unit postcode level.

A different type of dataset used was the distribution of surnames in the 19th century in Great Britain (i.e. excluding Northern Ireland), derived from the individual responses to the 1881 Census. This file was kindly supplied by Kevin Schürer, Director of the ESRC UK Data Archive at the University of Essex, and contained counts of surnames by Parish in the 1881 Census (Schürer, 2004). This file was aggregated to today's Postal Areas in a previous project at University College London (Surname Profiler, 2006). In that project, this dataset made it possible to trace internal migration movements in the changing

geographic pattern of names over time, while in the research described here it has been used to screen out names that have arrived in the UK during the late 19th and 20th centuries.

An additional dataset used, that is not considered a 'name dataset' and therefore is not included on Table 3, is a Geodemographics neighbourhood classification system, Mosaic, provided by Experian, which classifies the UK's 1.6 million unit postcodes into 61 types according to the demographics of the immediate residential neighbourhood. The neighbourhood types were clustered using both UK Census 2001 small area statistics as well as other publicly available and commercial datasets (Harris et al, 2005). In this research, the Mosaic dataset has made it possible to match the areas of highest concentration of certain names and relate them to neighbourhood types with higher presence of particular ethnic groups, religions, socioeconomic types, or urban/rural populations, as will be described in section 3.3.

Besides the UK data, other less detailed population files for other countries or periods have been sourced from electoral registers or telephone directories from five other countries (Ireland, Australia, New Zealand, the US, and Canada) at different levels of spatial disaggregation. The full details and characteristics of these datasets are listed in Table 3.

Country/ Territory	Name of Dataset	Year	Nominal Resolution (finest record)	Spatial Resolution (smallest area)	Data Provider	Population included in Dataset	Total Population Enumerated	% of Country's Total Pop.	Country's Total Population	No. unique Surnames	Avg. People/ Surname
Great Britain	Electoral Register & Consumer Dynamics	2004	Individual person	Postcode Unit	Experian UK	Residents registered to vote of age >17 (opt-in) + consumer database	46,336,087	77.5%	59,800,000	218,392	212
Great Britain	Electoral Register	1998	Family name	Postal Area	Experian UK	Surnames >100 occurrences (age >17)	37,278,477	63%	59,200,000	25,730	1,449
Great Britain	Census of population	1881	Family name	Postal Area (equivalent)	ESRC UK Data Archive	All census respondants	28,225,211	81%	35,026,108	44,545	634
Northern Ireland	Electoral register	2003	Family name	Postal Area	Experian UK	Residents registered to vote of age >17	n/a	n/a	n/a	n/a	n/a
Ireland (Republic of)	Electoral register	2003	Family name	County	Experian UK	Residents registered to vote of age >17	2,912,541	73%	4,015,676	n/a	n/a
Australia	Electoral register	2002	Family name	Standard Statistical Division (SSD)	Experian / Pacific Micromarketing	Residents registered to vote	7,784,676	38%	20,264,082	12,266	635
New Zealand	Telephone Directory	2002	Family name	Province	Experian / Pacific Micromarketing	Telephone subscribers	934,686	23%	4,076,140	n/a	n/a
United States	Telephone directory	1997	Family name	State	Ken Tucker	Names with >100 occurrences in the tel.directory	81,000,000	30%	266,490,000	145,242	558
Canada	Telephone directory	1996	Family name	National	Ken Tucker	Names with >100 occurrences in the tel.directory	9,150,000	28%	33,098,932	33,355	274

Table 3: Sources of Data- Reference Population- used to build the CEL classification

3 Techniques

This section will set down the basic methodological framework for the classification of names into CELs, while section 4 will draw in the task of explaining how the classification of the most common surnames and forenames has been performed.

The task of classifying the 281,422 surnames and 114,169 forenames most commonly present in Britain in 1998 and 2004 into cultural ethnic and linguistic groups (CEL) is one that cannot be approached manually or following traditional etymological methods. The Dictionary of American Family Names (DAFN) includes the 70,000 most common surnames in the US and their etymological explanation, and comprises three bulky volumes of over 2,000 pages which took ten years and more than twelve experts to prepare (Hanks, 2003) even using as it did a semi-automated initial classification system to allocate groups of names to linguistic experts (Tucker, 2003). Therefore, given the number of names to be classified and the scarce resources available, a different type of approach was required for the current UK project.

A set of classification rules was first investigated and then applied to this purpose through a dynamic and iterative process. This section summarises the core set of these rules and their sequence of use in the decision process, which have been substantially synthesised here for the purpose of clarity and brevity. The actual detailed process was much more complex, in terms of number of exceptions and iterations, and cannot be fully described in the space available in this paper.

Prior to explaining the actual classification process finally applied to the list of names, a summary of the major techniques used will be offered here, with the objective of giving meaning to the concepts employed in the next section and justifying their final selection after a preliminary evaluation.

3.1 'CEL-triage' between forenames and surnames

The 'CEL-triage' technique was first introduced by Tucker (2005), and consists in identifying clusters of names grouped by high frequencies of cross-occurrences between forename and surname in individuals (e.g. forenames will be considered Chinese if a high

proportion of their bearers also have Chinese surnames and vice versa). This is the mechanism used in DAFN to sort 70,000 surnames into 23 onomastic groups for etymology specialists to analyse.

CEL-triage, can be either user-induced or automatic. In the first case, the user selects a 'forename *A'* at random where the CEL is previously known, such as 'Pablo' (Spanish CEL), which will act as a 'seed' for building a new CEL. The user then finds the most common 'surnames *B'* that Pablos bear, such as Mateos, Garcia, Perez, etc, and then all the 'forenames *C'* associated with those 'surnames *B'* (e.g. Juan, Rosa, Javier, Marta, etc.). By repeating this process from the start for the 'forenames *C'* and conducting further iterations, the same forenames and surnames tend to be highly clustered around those individuals belonging to a same CEL category. This iterative process is illustrated in Figure 1, where only 2 of these cycles are shown (A-B and C-D).

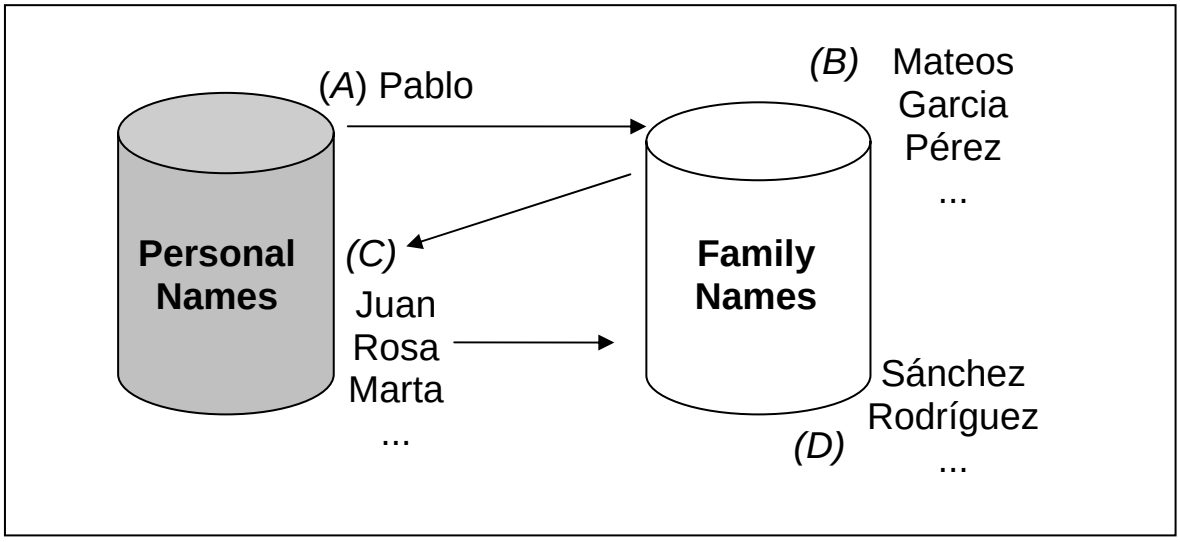


Figure 1: Diagram of the CEL-Triage technique, using the example of a cluster of Spanish Names

Therefore, after a few cycles one can find a set of several hundreds or even thousands of names belonging to a common CEL cluster, just by knowing one forename. In the automatic version of this technique, it is not even necessary to know the CEL of 'forename *A'*; the computer chooses a forename at random and automatically identifies clusters of common cross-occurrences through the same cycles described above. At the end of the automated process the user decides the most likely CEL of the whole cluster by looking up one or two names in a dictionary or through one of the other techniques described in this

section. Going back to the DAFN example, manual classification of an initial set of approximately 3,000 forenames into CELs allowed the authors to automatically assign a preliminary CEL to 85,000 forenames and over 100,000 surnames using this technique (Tucker, 2006).

This technique is perhaps the most useful method for the classification of large number of names into CELs, and it has indeed proved very reliable for classifying high frequency names. It works best with CEL groups that are distinctive, such as Japanese names. Amongst its limitations, are that it is less appropriate for names corresponding to well-established immigrant groups that are very integrated with the 'host population' (e.g. Jewish or Huguenot names in Britain), and for names with small frequencies.

3.2 Spatio-temporal analysis

A further technique is based on the analysis of spatio-temporal differences in the distribution of name frequencies and rates across the geographies and times available in the datasets. This implies the identification of significant differences in the total frequencies of names and/ or rates per million people between different areas within a country and between different countries or points in time. Once such significant differences are found, expert knowledge is required in the geography and history of the countries and their internal and international migration and patterns in order to justify the CEL categories associated with such migration or differential distributions. There are hundreds of different types of such patterns and just a few different types of examples will be mentioned here.

For example, in the UK those names which are proportionally more common in postal area 'NW' (North West London) than in any other of Britain's 120 postal areas include large numbers of Jewish names, whilst most Greek or Greek Cypriot names have postal area 'N' (North London) as their most common (see Figure 2). Likewise, if a name is more common pro rata in Wisconsin than in any other US state, this will support the contention that it is of German or Scandinavian origin, while other names that might appear to be Germanic yet are more common in New York than in any other US state are more likely to be Jewish than German. International comparisons are very useful, for example, most Chinese names are relatively common as a proportion of the total population in the US than in the UK, whereas most South Asian names are proportionally less common in the

US than in the UK. Moreover, names that are more common in Australia than in the US or the UK, and within Australia are even more common in New South Wales, are likely to be Vietnamese rather than Chinese.

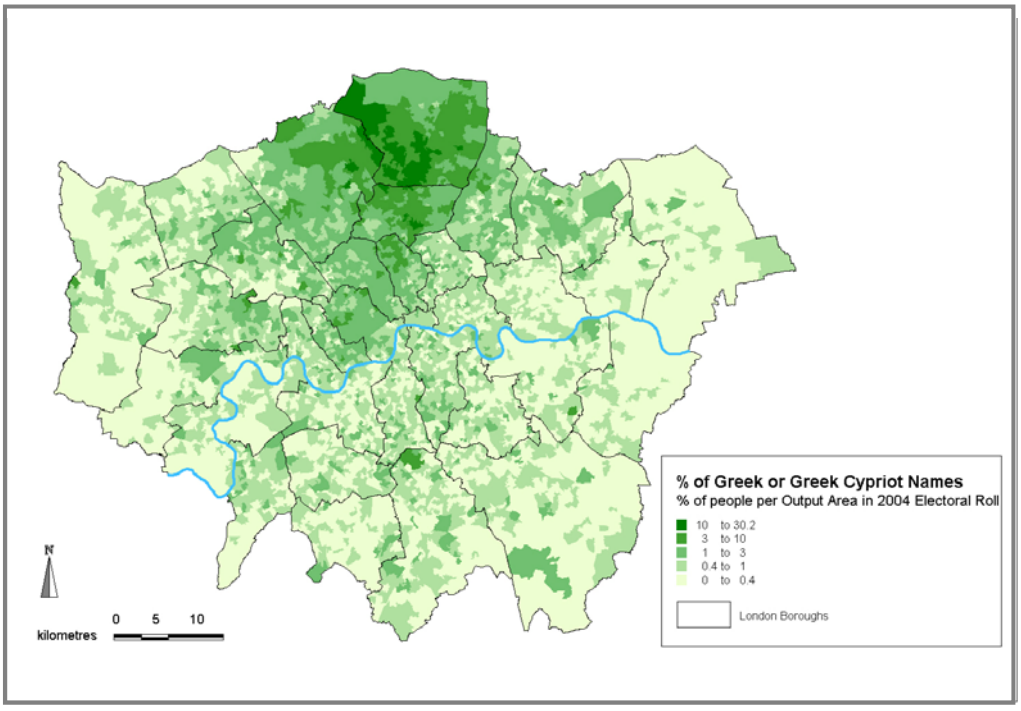


Figure 2: Map of distribution of Greek and Greek Cypriot Names in London by Ouput Area
The map shows percentages of people in each Output Area classified into 5 intervals

With regard to the temporal dimension of this type of analysis, names present today in Britain that did not appear in 1881 are likely to be of foreign origin (or of more recent invention). On the contrary, surnames of foreign origin that were present in Britain in 1881 are likely to have high numbers of British forenames today and therefore are unlikely to be properly picked up by the CEL-triage technique.

The spatio-temporal analysis technique has been especially useful to identify regional CEL groupings within a region or constituent country of the UK, such as Scotland, Wales, Northern Ireland, Cornwall, or the Channel Islands. It also very useful to identify the most frequent names in the ethnic minority groups that are highly concentrated in a few areas, such as South Asians in the UK. Its major limitations are that it is applicable only to names with frequencies of 100 or above, and that it requires detailed specialist knowledge of historic and current migrant settlement patterns by small area.

3.3 Geodemographic analysis

Geodemographics is defined as *'the study of population types and their dynamics as they vary by geographical area'* (Birkin and Clarke, 1998, 88). The geodemographic analysis used in this research entails identifying the socioeconomic types of neighbourhoods where a name is most commonly concentrated, and make inferences about the type of population living in them. The analysis of the UK 2004 Electoral Roll data using the geodemographic classification *Mosaic* proved useful in identifying non-British names which are highly concentrated in a few geodemographic types. However, a very similar result would have been achieved by just using UK Census data (Harris et al, 2005).

Some examples of geodemographic types in *Mosaic* representative of certain ethnic minorities are; Mosaic Type C20 'Asian Enterprise' has a particularly high proportion of residents classified by the census as South Asian and of Hindu or Sikh religion. D26 'South Asian Industry' is an example of a Mosaic Type with a very high proportion of South Asian Muslim residents. F36 'Metro Multiculture' by contrast is a Mosaic Type with a high concentration of more recent immigrant groups, only a small proportion of whom originate from South Asia. Other Mosaic types with higher proportions of minority ethnic groups than the national average are A01 'Global Connections', with Jewish and Armenian names, E28 'Counter Cultural Mix', D27 'Settled Minorities', which contains mostly Caribbeans, Greek Cypriots and Turks.

An extension to this geodemographic analysis of the UK is to analyse the percentage of people with a name that live in rural versus urban postcodes, or with Mosaic types of a 'high socioeconomic status'. This is based on the facts that most ethnic minorities are concentrated in urban areas (exceptions being some traditional groups in agri-business work such as Portuguese) and of that those certain groups that live in 'high status' postcodes tend to come to certain countries (such as Japan, Scandinavian countries, Saudi Arabia, etc.)

Geodemographic analysis is very useful for identifying certain non-British names that are typically very concentrated in a limited number of *Mosaic* types, particularly Jewish, South Asian and African names. However, this technique is less effective in assisting distinctions *within* those major non-British CELs (intra-South Asian or intra-African divisions) or less residentially concentrated non-British CELs.

3.4 Text mining

Text mining embraces a series of techniques that seek to capture similarities in the morphology of names, through text analysis, in order to relate them to a particular language of origin and thus a CEL. There are two basic techniques to find forms of commonalities between names; name stems and name endings on the one hand, and letter sequences and letter absences on the other.

The easiest way to group names by their stems is to sort them in alphabetical order, while to do so by their name endings the reverse of the name (i.e. the reverse of 'WEBBER' would be 'REBBEW') is first created and then sorted in alphabetical order. Once names are sorted by either their stems or endings, they are reviewed in order to isolate the main groups of common stems/endings (e.g. many names starting with 'ABD' are of Muslim origin and most with 'MAC' are Scottish or Irish, while most names ending in 'SKI' are Polish, 'SSON' are Swedish, 'OVA' are Russian or Czech, 'EZ' Spanish, or 'ULOS' or 'AKIS' Greek). The algorithm to process names in this way is as follows:

1. Sort all names in alphabetic order.
2. For each unclassified name do steps 3 to 8.
3. Look at the 10 previous and 10 subsequent neighbouring names in the list (20 neighbours).
4. Identify which CEL these neighbours are already assigned to.
5. Assign a weight to those neighbouring CELs by inverse distance to the target name; a weight of 1 (farther) to 10 (closer) in each direction from the name (a).
6. For each CEL present in the 20 neighbouring names, sum up their weights ($\sum a$)
7. Re-sort all the names using the reverse of the name (termed *anti-alphabetic order*)
8. Repeat the process once from 3 to 8 ($\sum b$)
9. Create a total score per name and CEL as follows:

$$S = \frac{\sum a + \sum b}{220} \cdot 100$$

Where S is the score, a and b are the result of step 5 for the alphabetic and anti-alphabetic rounds, and the denominator (255) is the maximum possible total score for the 10 neighbouring names (10!=55) in the 2 directions (55 x 2=

110) and 2 rounds (alphabetical and anti-alphabetical order, i.e. 110+110= 220)

10. Rank the unclassified names by the total score.
11. Select those name–CEL combinations with a total score of at least 40% and then allocate the CEL to the name.

An alternative method is to extract the first and last 2, 3, and 4 letters of a name, aggregate them and calculate their frequency in the name dataset, what makes it possible to locate the most common name stems and endings in a list of names. The CEL of each particular name stem or ending is then decided by using one of the other techniques (i.e. non-text mining), and is then applied to all names with such forms that remained unclassified after using the previous techniques.

The advantage of the sorting of names versus taking a discrete number of ending or stem letters, is that the former makes it possible to find in a single step patterns of names with a common origin even when they share 2, 3, 4 or more letters, while the latter might miss names that are not so obviously related (e.g. Basque names ending in 'BERRI' 'BARRI' or 'URI' are sorted together in their reverse form).

The second form of commonalities between names of the same CEL is letter sequences and letter absence. For example, for Spanish names linguists and statisticians have found that they never contain the letters 'K' or 'W' (Buechley, 1967), and the only double letters present are 'RR' and 'LL' (Word and Perkins, 1996). It is also known that the double letter 'AA' is the transcription into English of the Nordic letter 'Å', and therefore many names starting or containing 'AA' are likely to have originated in this region. However, this technique requires the development of large repertoire of letter sequences and absences and hence a good knowledge of each CEL language.

Text mining is useful for identifying non-British names which have been assimilated by the host community, for example those which, on the basis of CEL-triage analysis, appear to be British but which are for example of Scandinavian origin. It is also a useful strategy for classifying large numbers of low frequency names, such as Spanish or Italian names in the UK, and this reduces the number of names that would otherwise remain 'unclassified'. It is also useful to find different name variants that might have originated from the same

name (e.g. Mohamed, Mohammed, Muhammad, Mohammad). The main disadvantage of this technique is that it is not sufficiently reliable to override the results of the other methods, since there are exceptions to the text pattern rules (e.g. O'Brian could be misclassified as Armenian using text mining because of its ending in 'IAN', but is in reality of Irish origin). In other words, as mentioned, text mining is best used in classifying names not covered by other methods.

3.5 Name to Ethnicity data

This is the method followed by most of the name studies in epidemiology, as reviewed by Mateos (2007a). It is based on using population registers where the ethnicity (or a proxy) and the name of the person is already known in order to build appropriate name-to-ethnicity reference lists. Only two of the studies reviewed by Mateos (2007a) had access to a large enough population register to produce a reference list with a significant amount of unique surnames, in this case of over 20,000 surnames each (Lauderdale and Kestenbaum, 2000, Word and Perkins, 1996). These two studies satisfied the criterion established by Cook *et al* (1972) for minimum reference population size. Even then, these two major studies only aim to classify 7 ethnic groups.

The aim of this research is to classify over 250,000 surnames into CELs, and to apply the Cook *et al* (1972) criterion for reference population size, suggesting a minimum reference population of 3.35 million people together with their ethnicity. They also estimate a more robust size of 13.4 million people. As previously mentioned, it proved impossible to access a register of names in the UK which would include a large sample of the population surnames together with their ethnicities in sufficient numbers. Therefore, partial lists of names have been used for which one of the CEL dimensions was known, such as country of birth, or nationality.

One of the lists consulted is a list of surnames and forenames by nationality in Catalonia, Spain (IDESCAT, 2006). Furthermore, aggregated data of most common names by country of birth (COB) were obtained from patient registers from Camden Primary Care Trust in London, including names by COB with 4 or more occurrences (a minimum threshold for common names applied in order to preserve confidentiality).

This technique proved useful for 'seeding' new CELs out of pre-existing broader groups (specifically, of breaking Eastern European names down into Russian or Polish names), especially for 'rare' CELs, and thus needs to be used in combination with the CEL-triage technique. Independently sourced Name-to-ethnicity data is also very useful to validate the name classification.

However, this technique does also present its limitations, the major one being the limited number of names for which ethnicity or place of birth is known. Even where these lists exist, two problems might be encountered; the availability of only a small number of ethnic groups (e.g. the 16 UK Census categories), and the differential distribution of names between; a) periods of time; b) receiving countries of immigration; and c) regions within those countries, which all might introduce biases in the name to CEL attribution. This is explained in Mateos (2007a).

3.6 Lists of international name frequencies and genealogy resources

This technique complements the others and essentially consists of accessing new and more anecdotal sources of name frequency data upon which all of the previous techniques are based. It entails collating lists of names, and preferably their frequencies throughout as many countries as possible. This has only been possible because of a significant surge of interest in genealogy and family history through the Internet in the last few years. According to The Guardian, genealogy is now second only to pornography in generating internet traffic (The Guardian, 2004), and following this public thirst for information about ancestry, a range of data providers are willing to publish name data on the web, ranging from formal institutions such as national statistical offices, to amateur genealogist blogs. A previous project at University College London to show historic and current surname distributions in the UK has also leveraged this interest with over 3,000 daily visitors at www.spatial-literacy.org.

Amongst these types of lists, several resources are available from the official statistics offices or public registers in some countries, for example, lists of forenames and/or surnames and their frequencies in Belgium (Statbel, 2006), Denmark (Danmarks Statistik, 2006), Iceland (Statistics Iceland, 2006), Madrid and Catalonia in Spain (IDESCAT, 2006, Instituto de Estadística de la Comunidad de Madrid, 2006) and Germany (Gesellschaft für

deutsche Sprache, 2006). These lists have been used to re-apply the techniques previously mentioned, so that broader CELs such as Scandinavian, Central Europe, or Hispanic, could be broken down into finer CELs, for example making it possible to distinguish between Wallonian and Flemish names in Belgium, or Catalan and Castilian in Spain.

Other lists of names and their language of origin (with no frequencies) are available on the web and have allowed the classification of names from countries as diverse as Ghana, Romania and Albania. Furthermore, frequency data can be computed from electronic telephone directories from different countries, where available, in order to compile international comparison lists.

This method is especially useful for classifying names from CELs where the names overlap, (e.g. Albania, Croatia, Serbia) and to search for new CELs and seed them in the CEL-triage technique to split up broader groups. However, CEL-triage works best for high frequency names, and not so well for lower frequency ones, since a few individual surname-forename pairs can introduce a strong bias in the classification. Amongst other limitations is that the number of names available on the web is relatively small compared to an Electoral Register or a telephone directory, and their quality in linking names to a true language of origin very varied, with names 'claimed' as own from different countries or languages, so some further research and arbitration is necessary. This is where having name frequency with some geographical disaggregation is very useful (e.g. using telephone directories)

3.7 Researching individual names

As a last resort, when names cannot be classified using any of the methods presented above the last resource is to search for a particular name either in a name dictionary, or in a web search engine, such as Google, or in electronic telephone directories, to find particular associations between a name and a country or language through the contextual information in which they are found on the web. A similar technique to link geographic information found in miscellaneous web content, termed 'heuristics for geo-referencing web pages', has been developed by Silva *et al* (2006) to perform such associations automatically. The obvious limitations of this method are that it is time consuming, dictionaries are only available for a few names / countries, and different CELs have a very different presence on

the Internet (e.g. African names are misrepresented in the web), or duplicate and competing CELs are presented for some of the same names. However, this method has proved very useful to seed new CELs into the CEL-triage technique, where a forename or surname has a high proportion of corresponding names that are not classified – as, for example, to identify Fijian and Lao names in Australia or Ethiopian names in the UK.

4 Building the CEL Name classification

4.1 Stages in the creation of the classification

As previously discussed, the end objective of this research is to classify every surname and forename present in Britain in 2004 that has a frequency of 3 or more people. This means building and classifying a reference list of 281,422 surnames and 114,169 forenames into cultural ethnic and linguistic groups (CELs).

An ad-hoc methodological approach to classifying such name lists has been taken in the research reported here, following a series of empirical steps that were developed as the project evolved. This in practice means that the authors did not start with any pre-conceived notions of the optimal methods to classify all of the most frequent names according to their origins, and neither were all of the datasets available from the outset. Therefore a series of exploratory rules were tested and applied in a sequential process guided essentially by pragmatic considerations, not necessarily in the most logical order, the results of which were only evaluated at the end with reference to aggregate Census of Population data. This essentially ad-hoc approach shaped the way in which the final CEL Name classification was built, the techniques that were employed, and possibly the results that were obtained.

In order to build a name reference list, a series of data sources concerning several reference populations were used. These were initially described in section 2.3 and summarised in Table 3 as nine separate datasets (Great Britain- GB 2004, GB 1998, GB 1881, Northern Ireland, Republic of Ireland, United States, Canada; Australia, New Zealand). However, one of these datasets could not be sourced at the beginning of this research project; the GB Electoral Roll & Consumer Dynamics file for 2004. This dataset, which hereinafter will be called ‘GB04’, is the most detailed of the datasets in terms of the number of surnames and forenames that it contains and its resolution at the individual level. It only became

available at a late stage in the project. The other eight datasets were all available from the start, but only included surname data (no forenames) and the records were aggregations of individuals to some coarse level of geography (i.e. no individuals or neighbourhoods).

This non-availability of some data sources had implications for the research design, and the way in which the resulting name classification was built. If all of the datasets had been available from the start, and the project had set the ambitious objective of classifying such a large number of surnames and forenames at the outset, other methodological paths would have been followed. As it was, the final classification of 281,422 surnames and 114,169 forenames classified into cultural ethnic and linguistic groups (CELs) was built in three stages, and was made up of three distinct 'tiers' of names used as inputs into the classification. The characteristics of these three stages are summarised in the following paragraphs, and they are explained in detail in Sections 4.2 to 4.4.

4.1.1 Stage 1 and Tier 1 Names

The first dataset to be sourced was the GB Electoral Roll for 1998, consisting of 25,630 surnames with a frequency of 100 people or more nationally, together with their frequency by postal area, a geographical division defined by the first two digits of the unit postcode (e.g. 'BR' for Bristol), of which there are 120 in Great Britain, with an average population of 500,000 people. This dataset will be referred to in this paper as the 'GB98 dataset'. The aim of the first phase of the project was to classify these 25,630 unique surnames into British regions and major ethnic minority groups of origin. This table of most frequent surnames will be hereinafter termed 'Tier 1' names. Moreover, the other seven datasets that were obtained at approximately the same time as GB98 were used primarily to support the classification of the surnames in 'Tier 1', as will be detailed in section 4.2.2.

4.1.2 Stage 2 and Tier 2 Names

At the beginning of stage 2 the 'GB04' file was received from *Experian* (Electoral Roll & Consumer Dynamics for 2004), which included the forename, surname and unit postcode for each of the 46.3 million electors/residents. At this stage only, it was considered appropriate to add to the existing CEL Name classification the capability to classify forenames as well as surnames. Therefore, a list of the most common forenames, defined as those with a frequency of at least 9 people, was extracted from 'GB04' what produced a

file with 29,979 forenames, hereinafter termed ‘Tier 2’ names. The methodology to link Tier 2 to Tier 1 and classify it into CELs is detailed in section 4.3.

4.1.3 Stage3 and Tier 3 Names

At this point the CEL classification system is comprised of two files; Tier 1 with 25,630 surnames, and Tier 2 with 29,979 forenames, each of them assigned to a CEL Group and Type. As mentioned before, the names in these two files respectively cover 37.2 and 45.3 million residents in the UK 2004 file. At that moment, a decision was taken to expand the classification in order to classify the entire population of the Electoral Roll (46.3 million people) into ethnic groups, so that most ethnic minorities could be correctly covered by the CEL classification. ‘Tier 3’ names are thus comprised of all the names with 3 or more occurrences in the ‘GB04’ dataset and that are not included in either ‘Tier 1’ or ‘Tier 2’ files, comprising a total of 255,792 surnames and 84,192 forenames.

4.1.4 Classification of Tiers 1, 2 and 3

As a result, Tier 1 names contains the top 25,630 surnames, Tier 2 names the top 29,979 forenames, and Tier 3 names the rest of both surnames and forenames. The three Tiers combined comprise a total of 281,422 surnames and 114,169 forenames. The seven classification techniques described in section 3 were applied to each of the three Tiers of Names described in the previous section 4.1.1 to 4.1.3, according to a set of rules which will be summarised in the following three sections.

4.2 Tier 1 Names: Top Surnames

4.2.1 Data preparation

The names in the Tier 1 file were initially processed to eliminate data errors, such as inconsistent geographic indicators, invalid entries (e.g. ‘N/K’), and to standardise the format by trimming spaces, and unifying different dashes, apostrophes and other special characters used. These data cleansing steps were required to make sure that there was a common entry for each unique name across the seven datasets that were to be compared (see Table 3 all but GB2004). Other known errors in the data, such as the presence of name initials or honorifics (e.g. Prof., Ms., Dr., Sir), were kept in a separate field since they were judged to be able to provide some valuable information for later classification tasks.

Item	Attribute	Description	Variable Type	Geography
• *Items **a** to **j** are repeated for each surname and geography (a total of 9 sub-datasets: GB, NI, IE, US, CA, AU, NZ, All Ireland, British Isles)*				
a	Name Frequency	Number of occurrences	Integer	Countrywide
b	Name Rate	Rate of occurrences per million people,	Double	Countrywide
c	Top Area	Area with highest rate of occurrences per million people (e.g. Postal Area or State)	String	Countrywide
d	Top Area Rate	Rate of Occurrences per million people in Top Area **c**	Double	Top Area
e	2nd Top Area	Area with second highest rate of occurrences per million people	String	Countrywide
f	2nd Top Area Rate	Rate of Occurrences per million people in Next Top Area ***e***	Double	Next Top Area
g	Finer Top Area	Finer Area with highest rate of occurrences per million people (e.g. Postal District)	String	Countrywide (GB & AU only)
h	Finer Top Area Rate	Rate of Occurrences per million people in Finer Top Area ***c***	Double	Finer Top Area (GB & AU only)
i	Difference between Country Rates	Ratio of ***b*** between pairs of countries	Double	Selected Pairs of countries
j	Difference between Top Area Rates	Ratio of ***d*** between pairs of countries	Double	Selected Pairs of countries
*Items **k** to **s** include a unique value per surname in Tier 1*				
k	Temporal Change 1994/1881	Ratio of ***a*** between GB 1998 and GB 1881	Double	GB
l	Entries UK Gazetteer	Number of placename entries in the UK gazetteer	Integer	UK
m	Top UK Gaz. Area	UK County with highest number of entries in gazetteer	String	UK
n	Entries African/ Asian Gazetteer	Number of placename entries in the African or Asian gazetteer	Integer	Africa and Asia
o	Top African/ Asian Gaz. Area	African or Asian region / country with highest number of entries in gazetteer	String	Africa and Asia
p	Top *Mosaic* Type	Socioeconomic type of neighbourhood with highest rate of occurrences per million people	String	GB
q	Top *Mosaic* Rate	Rate of Occurrences per million people in Top *Mosaic* Type ***p***	Double	GB
r	Rurality	Percentage of names in rural postcodes	Double (%)	GB
s	High Status	Percentage of names in 'high status' postcodes (as defined by *Mosaic* Types)	Double (%)	GB

Table 4: List of attributes associated with each name in 'Tier 1'
Country codes used: UK= United Kingdom; GB = Great Britain (ex. NI); NI= Northern Ireland; IE= Republic of Ireland; All Ireland = IE+NI; British Isles= IE+UK; US= United States, CA= Canada; AU= Australia; NZ= New Zealand

The surname frequencies per postal area were converted to rates of number of names per 1 million people, in order to be able to compare them in a consistent way across all the geographies studied. Two additional geographies were created for the purposes of calculating additional name frequencies and rates, through the aggregation of some countries into bigger regions; *All Ireland*, including the Republic of Ireland and Northern Ireland, and *British Isles*, including the former plus Great Britain. The datasets at this stage included a total of 9 'countries or territories'; Great Britain- GB 2004, GB 1998, GB 1881, Northern Ireland, Republic of Ireland, United States, Canada; Australia, New Zealand, *All Ireland* and *British Isles*. These were linked together through the common surnames between them. For each surname in Tier 1, and for each of these 9 geographies, a set of 10 different statistical and geographic variables were calculated, the details of which are offered in Table 4 (items ***a*** to ***j***).

After these two steps of data cleansing and name frequencies and rates calculation, each name was linked to both a UK and a Worldwide place name gazetteer (Edina, 2006, National Geospatial Agency, 2006). This made it possible to evaluate if there were any gazetteer entries for each name, and if so count them by region of occurrence. The results of this search were stored for the UK gazetteer and for entries in Africa or Asia in the Worldwide gazetteer (items ***l*** to ***o*** in Table 4), since America, Europe and Oceania had substantial numbers of names in countries not corresponding to their apparent region of language origin (i.e. English or Spanish names).

Finally, an additional dataset was used, *Mosaic* neighbourhood classification which was previously described. A separate table was provided by Experian with the distribution per *Mosaic* Type for each surname. Using this table and for each surname in Tier 1, the *Mosaic* Type with the highest rate of names per million people was selected (item ***p*** in Table 4). This made it possible to calculate the percentage of names in rural versus urban areas, as well as in high socioeconomic status neighbourhoods (items ***r*** and ***s*** in Table 4)

As a result, Tier 1 names table included for any name, between 8 and 19 attributes for each of the 9 geographies studied, resulting in over 100 attribute combinations for some of the most common names present across all the geographies. These attributes are fully described in Table 4.

4.2.2 Classification rules applied to Tier 1 names

As stated, the original aim of this first phase was to classify the 25,630 surnames in Tier 1 into British regions and major ethnic minority groups, and therefore the techniques applied were spatio-temporal analysis and geodemographic analysis, combined with text mining. For the purpose of the CEL classification this phase consisted in separating British and non-British names, and to subdivide these two groups into finer categories. In this research the concept of 'British names' include all names that have either originated in the British Isles (comprised of the current UK and Ireland), or were introduced a long time ago as to be considered fully integrated into the British and Irish society. Such temporal distinction has been totally arbitrarily set to names that arrived in the British Isles before 1700, and thus before the Industrial Revolution. Defined as such, British names, were further subdivided into the following 7 CELs; English, Irish, Northern Ireland, Welsh, Scottish, Cornish and Channel Islands, with the addition of other CELs (Norman Huguenot, and Jewish names) that are sometimes considered together with the British CELs for the CEL-triage technique and other calculations, because of their high integration with British names. Such inclusion is made explicit in this document when these CELs are added to the British ones.

non-British names were originally divided into the following 12 major CEL groups present in the UK, which were relatively easy to isolate; African, Scandinavian, Greek or Cypriot, Jewish and Armenian, Hispanic, Rest of Europe, East Asian, Japanese, Muslim, Sikh, South Asian, and 'non-British Unclassified'. These 12 *CEL groups* were further subdivided into much finer CELs denominated *CEL types* (e.g. the Hispanic CEL group into Spanish, Portuguese, Catalan, Basque, and Galician CEL types).

Offering a detailed account of all the variables and thresholds considered in each decision to assign a CEL to a name is much beyond the scope of this paper. A summary of the main eight rules and decisions taken is offered here, what will help to understand how the classification was created and the decisions and assumptions taken. These eight rules have been numbered A.1 to A.8. A chart with the decision tree of these rules is offered in Figure 3 and should be used to accompany the text.

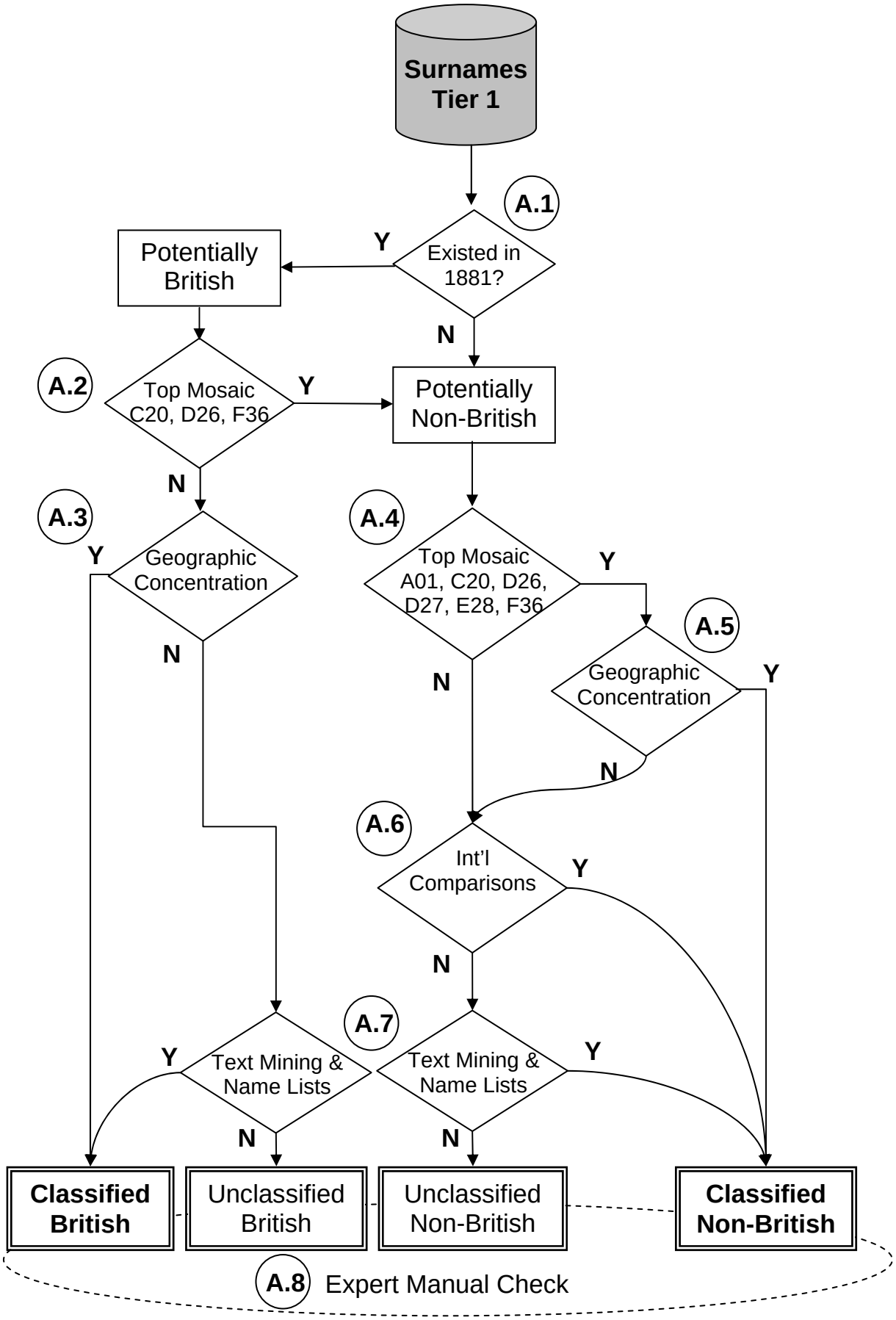


Figure 3: Classification Decision Tree for surnames Tier 1
The code in a circle relates to the reference of each rule in Tier 1, A.1 to A.8 which are described in the text.

4.2.2.1 Rule A.1

In order to split between British and non-British names, the main rule applied was to check whether a surname was present in the 1881 Census (potentially British) or not (potentially non-British, of which there were 1,674 surnames). Even if the surname was present in 1881, when the increase in its rate per million names between 1881 and 1996 was likely to have been high (over 100% on average) it was considered potentially non-British, adding over 1,000 more surnames (mainly European) to this list. As a result, 22,956 names were classified as British and 2,674 as non-British out of a total of 25,630 surnames.

4.2.2.2 Rule A.2

Geodemographic analysis was used to confirm a surname as British when the Top *Mosaic* Type was not C20 'Asian Enterprise', D26 'South Asian Industry', or F36 'Metro Multiculture', since these neighbourhoods present a high number of non-British population in the 2001 Census. Geodemographic analysis was used for reasons of convenience, since the data supplied by Experian was coded by *Mosaic*, but the 2001 Census raw scores could have been used as more direct indicators of ethnicity.

4.2.2.3 Rule A.3

Following the Spatiotemporal Analysis techniques, a British surname was assigned to a sub-British CEL type according to the region where the name was most concentrated both in 1881 and 1996 (Top Area and Next Top Area rates).

4.2.2.4 Rule A.4

The Top *Mosaic* Type of a potential non-British surname was used to identify specific CELs (letter ***p*** in Table 3). Those with C20 'Asian Enterprise' were pre-assigned the CEL 'South Asian', since this type has a particularly high proportion of residents classified by the census as South Asian and of Hindu or Sikh religion. Those surnames with D26 'South Asian Industry' to the CEL 'Muslim', since the majority of residents are of South Asian ethnicity and Muslim religion. Finally, *Mosaic* type F36 'Metro Multiculture' representing more recent immigrant groups, specially Black Africans, was provisionally assigned to the CEL 'African' when Top Areas were in South London, since these present a high concentration of this CEL. Other *Mosaic* types with high proportions of minority ethnic groups are A01 'Global Connections', with Jewish names, E28 'Counter Cultural Mix', and D27 'Settled Minorities', which contains mostly Caribbeans, Greek Cypriots and

Turks. However, to avoid mistakes in the assignment due to the ecological fallacy, this rule was best used in combination of rules A5, A6 and A7.

4.2.2.5 Rule A.5

Through spatial analysis of the top areas of potentially non-British surnames, groups of postal areas with much higher concentrations of non-British names were easily identified. Local knowledge of the postal geography of the UK, and specifically of London, allowed these groups of surnames to be provisionally assigned into CELs, for example Greek or Turkish names in postal area 'N', Jewish in 'NW', or South Asian in 'LE', but only if the top *Mosaic* type of a surname was also deemed to correspond to the Census Ethnicity and socioeconomic characteristics, as well as index of rurality, representative of a CEL (i.e. Jewish names are highly urban and more affluent than average).

4.2.2.6 Rule A.6

International comparisons of the relative frequency of a name (rate per million names) allowed to assign CELs to names with lower rates in the UK than in the US, Australia, Canada or New Zealand, based on the geographical distribution on those countries. For example, East Asian names are much more common in Australia than in Britain, while in the US are those of Scandinavian, East Asian, Jewish, or Hispanic origin, the last one much more common in the southern states than in the rest of the country.

4.2.2.7 Rule A.7

The surnames pre-assigned to a CEL through rules 1-6 were processed using text mining techniques to find particular patterns in their name stems and endings or letter sequence (for example identifying Scandinavian surnames ending in '-strom' or South Asian ones ending in '-dhu'). These techniques were also used to subdivide the 12 CEL groups into much finer CEL types, by finding common endings particular of a CEL type (such as Spanish '-EZ' or Greek Cypriot '-IDES'). Those same patterns were applied to the remaining unclassified surnames in both British and non-British groups to be able to allocate more surnames with a CEL. Finally, if an unclassified name had a concentration of entries in a placename gazetteer, this was used to assign the name to particular CEL in the world.

4.2.2.8 Rule A.8

Finally, all the 25,630 surnames, each assigned to a CEL, where distributed amongst university students and friends with an expert knowledge in each of the British CELs and the non-British CEL groups, for them to check any classification mistakes and to attempt further subdivisions of the broad non-British CEL groups into finer CEL types, according to linguistic, religious and geographic criteria. This process of giving each expert a pre-classified list of circa 1,000 surnames proved to be much more efficient than having attempted to give each of these experts the whole list of 25,630 unclassified surnames, avoiding the problem of substantial overlap between experts, and misclassifications due to fatigue and other human errors.

4.2.2.9 Outcome

At the end of this process of eight rules, of the 25,630 surnames in 'Tier 1', 2,978 surnames (11.6%) were classified as 'non-British' and assigned to 12 CEL groups and more than 50 CEL types, while the rest of the surnames (88.4%), were allocated to the 7 British CEL types or the 'unclassified' category. This Tier 1 file classified into CELs, covers a total population of 37,250,875 electors and residents in 2004, of which 3,834,722 of them have non-British surnames.

4.3 Tier 2 names: Top Forenames

4.3.1 Data preparation

At this stage access to the GB 2004 file (Electoral Roll & Consumer Dynamics) from the company *Experian* was obtained, a dataset which includes the forename, surname and unit postcode for each of the 46.3 million electors/residents. 'Tier 2' file was produced by aggregating the forenames in the 'GB04' file, and selecting those with at least 9 occurrences, what produced a file with 29,979 forenames.

However, since no other forenames dataset was available for any other geography, a new strategy was required to provide sufficient variables to classify these forenames. Using Tucker's (2005) CEL-triage technique, as described in section 3.1, the most efficient option was to use the existing CEL Group assigned to surnames in Tier 1, hereinafter called SCEL Group (for Surname CEL Group). For each forename in Tier 2, the proportion

of its bearers by SCEL was calculated. This is achieved by aggregating the individuals in the GB04 dataset by their forename, and then counting how many people there are for each SCEL Group according to their individual surnames and its corresponding SCEL Group in Tier 1. Finally the percentage of people per SCEL Group associated with each forename was calculated. In other words, the Tier 2 file contained a record for each of the 29,979 forenames, including a count and percentage of people with that forename whose surname is associated with an SCEL Group in Tier 1 file. For example, the entry for the forename *Pedro* in Tier 2 would be as follows (numbers are surnames counts with their percentage in brackets):

> *Forename:* ***Pedro*** *Total Frequency: 3,435*
>
> *SCEL Groups: British 245 (7.1%), Hispanic 2,410 (70.2%), European 35 (1.0%), 'None' 745 (21.7%)*

The SCEL Group 'None' represents the percentage of surnames associated with that forename that are not found in Tier 1 file. At this stage, there is a small number of forenames with a high proportion of SCEL Group 'None' since only a few of their surnames are in Tier 1 file, although most of forenames in Tier 2 file have surnames well represented in Tier 1.

4.3.2 Classification rules applied to Tier 2 names

As becomes obvious by now, the main method used to classify Tier 2 forenames into CELs is the CEL-Triage method, which was applied in a series of steps in combination with some of the other classification techniques described in section 3. The decision tree of the eleven rules applied to classify the forenames in Tier 2 file is illustrated in Figure 4, and the explanation of each rule is offered in the next paragraphs. As a result of this process, 29,979 forenames were classified into CELs, which hereinafter will be termed FCELs (for Forename CELs).

4.3.2.1 Rule B.1 – Unclassified forenames

Forenames with a frequency lower than 25 and a proportion of 80% or higher of their surnames not found in Tier 1 file (i.e. SCEL 'None') are classified as FCEL 'Unclassified_Rare'. Subsequent rules only apply to forenames not of the FCEL Type 'Unclassified_Rare'.

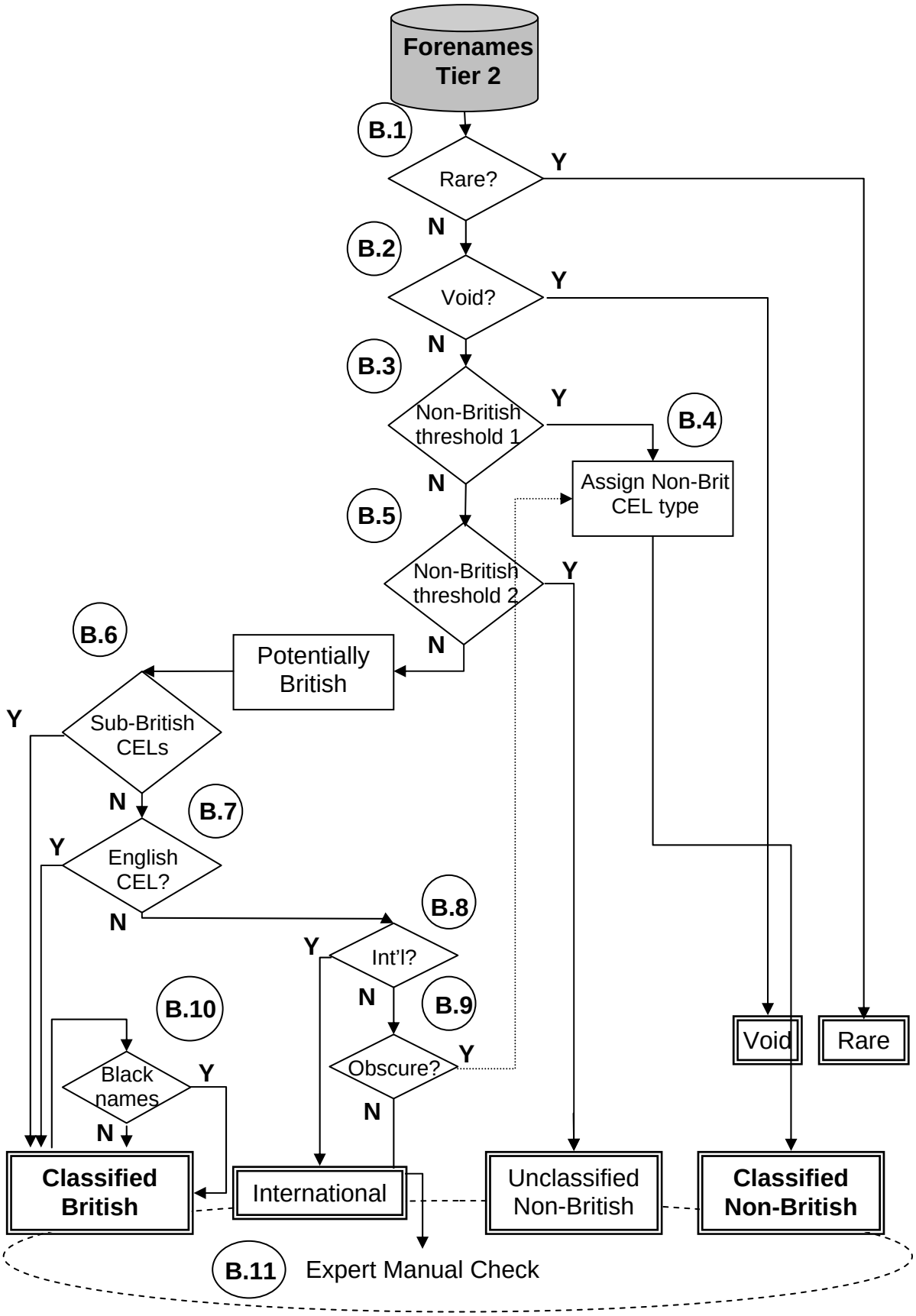


Figure 4: Classification Decision Tree for forenames Tier 2
The code in a circle relates to the reference of each rule in Tier 1, B.1 to B.11 which are described in the text.

4.3.2.2 Rule B.2 – Void forenames

Forenames with entries that do not seem proper names are identified and coded as FCEL 'Void'. Examples of these types of entries are records where no entry of any sort is found, (182,457 people), titles or honorifics such as Mr (17,692 people), Mrs, Count, Countess etc.. single letters which appear to be initials such as A, B, C etc. (2.76 million people); or two letter combinations with no vowel which also appear to be initials (eg JK, DL etc). Note that some two digit combinations with a vowel, such as 'Ho' (as in Ho Chi Minh) and 'Al' are valid forenames. These valid two character forenames are often identifiable from their greater frequency of occurrences than other two digit combinations, and because they might have a high number of unclassified surnames.

Subsequent rules (B.3 to B.10) only apply to forenames which are not of the FCEL 'Unclassified' or to the FCEL 'Void'. However, it has been appreciated that both Unclassified and Void entries are disproportionately more common in *Mosaic* types containing above average proportions of ethnic minorities, a fact that would require further analysis to reveal specific naming and recording practices in these minorities.

4.3.2.3 Rule B.3 – non-British or Jewish FCEL Groups

For all forenames which have at least 2 occurrences in one of the non-British SCEL Groups (excluding Jewish), and if so this frequency is 5% or higher of the occurrences in the combined British and Jewish SCEL Groups, the FCEL Group is made equal to that of the SCEL Group. If more than one SCEL Group meets this threshold criterion the FCEL Group is assigned to the SCEL Group with the largest number of occurrences. If two SCEL Groups have an equal number of forenames occurrences, the FCEL Group is then assigned to the SCEL Group with the smaller or smallest total number of occurrences on the entire electoral roll file. In all of the above situations the SCEL Group 'None' is not taken into account. For example:

Forename: ***Ourania*** *Total Frequency: 88*

SCEL Groups: British 10 (11.4%), 'Greek or Greek Cypriot' 24 (27.3%), 'None' 52 (61.4%)

Given that the 'Greek or Greek Cypriot' frequency (24) is more than 5% of the number with a British of Jewish SCEL (10), the name 'Ourania' qualifies as belonging to FCEL Group 'Greek or Greek Cypriot'.

4.3.2.4 Rule B.4 – non-British FCEL Types

Where on the basis of rule B.3 a forename has been assigned to a non-British FCEL Group, a further calculation was done to identify which SCEL Type, within the FCEL Group assigned in Rule B.3, had the highest number of occurrences of that forename, and then assigned it to the corresponding FCEL Type.

Taking the example of the forename 'Ourania', the following calculation is done:

Forename: ***Ourania*** *Total Frequency: 88*

FCEL Group assigned in Rule B.3: 'Greek or Greek Cypriot' 24

SCEL types: 'Greek' 15 (62.5%), 'Greek Cypriot' 9 (37.5%)

After which the forename is finally assigned to the Greek FCEL Type.

4.3.2.5 Rule B.5 – non-British Unclassified

Where, on the basis of rules B.3 and B.4, a P_ Name has not been assigned to a non-British FCEL Type, the following rule is applied to distinguish non-British from British forenames; if the percentage of occurrences of the SCEL Group 'non-British Unclassified' for that forename is equal or greater than 50%, then both FCEL Group and Type are assigned to 'non-British Unclassified'. This is illustrated through the following example:

Forename: ***Joris*** *Total Frequency: 35*

SCEL Groups: British 2 (5.7%), 'non-British Unclassified' 28 (80.0%), 'None' 5 (14.3%)

In all likelihood names such as 'Joris' that meet this criteria are surnames from a CEL which is not included in the list of CELs used in the analysis. In the case of 'Joris' this name, according to the Oxford Dictionary of First Names, originated from Frisia. However, at this stage it is decided to leave these rather obscure non-British forenames in the 'non-British Unclassified' FCEL for further work or their surnames in Tier 3.

4.3.2.6 Rule B.6 - British and Jewish FCEL Types

Where on the basis of rules B.3, B.4, and B.5, a forename has not been assigned to a non-British or Jewish FCEL Type, then it is assumed that the forename is likely to be of British or Jewish origin. [1]

[1] The reason why Jewish CEL Type is treated together with British CEL types is because although Jewish S_Names are quite distinct, they carry a high proportion of British F_Names due to a long history of integration into British society, compared to the rest of the non-British CELs (bearing in mind that, as stated before, the 'British' CEL include all names originated in the British Isles or imported up to around 1700).

To determine which British or Jewish FCEL Type it should be assigned to, a series of calculations are involved. For the purpose of this rule, only British and Jewish CELs are considered and all other non-British CELs are ignored. Firstly, the proportion of the occurrences of each forename in each British or Jewish SCEL Type is calculated (variable '*a*'). Secondly, the same calculation is repeated as a summary of all forenames in Tier 2 file, giving the overall GB average proportions by SCEL Type (variable '*b*'). Thirdly, the proportions of each particular forename in each British or Jewish SCEL Type are divided by the overall average for all forenames (ratio '*c'* =variable '*a*' divided by '*b*'). Finally, in those cases where a forename has a proportion of occurrences in a British or Jewish SCEL Type equal or higher than twice the national average (*a/b =>2*), the SCEL type with the highest value of *ratio 'c'* is assigned to that forename's FCEL Type. This is illustrated in the following example:

Forename: ***Lorcan*** *Frequency- Total: 90;, British & Jewish:* ***83***

	English	*Welsh*	*Scottish*	***Irish***	*Jewish*
a) Lorcan SCEL Types	30%	8%	18%	41%	2.4%
b) GB Average SCEL Types	69.4%	11.4%	10.3%	6.9%	2%
c) *Ratio c=a/b*	0.43	0.74	1.75	**5.94**	1.20

Those forenames which do not reach a *ratio c* of a least 2, are most likely English, since the average of 69.4% SCEL Types prevents that they meet rule B.6, and thus are dealt with in the next rule B.7. Therefore B.6 identifies Irish, Scottish, Welsh and Jewish FCEL Types.

4.3.2.7 Rule B.7 – English FCEL Type

A forename considered for rule B.6, and hence provisionally considered as potentially British or Jewish, but that did not meet the threshold of *'ratio c'>= 2*, thus not being assigned to a FCEL Type, is likely to be an English name, since as mentioned in the previous rule they cannot meet rule B.6 by definition, and is the default most common CEL in the UK. To confirm this, a test is applied to establish whether the combined proportions of the forename associated with non-British SCELs (excluding Jewish) is greater or less than one third of the total occurrences. If it is below one third the forename is then assigned to the FCEL 'English'. If it is equal or above one third it is assigned to a temporary classification 'For Later Review', since the 'non-British' SCELs indicate it may not be a British forename after all.

4.3.2.8 Rule B.8 – International FCEL Type

Taking the set of forenames in the temporary classification 'For Later Review' from rule B.7 those with no occurrences or with only one occurrence in any single non-British SCEL Type, other than 'non-British Unclassified', are assigned to the FCEL Type 'International'. An example of these types of forenames is *Marinda*.

Forename: ***Marinda*** *Total Frequency: 56*

SCEL Groups [SCEL Types]:

British or Jewish 36 (64%) [English 29, Welsh 2, Scottish 3, Irish 2, Jewish 0]

non-British 20 (32%) ['Polish' 1, 'Somali' 1,'non-British Unclassified' 18]

The FCEL Type International is comprised of names that either originated in several different countries or which are widely adopted in several of them as to distinguish a unique origin. Another meaningful example of this FCEL Type is the forename *Felix*.

4.3.2.9 Rule B.9 – Obscure non-British FCELs

A further test is applied to the set of forenames still remaining unclassified, and with a minimum of two occurrences in any one non-British SCEL Type. These forenames are now subject to rules B.3 and B.4, except for the criteria of a having a non-British frequency of 5% or higher of the occurrences in the combined British and Jewish SCEL Groups. This is illustrated with the example *Nelson*:

Forename: ***Nelson*** *Total Frequency: 1628*

SCEL Groups [SCEL Types]:

British or Jewish 1139 (69.96%) [English 756, Welsh 143, Scottish 135, Irish 95, Jewish 2]

non-British 489 (30.04%):

*Hispanic 55,'non-British Unclassified' 387, Others 47 [****Portuguese*** *42, Spanish 11, Others]*

From these figures it can be seen that the largest 'non-British' SCEL Group is Hispanic, with 55 occurrences. This number, as a proportion of those with a British or Jewish SCEL, is just under the 5% threshold so *Nelson* does not qualify in the initial Rule B.3 as a non-British forename. However, the proportion of non-British SCELs (30.04%) is too low for the name to be considered an 'non-British Unclassified' in Rule B.5. The name is not especially associated with Irish, Scottish, Welsh or Jewish SCEL Types in rule B.6, and, because the proportion of occurrences with an English SCEL is slightly below average and non-British SCELs just above 30%, the name does not qualify as English either. However

with 42 of the 55 Hispanic SCEL occurrences classified as Portuguese, the name Nelson qualifies as Portuguese. [2]

4.3.2.10 Rule B. 10 – Black British forenames

Finally, those forenames which by this stage have been assigned a FCEL Type English, Irish, Scottish, Welsh or 'International' (but not Jewish) are selected. For each of these forenames, the proportion of people that are resident in postcodes classified by the *Mosaic* geodemographic classification as being predominantly non-White British based on Census data. These are *Mosaic* Type codes: A01 – Global Connections; C20 – Asian Enterprise; D26 – South Asian Industry; D27 – Settled Minorities; E28 – Counter Cultural Mix; F36 – Metro Multiculture, where overall, 7.2% of British households lived in such types of neighbourhood at the time of the 2001 Census. When the proportion of occurrences of a forename in such neighbourhoods exceeds four times the national average, i.e. exceeds 28.8% of all occurrences, and the proportion of occurrences with British or Jewish SCEL Types is equal to 80% or more, the name is assigned to the FCEL 'Black British'. Note that there is no corresponding SCEL for 'Black' names, since most of these names are associated with Caribbean immigrants or their descendants, most of whom hold British surnames. However there are many forenames that are highly frequent amongst Blacks, a fact also found in the US by Levitt and Dubner (2005).

The name *Hyacinth* is an example of a forename with has been assigned to the FCEL code 'Black' on the basis of rule B.10:

Forename: ***Hyacinth*** *Total Frequency: 1066*

SCEL Groups [SCEL Types]:

British or Jewish 865 (81.1%) [English 712, Welsh 42, Scottish 63, Irish 35, Jewish 13]

non-British 201 (18.9%):

Hispanic 7,'non-British Unclassified' 169, Others [Lithuania 5, Portuguese 4, Spanish 3, Others]

Applying Rules B.1 to B.9 *Hyacinth* is assigned the FCEL Type 'English'. However 40.8% of all occurrences of 'Hyacinth' are resident in one of the six disproportionately Non-White British *Mosaic* neighbourhood types, significantly above the threshold of 28.8% required under rule B.10. The name 'Hyacinth' therefore is one example of many

[2] This assignment may seem odd until one remembers that both Nelson and Wellington became heroes among the Portuguese as a result of their actions in the Peninsula War. Both Nelson and Wellington have become popular personal names in Brazil as well as in Portugal but, for obvious reasons, not among the Spanish speaking populations of the Hispanic SCEL Group.

forenames found among people with British surnames but who live predominantly in Non-White British residential neighbourhoods.

4.3.2.11 Outcome

At the end of these 10 rules described for the Tier 2 file, the 29,979 forenames were classified into FCEL Groups and Types (13,600 British Group and 16,379 non-British) including a small proportion of them being assigned to the 'Unclassified' and 'non-British Unclassified' FCELs (434). These forenames covered 45.3 million people out of the 46.3 million in the UK 2004 file.

4.4 Tier 3: Rest of Names

At this stage the CEL classification system is comprised of two files; Tier 1 with 25,630 surnames, and Tier 2 with 29,979 forenames, each of them assigned to a CEL Group and CEL Type. As mentioned before, the names in these two files respectively cover 37.2 and 45.3 million residents in the GB04 file. Since the purpose of this research project is to classify the entire population into ethnic groups, an additional effort has to be done to classify the remaining surnames and forenames as to cover the 46.3 million people in the GB04 file.

Tier 3 is thus comprised of all the names with 3 or more occurrences in the GB04 dataset and which are not included in either Tier 1 or Tier 2 files, comprising a total of 255,792 surnames and 84,192 forenames. As can be appreciated, the task of classifying Tier 3 involves a substantially higher number of names, but with very low frequencies, most of them with a non-British origin, what requires a different approach to the one previously used. While the rules applied to classify the names in Tiers 1 and 2 where performed in a single cycle, the processing of Tier 3 will be done through a series of repetitive cycles.

4.4.1 Classification by CEL-Triage

There is a known difference in the frequency distribution of surnames and forenames, the latter with an average number of people higher than the former. This is explained by a relatively smaller pool of names from where a society selects their children's forenames, together with the temporal effects in their naming fashions, compared with the fixed nature of surnames, a proportion of which disappear due to a process of 'natural selection' (Manni

et al, 2005). This feature of names has been noted in different countries, for example in the U.S. (Tucker, 2003) and Spain (Mateos, 2007b).

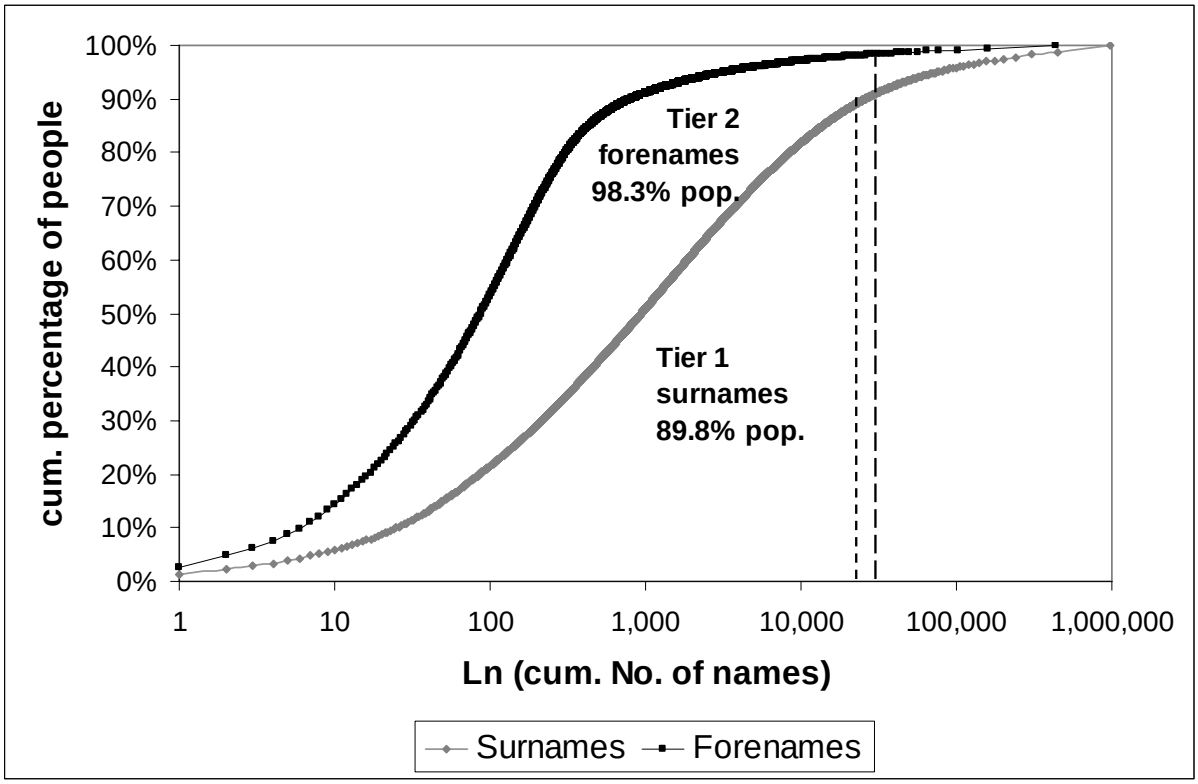


Figure 5: Graph of cumulative number of surnames and forenames (log scale) against cumulative percentage of population in the UK 2004 Electoral Roll

Figure 5 illustrates this difference in the frequency distribution of forenames and surnames for the UK Electoral Roll. The graph shows the cumulative number of surnames or forenames, on a logarithmic scale on the x-axis, against the percentage of population covered, in the y-axis. If the logarithmic scale is not used both curves are so highly positively skewed that no difference is appreciated. The vertical dotted lines represent the cut-off points of both Tier 1 surnames (25,630) and Tier 2 forenames (29,979), and hence the area to the left of these dotted lines represent the total population classified by Tier 1 (89.8%) and Tier 2 (98.3%). The area between the two curves actually represents the number of people in the electoral roll for which their forename is classified but their surname is not.

This difference between the degree of skewness in the frequency distribution of surnames and that of forenames actually permitted the classification of names in Tier 3 in a relatively effortless way, by using the CEL-triage technique described in section 3.1. Since the

25,630 surnames in Tier 1 cover 37.2 million residents, while the 29,979 forenames in Tier 2 cover 45.3 million residents, there is an obvious potential to find out more information about the surnames of the 8.1 million people whose forename appears in Tier 2 (known FCEL) but whose surnames are still unclassified and thus present in Tier 3 (unknown SCEL). These (several thousands) surnames were classified into SCELs using the FCEL distribution of that 8.1 million people, that is, the CEL-triage technique using Tier 1 and Tier 2 SCELs and FCELs. Such classification was performed in a step-wise approach through a series of iterations of the same process that is summarised in Figure 6.

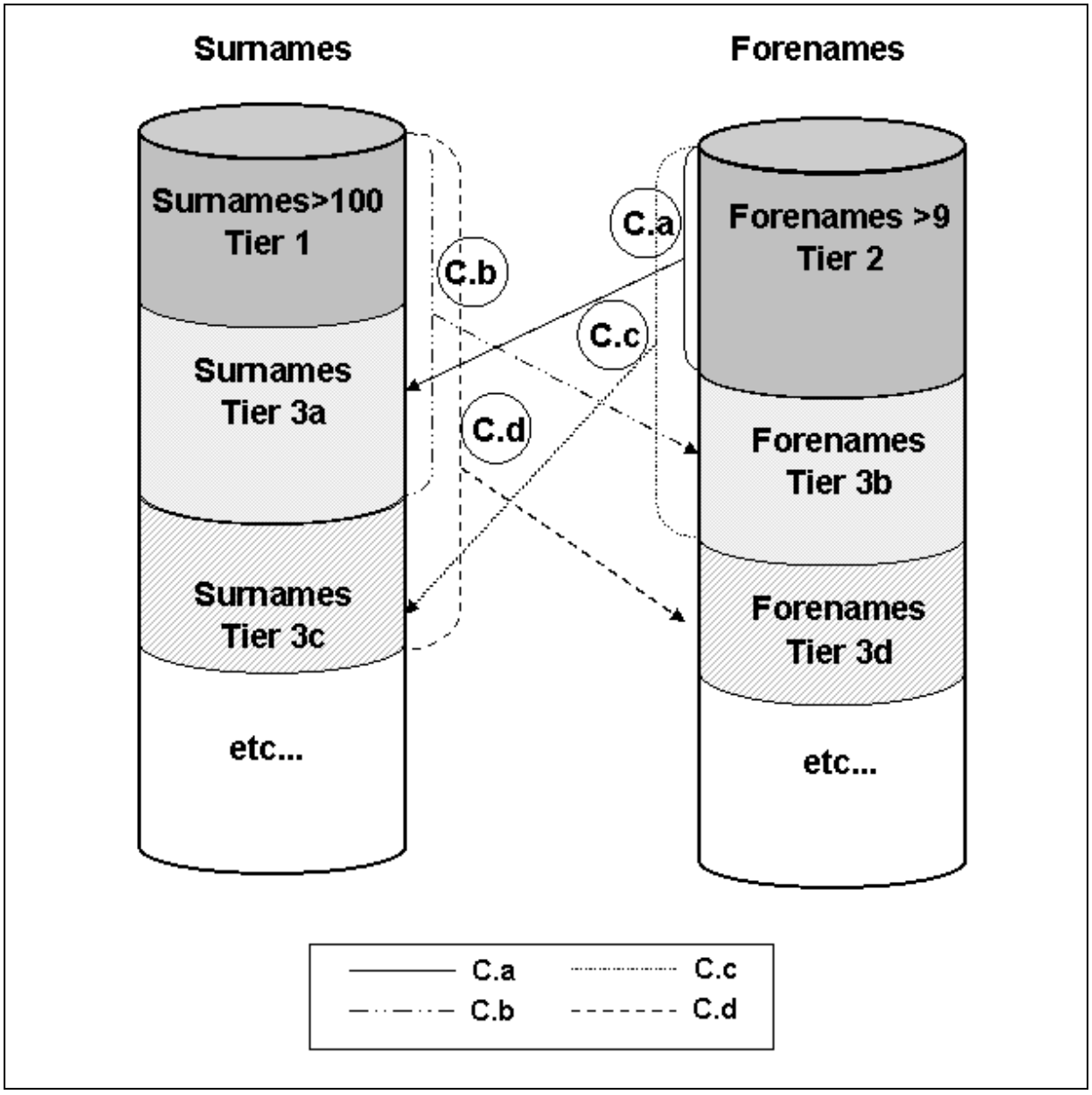


Figure 6: Iterative processing of name classification cycles in Tier 3
Cycles start at 'C.a' which runs from right to left producing set 'Tier 3a', then 'C.b' from left to right producing 'Tier 3b', etc.

Therefore, each iteration or cycle, which are termed 'C.a.' to 'C.d' in Figure 6, aimed to expand the number of names classified, leveraging upon the mentioned difference between the frequency of forenames and surnames, and using all of the names whose CEL were already known at each step (i.e. using Tier 1, Tier 2 and Tier 3 files). As opposed to the classification processes of Tier 1 and 2 previously described, in Tier 3, the cross-CEL distributions for the whole dataset of names (i.e. the count and percentage of each SCEL and FCEL Group and Type associated with each name) were re-calculated at the end of each cycle, thus making the process a dynamic one.

These cycles were run several times, as shown in Figure 6, and as the volume of both surnames and forenames classified grew, it offered new light on previously unclassified names in Tier 3. Finally, the process stopped when most of the names in Tier 3 were classified and a few remained unconnected with the rest of names. This process could even result in the change of a CEL allocation in Tier 1 or Tier 2, when the CEL distributions of the names in Tier 3 pointed to errors in the CEL allocations previously made.

Some of the final unclassified names result from errors in the file such as non-person names (business or building names), names in the wrong fields of the database, name initials or honorifics used, parts of the name missing, or transcription errors. The number that are genuinely unknown names is certainly much lower, but this could only be demonstrated by testing the CEL classification against a good quality population register, currently non-existent in the UK.

4.5 Name-to-CEL tables and scores

At the end of Tier 1, 2 and 3 processes all the classified name files were merged into two separate tables: a surname-to-CEL table with 281,422 surnames, and a forename-to-CEL table with 114,169 forenames. For each name in these two tables the following fields were available: the name's frequency in the UK 2004 file, the CEL Group, and the CEL Type. These two tables will be hereinafter referred to as Name-to-CEL tables.

The set of rules outlined in sections 4.2 to 4.4 are used in order to assign a categorical SCEL classification to each surname and an FCEL to each forename. Such a categorical

assignment is necessary in order to improve and indeed maximise the accuracy of the CEL categories as well as the assignment of personal and surnames. However, once the process of assignment has been finalised then it is natural that in addition to the resulting categorical assignment of names to SCELs and FCELs the use of a proportional assignment is also considered.

This proportional assignment is useful for understanding the large number of names that are associated with more than one cultural, ethnic or linguistic origin. For example, this is the case with the name 'Gill', which has dual origins in Britain and in the Indian Subcontinent and can introduce a bias is assigned only to a single CEL. Proportional assignment is also useful in instances where the actual boundary between different CEL is imprecise, whether geographical, linguistic, religious, or cultural. An example of geographical boundary impreciseness is for instance between the Netherlands and Germany where many names are common in both cultures. Proportional assignment is also useful in situations where one wants to maximise the chance of correctly identifying the final CEL of a person's full name using multiple sources of information, such as for example both elements of the name (forename and surname) in combination with the unit postcode of that person's residence. These aspects of the application of proportional assignment to the classification of actual people by CEL (as opposed to the classification of particular names) is dealt with in section 4.6 while the creation of name scores to facilitate proportional assignment will be explained here.

In order to facilitate the process of proportional assignment of CELs to a person, Name-to-CEL scores need to be created. These scores will represent the degree to which a CEL allocated to a name is actually representative of that name's origin. Going back to the example of 'Gill', ideally this surname should be accompanied by a low score of a South Asian SCEL, so that the FCEL of the person can easily override it if the forename of the person if it is not of South Asian origin.

This rationale is illustrated in the following example for two different persons with the Surname Gill:

John Gill - FCEL; English (57%) SCEL; Indian Sikh (16%) = Person CEL; English

Rabindra Gill - FCEL; North Indian (56%) SCEL; Indian Sikh (16%)

= Person CEL; North Indian

The figures given in brackets are the percentage of people with that surname whose forenames' FCELs are equal to that surname's SCEL, and vice versa for forenames. Name-to-CEL scores were created dividing this percentage by the average percentage for all names of that CEL in the UK. This produces a ratio indicating whether a surname or forename is more common amongst one CEL than the average in the general population. This ratio also remove the effect of an excessively high percentage of British names present in any CEL, due to a 'host country name assimilation phenomena' pointed out by (Tucker, 2003). Since the percentages used for this calculation contained three decimal digits, the resulting ratio between two percentages had potential values between 0 and 100,000. These ratios were finally transformed to a scale between 0 and 1 using a logarithmic transformation function.

Exactly the same process was carried out separately for forenames and surnames, as well as for CEL Types and for CEL Groups (the groupings of CEL Types based on onomastic criteria and shown in Table 2). These scores were added to each entry in the Name-to-CEL tables.

Once the Surname-to-CEL and Forename-to-CEL tables were completed each of them was comprised of the following information per name:

- Name
- GB Frequency (total count of occurrences in the GB04 file)
- CEL Type (out of 185 possible types described in Table 2)
- CEL Group (fixed grouping of CEL Types based on onomastic criteria)
- CEL Type score (a number between 0 and 1 with two decimal digits)
- CEL Group score (a number between 0 and 1 with two decimal digits)

4.6 Creating a person level CEL allocation system

This section explains the process by which the CEL classification created can be applied to a list of names in a Target Population, in order to classify people with their most likely CEL. This person level CEL will be termed PCEL (for Person CEL), as opposed to the two separate FCEL and SCEL of its name components.

The two Name-to-CEL tables explained in the previous section, are then used as 'dictionaries' against which a person's full name can be assigned to a PCEL, taking into account both the person's FCEL and SCEL. An individual's full name is evaluated as per the following algorithm of 6 cases evaluated in order from 1 to 6:
(Algorithm presented as pseudo-code, with comments tagged as '##' and in italics)

Evaluate if both CEL Types are the same

CASE1 SCEL Type = FCEL Type, then:

Assign PCEL

PCEL = SCEL Type = FCEL Type

Evaluate if CEL Groups are the same and if so assign that CEL Group

CASE2 SCEL Group = PCEL Group, then

PCEL= SCEL Group= PCEL Group

If the absolute difference between scores is small then assign PCEL to the CEL Group with the highest score

CASE3 |SCEL Type score - FECL Type score| < 0.05, then

PCEL= MAX(SCEL or PCEL Group score)

Evaluate if both SCEL and FCEL exist for that person and assign PCEL to the CEL Type with the highest score

CASE4 SCEL AND FCEL exist, then

PCEL= MAX(SCEL or PCEL Type score)

If only one CEL component is present, then assign at the CEL Group Level

CASE5 SCEL or FCEL = 'UNCLASSIFIED' then

PCEL= SCEL Group or PCEL Group

Else, set the PCEL as unclassified

ELSE PCEL= 'UNCLASSIFIED'

At the end of this process each person's full name will have an overall PCEL assigned to it, at the CEL Type or CEL Group level, or remains unclassified. Furthermore, apart from selecting the most likely CEL for a person, the classification also provides a final CEL score for the person. This will be useful when analysing the final results since the user can set a minimum threshold from which to choose people-to-CEL assignments depending on the sensitivity of each specific application of this methodology. In other words, one can choose to aim for precision in the classification and to select a small group of individuals

that have a very high PCEL scores, and thus with a high probability of belonging to a specific CEL, or to aim to maximise coverage and include lower score names, but classifying more individuals. A similar approach is proposed by Word and Perkins (1996) for a Spanish surnames list and by Lauderdale and Kestenbaum (2000) for an Asian surnames list.

The PCEL score for the person is calculated as follows, depending on which case in the previous algorithm was the PCEL assigned to:

For coincident SCEL and FCEL the scores are added

PCEL under CASE1, CASE2 and CASE5

PCEL score = SCEL score + FCEL score (either Type or Group as used above)

For divergent SCEL and FCEL the scores are subtracted

PCEL under CASE3 and CASE4

PCEL score = |SCEL Type score - FECL Type score|

Else assign a score of 0

ELSE PCEL score= 0

At the end of the individuals' classification process, the list of people's full names in the target population is classified with a PCEL and a PCEL score in a scale from 0 to 2, with an indication to whether the PCEL assignment was performed at the CEL Type (Cases 1 and 4 above) or CEL Group (Cases 2, 3 and 5 above) levels.

5 Validating the CEL Name classification

The final step in the construction of the CEL methodology reported here is to evaluate the extent to which the CEL classification is effective in classifying ethnicity. A selection of the best-practice examples of evaluations of name classifications found in the literature is presented in a systematic review by Mateos (2007a). This review compared thirteen studies, most of which emanate from the public health literature, which each validated their name classifications at the individual level. This external validation used lists of individuals that included ethnicity, country of birth or nationality, as part of using patient registers. Such external validation at the individual level will only be carried out during a subsequent stage of the research project presented here, and hence will not be covered by this paper.

The type of validation attempted here follows a 'geography tradition' and seeks to evaluate the ability of the CEL classification to correctly identify ethnicity at the level of the small area aggregation (as opposed to unique human individuals). The validation uses the ethnicity data reported in the UK Census of Population at small area (Census Output Area), which is compared against the CEL classification of the same areas using the names in the GB Electoral Roll.

5.1 Data preparation

This validation requires the use of two datasets to be compared; Census 2001 Key Statistics KS06 table (Ethnic Group), and a new dataset to be prepared by coding the GB Electoral Roll by CEL Type, which is then aggregated by the Census Output Area geography, and by the Census 16 ethnic groups.

Each of the 46.3 million adults in the GB 2004 Electoral Roll file, described in 2.3, was classified by CEL Type, using the Name-to-CEL tables and the algorithms explained in section 4.6. This comprises, to our knowledge, the first attempt ever to classify the whole population of Great Britain by the cultural ethnic and linguistic origin of their name, and the results are presented in Table 5 summarized by CEL Group. The individual records were then aggregated by unit postcode, calculating the number of people per CEL Type in each unit postcode, what produced a table of 1.4 million records and 185 columns.

Two further steps are required in order to make the electoral roll dataset comparable with the Census; the aggregation of the 185 CEL Types to 16 Census ethnic groups, and of the unit postcode geography to Census Output Areas. The first step is achieved through the CEL Type lookup table previously used that relates the 185 CEL Types to various attributes, one of them being the sixteen 2001 Census ethnic groups, and which appears in Table 7 in the Appendix. Therefore, the 185 columns are now aggregated into the mentioned 16 ethnic groups.

The second step requires the use of an additional dataset, the postcode directory maintained by Office of National Statistics, named the National Statistics Postcode Directory (NSPD) (formerly known as the All Fields Postcode Directory (AFPD) and previously the *Gridlink*

Postcode Directory) (Office for National Satistics, 2006). The NSPD provides a lookup table between every unit postcode in the UK to a set of higher level geographies to which it belongs, which of interest to this validation are; Census Output Area (OA), Lower level Super Output Area (LSOAs), Ward, and Local Authority (LA). Therefore, the 1.4 million records were separately aggregated by each of these four geographic levels, generating four tables of different spatial resolutions; 218,037 OAs, 40,883 LSOAs, 10,072 Wards, and 408 Local Authorities in Great Britain (i.e. excluding Northern Ireland). Each of these tables contains counts of persons in the Electoral Roll by ethnic group based on CEL but expressed as their 2001 Census ethnic group equivalent. To avoid confusion, this four tables will be generally refered to as the CEL-GB04 datasets.

CEL Group	***People***	***%***
ENGLISH	29,455,761	67.6%
CELTIC	10,485,126	24.1%
MUSLIM	987,422	2.3%
EUROPEAN	735,105	1.7%
SOUTH ASIAN	475,834	1.1%
SIKH	275,939	0.6%
NORDIC	222,859	0.5%
HISPANIC	186,381	0.4%
EAST ASIAN	159,668	0.4%
AFRICAN	149,076	0.3%
GREEK	102,646	0.2%
JEWISH AND ARMENIAN	80,650	0.2%
JAPANESE	5,829	0.0%
INTERNATIONAL	35,763	0.1%
VOID	210,803	0.5%
UNCLASSIFIED	20,942	0.0%
TOTAL	43,589,804	100.0%
Total valid CELs	43,322,296	99.4%
Total non-valid CELs	267,508	0.6%

Table 5: Summary of CEL Groups in the GB 2004 Electoral Roll

5.2 Data Analysis: Validation of CEL Vs Census Ethnicity at small area

The comparison between the CEL-GB04 datasets and the Census ethnic groups by small area (2001 Census Key Statistics KS06 table), is then achieved by linking both datasets at each of the geographical levels for which the validation will be performed; OA, LSOA, Ward and LA. The idea of performing the validation at four different geographical scales is to evaluate how sensitive is the CEL classification to changes in scale, what will then determine the optimum geographical level of its applications.

This analysis entailed calculating correlation coefficients between the CEL-GB04 dataset and the Census ethnicity responses at the four different levels of geography. Since the two datasets do no use the same denominator (the CEL file only includes adults entitled to vote while the Census enumerates all of the resident population), the comparison was performed using the proportion of people in each ethnic group for each geographical unit, upon which a correlation matrix was calculated using Pearson's correlation coefficient (Robinson, 1998). A summary of the results is offered in Table 6, which summarises the correlation coefficients at the different levels of geography for which they were calculated (OA, LSOA, Ward and Borough), and for 11 Census ethnic groups, after removing the four 'Mixed' and the 'Not Stated' categories. The number of geographical units at each level and their population size are indicated at the bottom of the table.

	Geographical Unit of Comparison			
Ethnic Group	***OA***	***LSOA***	***WARD***	***LA***
A) White - British	**0.88**	**0.93**	**0.93**	**0.95**
B) White - Irish	0.32	0.37	0.42	0.46
C) White - Any other White background	**0.74**	**0.85**	**0.88**	**0.93**
H) Asian or Asian British - Indian	**0.92**	**0.95**	**0.96**	**0.98**
J) Asian or Asian British - Pakistani	**0.90**	**0.93**	**0.93**	**0.91**
K) Asian or Asian British - Bangladeshi	**0.91**	**0.93**	**0.95**	**0.98**
L) Asian or Asian British - Any other Asian background	-0.06	0.11	0.24	0.62
M) Black or Black British - Caribbean	0.32	**0.77**	**0.91**	**0.98**
N) Black or Black British - African	**0.83**	**0.95**	**0.97**	**0.99**
R) Other Ethnic Groups - Chinese	0.65	**0.79**	**0.84**	**0.97**
S) Other Ethnic Groups - Any other ethnic group	0.38	0.66	**0.77**	**0.88**
Number of Units valid for analysis	218,037	40,883	10,072	408

Table 6: Summary of Pearson's correlation coefficients between the CEL-GB04 and 2001 Census datasets

All correlations are significant at the 0.01 level (2-tailed). Correlations ≥ 0.7 are highlighted in bold
OA = Output Area, LSOA= Lower Super Output Area, LA= Local Authority

5.3 Discussion of results

Most of the ethnic categories present a high degree of correlation between the two datasets, which generally increases with area size, but that it is still strong at OA level. This is however expected as per the definition of the classic *modifiable areal unit problem* (MAUP) (Openshaw, 1984). There are however some groups for which anomalies occur; the 'White-Irish' (B), 'Any other Asian background' (L), and 'Any Other Ethnic Group' (S) categories, and to a lesser extent the 'Black-Caribbean' (M) group, which present an overall low correlation. The main reasons for this divergence are, on the one hand, the inherent vagueness of some Census categories ('Mixed', 'Other') together with their lack of exact correspondence to the CEL taxonomy, and on the other hand, some problems detected in the distinction of Irish and Caribbean names, that are due to historic differences and a high degree of assimilation with the White-British majority. Nevertheless, the correlation coefficient of all the categories is significant at the 0.01 level (2-tailed). The stronger correlations (≥ 0.7) are highlighted in bold in Table 6, including 6 out of 11 categories at OA level, and especially at the LSOA level (1,500 persons in average) and coarser geographies. Some groups perform extraordinarily well across all scales; 'White British' (A), 'White-Other' (C), 'Indian' (H), 'Pakistani' (J), 'Bangladeshi' (K), 'Black African' (N) and to a lesser extent Chinese (R), probably indicating the robustness of their Census ethnic categories as well as a strong linkage between current self-perception of ethnic identity and name origins for those groups. With the exception of 'White-Other' and 'Indian', these are also the ethnic groups which performed best in the validation using the HES dataset in Camden and Islington, and area with fewer residents from the 'Indian' ethnic group. However, at LSOA and higher geographies, all groups except for the 'Other' mentioned (L & S) perform very well when compared with the Census geographical distribution.

6 Applications of the CEL classification

The preliminary evaluation of the CEL Name classification presented in this paper suggests that it is successful in classifying neighbourhoods by ethnic group. Further analysis, currently in progress, will seek to establish its validity in classifying individuals in patient lists. However, there are other fields of study in which this methodology may be

also applied, in order to provide better insights into the reality of ethnic minorities in terms of our finely detail defined cultural, ethnic and linguistic groups (CELs).

We believe that name classification methods have great potential in broader population studies that have ethnic dimensions. These include: ethnic group population forecasting by small area (Large and Ghosh, 2006); monitoring migration (Stillwell and Duke-Williams, 2005); detecting Census undercount (Graham and Waterman, 2005); measuring residential segregation (Simpson, 2004); analysing the geography of ethnic inequalities (Dorling and Rees, 2003) or of mortality and morbidity (Boyle, 2004); evaluating equal opportunity policies (Johnston et al, 2004) and political empowerment processes (Clark and Morrison, 1995); and improving public and private services to ethnic minorities (Van Ryn and Fu, 2003). Each of these research and public policy areas remains characterised by a lack of appropriate, timely and detailed data on ethnicity, a problem that is increasing as the last round of Census data age and new migration flows are changing the composition and demands for public services. Improved methods in these areas are thus of key policy importance in today's multi-cultural society.

Beyond population studies, some of the potential areas of application of the CEL Name classification can be found in public services provision, private sector marketing and the implementation of equal opportunities legislation. These areas include: the identification of and communication with ‘hard-to-reach’ ethnic groups through appropriate channels and at fine geographical level; classification of employee names files or audit of take up of public services; monitoring progress towards equal opportunities objectives; and devising strategies for sampling ethnic minorities in surveys, for example through the classification of names in the electoral roll or telephone directory. There are a number of precedents to this last area of application in the literature, with a range of studies that have typically used telephone directories to select names from a particular ethnic group as a sampling strategy for their surveys, showing the usefulness of the name-based approach to classify Vietnamese (Hinton et al, 1998, Rahman et al, 2005), Korean (Hofstetter et al, 2004), Cambodian (Tu et al, 2002), Chinese (Hage et al, 1990, Lai, 2004), South Asian (Chaudhry et al, 2003), Japanese (Kitano et al, 1988), Irish (Abbotts et al, 1999), Jewish (Himmelfarb et al, 1983) and Lebanese (Rissel et al, 1999) names, in the U.S., Canada, U.K. and Australia.

In the applications listed here name classifications have proved very useful in segmenting populations when no other ethnicity data are available, or where a much finer definition of ethnic groups is required than is available from the current 2001 Census ethnicity classification – as, for example, when identifying an increase in Polish migrants since 2004 in emergency admissions to hospital (Leaman et al, 2006). Furthermore, the classification can be tailored to each application by selecting a specific minimum name score threshold, which provides a measure of the probability of ethnicity of a name. The sensitivity of the application will determine whether the objective is to maximize the number of population classified (low score) or to minimize potential classification errors (high score). Another way by which the user can tailor the classification is to aggregate the 185 CEL Types in different ways that are best suited to a particular purpose, for example by language, religion, or geography, as well as different ethnicity groupings.

7 Conclusion

The CEL Name-based ethnicity classification methodology presented in this paper offers a series of potential advantages to identify and quantify ethnicity over traditional information sources such as censuses of population. Amongst these advantages: it can develop a more detailed and meaningful classification of people's origins (185 fine categories or CELs based on a very large number of languages versus just 10 to 16 ethnic groups in the Census); it offers improved updating (annually through registers with substantial population coverage, such as electoral or patient registers); it accommodates changing perceptions of identity better than ethnicity self-classification (through independent assignment of ethnicity and or cultural origins according to name); and is made available, subject to confidentiality safeguards, at the individual or household level (rather than an aggregated Census area). Moreover, according to the literature its main advantage remains its capability to provide an ethnicity classification where self-reported ethnicity is not available (which is the case in most population registers and datasets about individuals), at a fraction of the cost of other methods. These are the reasons why name-based ethnicity classifications are considered to have an important potential, and research on improvements to the original methods has been encouraged by Bhopal et al (2004) and Peach and Owen (2004).

This paper has presented the work in progress towards achieving a name classification system that covers all potential ethnic groups in the UK. Departing from the previous methods developed in this area, the research project described here has taken on the challenge of building a name-based ethnicity classification using a very large reference population, maximised for total population coverage in the entire UK. As such, the method's objective is to classify complete populations into all of the potential ethnic groups present in UK society at present time. It has also developed name-to-ethnicity probability scores, to be used in the arbitration between potentially conflicting ethnicities arising from the different name components of a person, as well as to facilitate the tailoring of the classification to the sensitivity of each specific application.

Finally, this paper has validated the CEL Name classification using the 2001 Census of Population, through which the initial effectiveness of this method has been demonstrated for the main ethnic groups at the neighbourhood level. Further external validation at the individual level is required to compare this methodology with methods that are similar, but significantly more limited in scope, that have previously been published in the literature.

It is now in the hands of social scientists, health researchers, public service practitioners, and a realm of other disciplines, to test this method in their own fields. It is our contention that a wide range of applications might benefit from its innovative classification power, and that a range of new applications may also both develop and contribute to improvement of the methodology.

8 Appendix

CEL Type Master Lookup Table

Lookup table between CEL Types and various onomastic groups (CEL Group), languages, religions, geographical area, and Census categories

CEL Type Code	CEL Group	CEL Type	People GB 2004	Geographical Area	Religion	Major Language	1991 Census Ethnic Group	2001 Census Ethnic Group
AF110	AFRICAN	AFRICA	3,316	AFRICA	CHRISTIAN: PROTESTANT	Not Applicable	2- Black - African	N) Black or Black British - African
AF211	AFRICAN	BLACK SOUTHERN AFRICA	5,198	AFRICA	CHRISTIAN: PROTESTANT	Zulu	2- Black - African	N) Black or Black British - African
AF212	AFRICAN	BOTSWANA	8	AFRICA	CHRISTIAN: PROTESTANT	Tswana	2- Black - African	N) Black or Black British - African
AF213	AFRICAN	CONGO	1,164	AFRICA	CHRISTIAN	Luba-Kasai	2- Black - African	N) Black or Black British - African
AF214	AFRICAN	MADAGASCAR	2	AFRICA	CHRISTIAN: CATHOLIC	Malagasy	2- Black - African	N) Black or Black British - African
AF215	AFRICAN	MALAWI	23	AFRICA	CHRISTIAN: PROTESTANT	Nyanja	2- Black - African	N) Black or Black British - African
AF216	AFRICAN	NAMIBIA	1	AFRICA	CHRISTIAN: CATHOLIC	Afrikaans	0- White	C) White - Any other White background
AF217	AFRICAN	OTHER AFRICAN	1,575	AFRICA	CHRISTIAN	Not Applicable	2- Black - African	N) Black or Black British - African
AF218	AFRICAN	SWAZILAND	5	AFRICA	CHRISTIAN: PROTESTANT	Swati	2- Black - African	N) Black or Black British - African
AF219	AFRICAN	ZAIRE	41	AFRICA	CHRISTIAN: PROTESTANT	Luba-Kasai	2- Black - African	N) Black or Black British - African
AF220	AFRICAN	ZAMBIA	274	AFRICA	CHRISTIAN: PROTESTANT	Bemba	2- Black - African	N) Black or Black British - African
AF221	AFRICAN	ZIMBABWE	991	AFRICA	CHRISTIAN: PROTESTANT	Shona	2- Black - African	N) Black or Black British - African
AF322	AFRICAN	BURUNDI	18	AFRICA	CHRISTIAN	Rundi	2- Black - African	N) Black or Black British - African
AF323	AFRICAN	ETHIOPIA	1,238	AFRICA	CHRISTIAN: OTHER	Amharic	2- Black - African	N) Black or Black British - African
AF324	AFRICAN	KENYAN AFRICAN	1,197	AFRICA	CHRISTIAN: PROTESTANT	Gikuyu	2- Black - African	N) Black or Black British - African
AF325	AFRICAN	RWANDA	18	AFRICA	CHRISTIAN	Rwanda	2- Black - African	N) Black or Black British - African
AF326	AFRICAN	SUDAN	5	AFRICA	MUSLIM	Arabic	2- Black - African	N) Black or Black British - African
AF327	AFRICAN	TANZANIA	104	AFRICA	CHRISTIAN: PROTESTANT	English	2- Black - African	N) Black or Black British - African
AF328	AFRICAN	UGANDA	1,018	AFRICA	CHRISTIAN: PROTESTANT	Ganda	2- Black - African	N) Black or Black British - African
AF429	AFRICAN	BENIN	7	AFRICA	MUSLIM	French	2- Black - African	N) Black or Black British - African
AF430	AFRICAN	CAMEROON	72	AFRICA	CHRISTIAN	Fulfulde	2- Black - African	N) Black or Black British - African
AF431	AFRICAN	GAMBIA	11	AFRICA	MUSLIM	Wolof	2- Black - African	N) Black or Black British - African

CEL Type Code	CEL Group	CEL Type	People GB 2004	Geographical Area	Religion	Major Language	1991 Census Ethnic Group	2001 Census Ethnic Group
AF432	AFRICAN	GHANA	46,095	AFRICA	CHRISTIAN	Akan	2- Black - African	N) Black or Black British - African
AF433	AFRICAN	GUINEA	15	AFRICA	MUSLIM	French	2- Black - African	N) Black or Black British - African
AF434	AFRICAN	IVORY COAST	122	AFRICA	MUSLIM	Baoulé	2- Black - African	N) Black or Black British - African
AF435	AFRICAN	LIBERIA	5	AFRICA	MUSLIM	Kpelle	2- Black - African	N) Black or Black British - African
AF436	AFRICAN	NIGERIA	88,243	AFRICA	CHRISTIAN	Yoruba	2- Black - African	N) Black or Black British - African
AF437	AFRICAN	SENEGAL	37	AFRICA	MUSLIM	Wolof	2- Black - African	N) Black or Black British - African
AF438	AFRICAN	SIERRA LEONE	6,155	AFRICA	MUSLIM	English	2- Black - African	N) Black or Black British - African
CL110	CELTIC	CELTIC	45,653	BRITISH ISLES	CHRISTIAN	English	0- White	A) White - British
CL211	CELTIC	IRELAND	3,172,876	BRITISH ISLES	CHRISTIAN: CATHOLIC	English	0- White	B) White - Irish
CL212	CELTIC	NORTHERN IRELAND	223,988	BRITISH ISLES	CHRISTIAN	English	0- White	A) White - British
CL213	CELTIC	SCOTLAND	4,749,864	BRITISH ISLES	CHRISTIAN: PROTESTANT	English	0- White	A) White - British
CL314	CELTIC	WALES	3,065,041	BRITISH ISLES	CHRISTIAN: PROTESTANT	Welsh	0- White	A) White - British
EA110	EAST ASIAN	EAST ASIA	627	EAST ASIA	BHUDDIST	Chinese, Mandarin	7- Chinese	R) Other Ethnic Groups - Chinese
EA211	EAST ASIAN	SOUTH EAST ASIA	371	EAST ASIA	BHUDDIST	Chinese, Min Nan	7- Chinese	R) Other Ethnic Groups - Chinese
EA212	EAST ASIAN	CHINA	21,185	EAST ASIA	BHUDDIST	Chinese, Mandarin	7- Chinese	R) Other Ethnic Groups - Chinese
EA213	EAST ASIAN	EAST ASIAN CARIBBEAN	2	AMERICAS	BHUDDIST	Chinese, Mandarin	7- Chinese	R) Other Ethnic Groups - Chinese
EA215	EAST ASIAN	HONG KONG	119,566	EAST ASIA	BHUDDIST	Chinese, Cantonese	7- Chinese	R) Other Ethnic Groups - Chinese
EA218	EAST ASIAN	MALAYSIAN CHINESE	3,238	EAST ASIA	BHUDDIST	Chinese, Min Nan	7- Chinese	R) Other Ethnic Groups - Chinese
EA221	EAST ASIAN	SINGAPORE	583	EAST ASIA	BHUDDIST	Chinese, Min Nan	7- Chinese	R) Other Ethnic Groups - Chinese
EA225	EAST ASIAN	TIBET	13	EAST ASIA	BHUDDIST	Tibetan	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
EA316	EAST ASIAN	INDONESIA	116	EAST ASIA	MUSLIM	Javanese	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
EA317	EAST ASIAN	MALAYSIA	2,092	EAST ASIA	MUSLIM	Malay	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
EA319	EAST ASIAN	MYANMAR	1,601	EAST ASIA	BHUDDIST	Burmese	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
EA323	EAST ASIAN	SOUTH KOREA	2,315	EAST ASIA	BHUDDIST	Korean	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
EA324	EAST ASIAN	THAILAND	407	EAST ASIA	BHUDDIST	Thai	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
EA326	EAST ASIAN	VIETNAM	15,723	EAST ASIA	BHUDDIST	Vietnamese	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
EA327	EAST ASIAN	CAMBODIA	59	EAST ASIA	BHUDDIST	Khmer, Central	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group

CEL Type Code	CEL Group	CEL Type	People GB 2004	Geographical Area	Religion	Major Language	1991 Census Ethnic Group	2001 Census Ethnic Group
EA328	EAST ASIAN	LAOS	120	EAST ASIA	BHUDDIST	Lao	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
EA414	EAST ASIAN	FIJI	10	EAST ASIA	HINDU	Hindustani	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
EA420	EAST ASIAN	POLYNESIA	54	EAST ASIA	CHRISTIAN	Tahitian	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
EA422	EAST ASIAN	SOLOMON ISLANDS	8	EAST ASIA	CHRISTIAN	English	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
EA429	EAST ASIAN	HAWAII	2	EAST ASIA	CHRISTIAN	Hawaii Creole English	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
EA430	EAST ASIAN	MAORI	1	EAST ASIA	CHRISTIAN	Maori	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
EA431	EAST ASIAN	MAURITIUS	2	EAST ASIA	HINDU	Hindi	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
EA432	EAST ASIAN	SAMOA	10	EAST ASIA	CHRISTIAN	Samoan	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
EA433	EAST ASIAN	TONGA	9	EAST ASIA	CHRISTIAN	Tongan	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
EA434	EAST ASIAN	TUVALU	2	EAST ASIA	CHRISTIAN	Tuvaluan	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
EN110	ENGLISH	ENGLAND	31,118,965	BRITISH ISLES	CHRISTIAN: PROTESTANT	English	0- White	A) White - British
EN211	ENGLISH	CORNWALL	107,068	BRITISH ISLES	CHRISTIAN: PROTESTANT	English	0- White	A) White - British
EN213	ENGLISH	CHANNEL ISLANDS	23,995	BRITISH ISLES	CHRISTIAN: PROTESTANT	English	0- White	A) White - British
EN314	ENGLISH	BRITISH SOUTH AFRICA	45	AFRICA	CHRISTIAN: PROTESTANT	English	0- White	A) White - British
EN315	ENGLISH	BLACK CARIBBEAN	23,665	AMERICAS	CHRISTIAN: PROTESTANT	English	1- Black - Caribbean	M) Black or Black British - Caribbean
EU110	EUROPEAN	EUROPEAN	31,341	CENTRAL EUROPE	CHRISTIAN	German	0- White	C) White - Any other White background
EU211	EUROPEAN	BELGIUM	815	CENTRAL EUROPE	CHRISTIAN	French	0- White	C) White - Any other White background
EU212	EUROPEAN	BELGIUM (FLEMISH)	4,417	CENTRAL EUROPE	CHRISTIAN: PROTESTANT	Vlaams	0- White	C) White - Any other White background
EU213	EUROPEAN	BELGIUM (WALLOON)	618	CENTRAL EUROPE	CHRISTIAN: CATHOLIC	French	0- White	C) White - Any other White background
EU214	EUROPEAN	NETHERLANDS	20,495	CENTRAL EUROPE	CHRISTIAN	Dutch	0- White	C) White - Any other White background
EU215	EUROPEAN	AFRIKAANS	7,805	AFRICA	CHRISTIAN: PROTESTANT	Afrikaans	0- White	C) White - Any other White background
EU316	EUROPEAN	FRANCE	125,754	CENTRAL EUROPE	CHRISTIAN: CATHOLIC	French	0- White	C) White - Any other White background

CEL Type Code	CEL Group	CEL Type	People GB 2004	Geographical Area	Religion	Major Language	1991 Census Ethnic Group	2001 Census Ethnic Group
EU317	EUROPEAN	BRETON	640	CENTRAL EUROPE	CHRISTIAN: CATHOLIC	French	0- White	C) White - Any other White background
EU318	EUROPEAN	CANADA	299	AMERICAS	CHRISTIAN: PROTESTANT	English	0- White	C) White - Any other White background
EU319	EUROPEAN	FRENCH CARIBBEAN	3	AMERICAS	CHRISTIAN: CATHOLIC	French	1- Black - Caribbean	M) Black or Black British - Caribbean
EU420	EUROPEAN	GERMANY	129,190	CENTRAL EUROPE	CHRISTIAN	German	0- White	C) White - Any other White background
EU421	EUROPEAN	SWITZERLAND	128	CENTRAL EUROPE	CHRISTIAN	Schwyzerdütsch	0- White	C) White - Any other White background
EU522	EUROPEAN	ITALY	229,931	SOUTHERN EUROPE	CHRISTIAN: CATHOLIC	Italian	0- White	C) White - Any other White background
EU523	EUROPEAN	MALTA	8,027	SOUTHERN EUROPE	CHRISTIAN: CATHOLIC	Maltese	0- White	C) White - Any other White background
EU624	EUROPEAN	ESTONIA	778	EASTERN EUROPE	CHRISTIAN: PROTESTANT	Estonian	0- White	C) White - Any other White background
EU625	EUROPEAN	LATVIA	1,559	EASTERN EUROPE	CHRISTIAN: CATHOLIC	Latvian	0- White	C) White - Any other White background
EU626	EUROPEAN	LITHUANIA	1,790	EASTERN EUROPE	CHRISTIAN: CATHOLIC	Lithuanian	0- White	C) White - Any other White background
EU727	EUROPEAN	BALKAN	16,274	EASTERN EUROPE	CHRISTIAN	Serbian	0- White	C) White - Any other White background
EU728	EUROPEAN	SERBIA	5,279	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Serbian	0- White	C) White - Any other White background
EU729	EUROPEAN	BOSNIA AND HERZEGOVINA	1,034	EASTERN EUROPE	MUSLIM	Bosnian	0- White	C) White - Any other White background
EU730	EUROPEAN	MONTENEGRO	44	EASTERN EUROPE	MUSLIM	Serbian	0- White	C) White - Any other White background
EU731	EUROPEAN	MACEDONIA	371	EASTERN EUROPE	CHRISTIAN: GREEK ORTHODOX	Macedonian	0- White	C) White - Any other White background
EU732	EUROPEAN	SLOVENIA	1,282	EASTERN EUROPE	CHRISTIAN: CATHOLIC	Slovenian	0- White	C) White - Any other White background
EU733	EUROPEAN	CROATIA	1,362	EASTERN EUROPE	CHRISTIAN: CATHOLIC	Croatian	0- White	C) White - Any other White background
EU734	EUROPEAN	ALBANIA	3,440	EASTERN EUROPE	CHRISTIAN: GREEK ORTHODOX	Albanian	0- White	C) White - Any other White background
EU835	EUROPEAN	POLAND	155,743	EASTERN EUROPE	CHRISTIAN: CATHOLIC	Polish	0- White	C) White - Any other White background
EU836	EUROPEAN	CZECH REPUBLIC	4,357	EASTERN EUROPE	CHRISTIAN: CATHOLIC	Czech	0- White	C) White - Any other White background
EU837	EUROPEAN	SLOVAKIA	524	EASTERN EUROPE	CHRISTIAN	Slovak	0- White	C) White - Any other White background
EU838	EUROPEAN	HUNGARY	11,768	EASTERN EUROPE	CHRISTIAN: CATHOLIC	Hungarian	0- White	C) White - Any other White background

CEL Type Code	CEL Group	CEL Type	People GB 2004	Geographical Area	Religion	Major Language	1991 Census Ethnic Group	2001 Census Ethnic Group
EU839	EUROPEAN	BULGARIA	109	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Bulgarian	0- White	C) White - Any other White background
EU840	EUROPEAN	ROMANIA	744	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Romanian	0- White	C) White - Any other White background
EU841	EUROPEAN	ROMANIA BANAT	29	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Romanian	0- White	C) White - Any other White background
EU842	EUROPEAN	ROMANIA DOBREGA	28	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Romanian	0- White	C) White - Any other White background
EU843	EUROPEAN	ROMANIA MANAMURESCRIANA	331	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Romanian	0- White	C) White - Any other White background
EU844	EUROPEAN	ROMANIA MOLDOVA	200	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Romanian	0- White	C) White - Any other White background
EU845	EUROPEAN	ROMANIA MUNTENIA	364	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Romanian	0- White	C) White - Any other White background
EU846	EUROPEAN	ROMANIA TRANSILVANIA	835	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Romanian	0- White	C) White - Any other White background
EU947	EUROPEAN	RUSSIA	11,118	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Russian	0- White	C) White - Any other White background
EU948	EUROPEAN	BELARUS	27	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Belarusan	0- White	C) White - Any other White background
EU949	EUROPEAN	UKRAINE	3,948	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Ukrainian	0- White	C) White - Any other White background
EU950	EUROPEAN	AZERBAIJAN	12	CENTRAL ASIA	MUSLIM	Azerbaijani, North	0- White	C) White - Any other White background
EU951	EUROPEAN	GEORGIA	185	CENTRAL ASIA	CHRISTIAN: RUSSIAN ORTHODOX	Georgian	0- White	C) White - Any other White background
GR110	GREEK ORTHODOX	GREECE	29,134	SOUTHERN EUROPE	CHRISTIAN: GREEK ORTHODOX	Greek	0- White	C) White - Any other White background
GR211	GREEK ORTHODOX	GREEK CYPRUS	79,304	SOUTHERN EUROPE	CHRISTIAN: GREEK ORTHODOX	Greek	0- White	C) White - Any other White background
GR212	GREEK ORTHODOX	GREEK ORTHODOX	932	SOUTHERN EUROPE	CHRISTIAN: GREEK ORTHODOX	Greek	0- White	C) White - Any other White background
HI110	HISPANIC	HISPANIC	6,084	SOUTHERN EUROPE	CHRISTIAN: CATHOLIC	Spanish	0- White	C) White - Any other White background
HI211	HISPANIC	PORTUGAL	86,930	SOUTHERN EUROPE	CHRISTIAN: CATHOLIC	Portuguese	0- White	C) White - Any other White background
HI212	HISPANIC	BRAZIL	1,949	AMERICAS	CHRISTIAN: CATHOLIC	Portuguese	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
HI213	HISPANIC	ANGOLA	458	AFRICA	CHRISTIAN: CATHOLIC	Portuguese	2- Black - African	N) Black or Black British - African
HI214	HISPANIC	GOA	990	SOUTH ASIA	CHRISTIAN: CATHOLIC	Portuguese	4- Indian	H) Asian or Asian British - Indian

CEL Type Code	CEL Group	CEL Type	People GB 2004	Geographical Area	Religion	Major Language	1991 Census Ethnic Group	2001 Census Ethnic Group
HI315	HISPANIC	SPAIN	80,180	SOUTHERN EUROPE	CHRISTIAN: CATHOLIC	Spanish	0- White	C) White - Any other White background
HI316	HISPANIC	CASTILLIAN	10,775	SOUTHERN EUROPE	CHRISTIAN: CATHOLIC	Spanish	0- White	C) White - Any other White background
HI317	HISPANIC	LATIN AMERICA	3,644	AMERICAS	CHRISTIAN: CATHOLIC	Spanish	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
HI318	HISPANIC	PHILIPPINES	1,976	EAST ASIA	CHRISTIAN: CATHOLIC	Filipino	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
HI419	HISPANIC	BASQUE	1,568	SOUTHERN EUROPE	CHRISTIAN: CATHOLIC	Basque	0- White	C) White - Any other White background
HI520	HISPANIC	CATALAN	3,105	SOUTHERN EUROPE	CHRISTIAN: CATHOLIC	Catalan	0- White	C) White - Any other White background
HI621	HISPANIC	GALICIAN	511	SOUTHERN EUROPE	CHRISTIAN: CATHOLIC	Galician	0- White	C) White - Any other White background
IN110	INTERNATIONAL	INTERNATIONAL	15,799	UNCLASSIFIED	Not Applicable	Not Applicable	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
IN211	INTERNATIONAL	UNCLASSIFIED	511		Not Applicable	Not Applicable	9- Unknown	Z) Unknown
JA110	JEWISH AND ARMENIAN	JEWISH AND ARMENIAN	72	DIASPORIC	Not Applicable	Not Applicable	0- White	C) White - Any other White background
JA211	JEWISH AND ARMENIAN	JEWISH	80,522	DIASPORIC	JEWISH	Hebrew	0- White	C) White - Any other White background
JA212	JEWISH AND ARMENIAN	SEPHARDIC JEWISH	821	DIASPORIC	JEWISH	Ladino	0- White	C) White - Any other White background
JA313	JEWISH AND ARMENIAN	ARMENIAN	4,353	CENTRAL ASIA	CHRISTIAN: ORTHODOX_CALCEDONIAN	Armenian	0- White	C) White - Any other White background
JP110	JAPANESE	JAPAN	6,335	EAST ASIA	BHUDDIST	Japanese	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML110	MUSLIM	MUSLIM	103,514	MIDDLE EAST	MUSLIM	Arabic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML211	MUSLIM	MIDDLE EAST	672	MIDDLE EAST	MUSLIM	Arabic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML212	MUSLIM	IRAN	10,312	MIDDLE EAST	MUSLIM	Farsi	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML213	MUSLIM	IRAQ	262	MIDDLE EAST	MUSLIM	Arabic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML214	MUSLIM	JORDAN	55	MIDDLE EAST	MUSLIM	Arabic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML215	MUSLIM	KUWAIT	3	MIDDLE EAST	MUSLIM	Arabic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML216	MUSLIM	LEBANON	3,107	MIDDLE EAST	MUSLIM	Arabic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group

CEL Type Code	CEL Group	CEL Type	People GB 2004	Geographical Area	Religion	Major Language	1991 Census Ethnic Group	2001 Census Ethnic Group
ML217	MUSLIM	OMAN	5	MIDDLE EAST	MUSLIM	Arabic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML218	MUSLIM	SAUDI ARABIA	186	MIDDLE EAST	MUSLIM	Arabic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML219	MUSLIM	SYRIA	142	MIDDLE EAST	MUSLIM	Arabic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML220	MUSLIM	UNITED ARAB EMIRATES	14	MIDDLE EAST	MUSLIM	Arabic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML221	MUSLIM	YEMEN	6	MIDDLE EAST	MUSLIM	Arabic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML322	MUSLIM	KAZAKHSTAN	11	CENTRAL ASIA	MUSLIM	Kazakh	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML323	MUSLIM	KYRGYZSTAN	2	CENTRAL ASIA	MUSLIM	Kirghiz	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML324	MUSLIM	TURKMENISTAN	8	CENTRAL ASIA	MUSLIM	Turkmen	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML325	MUSLIM	UZBEKISTAN	3	CENTRAL ASIA	MUSLIM	Uzbek	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML326	MUSLIM	AFGHANISTAN	3,687	CENTRAL ASIA	MUSLIM	Farsi	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML427	MUSLIM	BANGLADESH	179,401	SOUTH ASIA	MUSLIM	Bengali	6- Bangladeshi	K) Asian or Asian British - Bangladeshi
ML428	MUSLIM	MUSLIM INDIA	25,704	SOUTH ASIA	MUSLIM	Punjabi	4- Indian	H) Asian or Asian British - Indian
ML429	MUSLIM	PAKISTAN	508,699	SOUTH ASIA	MUSLIM	Punjabi	5- Pakistani	J) Asian or Asian British - Pakistani
ML430	MUSLIM	PAKISTANI KASHMIR	91,472	SOUTH ASIA	MUSLIM	Kashmiri	5- Pakistani	J) Asian or Asian British - Pakistani
ML431	MUSLIM	MALAYSIAN MUSLIM	220	EAST ASIA	MUSLIM	Malay	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML532	MUSLIM	ALGERIA	2,585	AFRICA	MUSLIM	Arabic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML533	MUSLIM	EGYPT	479	AFRICA	MUSLIM	Arabic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML534	MUSLIM	TUNISIA	39	AFRICA	MUSLIM	Arabic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML535	MUSLIM	LIBYA	38	AFRICA	MUSLIM	Arabic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML536	MUSLIM	MOROCCO	572	AFRICA	MUSLIM	Arabic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML637	MUSLIM	WEST AFRICAN MUSLIM	2,399	AFRICA	MUSLIM	Arabic	2- Black - African	N) Black or Black British - African
ML638	MUSLIM	SOMALIA	33,260	AFRICA	MUSLIM	Somali	2- Black - African	N) Black or Black British - African
ML639	MUSLIM	SUDAN	468	AFRICA	MUSLIM	Arabic	2- Black - African	N) Black or Black British - African

CEL Type Code	CEL Group	CEL Type	People GB 2004	Geographical Area	Religion	Major Language	1991 Census Ethnic Group	2001 Census Ethnic Group
ML640	MUSLIM	ERITREA	1,397	AFRICA	CHRISTIAN: OTHER	Tigré	2- Black - African	N) Black or Black British - African
ML741	MUSLIM	TURKEY	50,706	MIDDLE EAST	MUSLIM	Turkish	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML742	MUSLIM	TURKISH CYPRUS	1,205	MIDDLE EAST	MUSLIM	Turkish	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
ML743	MUSLIM	BALKAN MUSLIM	10	EASTERN EUROPE	MUSLIM	Bosnian	0- White	C) White - Any other White background
ND110	NORDIC	NORDIC	6,377	NORTHERN EUROPE	CHRISTIAN: PROTESTANT	Not Applicable	0- White	C) White - Any other White background
ND211	NORDIC	DENMARK	20,561	NORTHERN EUROPE	CHRISTIAN: PROTESTANT	Danish	0- White	C) White - Any other White background
ND212	NORDIC	ICELAND	115	NORTHERN EUROPE	CHRISTIAN: PROTESTANT	Icelandic	0- White	C) White - Any other White background
ND213	NORDIC	SWEDEN	19,090	NORTHERN EUROPE	CHRISTIAN: PROTESTANT	Swedish	0- White	C) White - Any other White background
ND214	NORDIC	NORWAY	186,375	NORTHERN EUROPE	CHRISTIAN: PROTESTANT	Norwegian	0- White	C) White - Any other White background
ND315	NORDIC	FINLAND	5,685	NORTHERN EUROPE	CHRISTIAN: PROTESTANT	Finnish	0- White	C) White - Any other White background
SA110	SOUTH ASIAN	SOUTH ASIA	12,699	SOUTH ASIA	BHUDDIST	Hindi	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
SA211	SOUTH ASIAN	INDIA HINDI	319,677	SOUTH ASIA	HINDU	Hindi	4- Indian	H) Asian or Asian British - Indian
SA212	SOUTH ASIAN	INDIA NORTH	75,282	SOUTH ASIA	HINDU	Hindi	4- Indian	H) Asian or Asian British - Indian
SA213	SOUTH ASIAN	INDIA SOUTH	302	SOUTH ASIA	BHUDDIST	Hindi	4- Indian	H) Asian or Asian British - Indian
SA214	SOUTH ASIAN	HINDU NOT INDIAN	22,106	SOUTH ASIA	HINDU	Hindi	4- Indian	H) Asian or Asian British - Indian
SA315	SOUTH ASIAN	SRI LANKA	53,919	SOUTH ASIA	BHUDDIST	Sinhala	4- Indian	L) Asian or Asian British - Any other Asian background
SA316	SOUTH ASIAN	BANGLADESH HINDU	2,974	SOUTH ASIA	HINDU	Bengali	6- Bangladeshi	K) Asian or Asian British - Bangladeshi
SA317	SOUTH ASIAN	BHUTAN	3	SOUTH ASIA	BHUDDIST	Dzongkha	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
SA318	SOUTH ASIAN	NEPAL	150	SOUTH ASIA	HINDU	Nepali	8- Any other ethnic group	L) Asian or Asian British - Any other Asian background
SA419	SOUTH ASIAN	MAURITIUS	1,765	SOUTH ASIA	HINDU	Morisyen	4- Indian	L) Asian or Asian British - Any other Asian background
SA420	SOUTH ASIAN	SEYCHELLES	71	SOUTH ASIA	CHRISTIAN: CATHOLIC	Seselwa Creole French	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group
SA421	SOUTH ASIAN	KENYAN ASIAN	1,121	AFRICA	HINDU	Hindi	4- Indian	L) Asian or Asian British - Any other Asian background

CEL Type Code	CEL Group	CEL Type	People GB 2004	Geographical Area	Religion	Major Language	1991 Census Ethnic Group	2001 Census Ethnic Group
SA522	SOUTH ASIAN	ASIAN CARIBBEAN	581	AMERICAS	HINDU	Hindi	4- Indian	L) Asian or Asian British - Any other Asian background
SA523	SOUTH ASIAN	GUYANA	911	AMERICAS	HINDU	Hindi	4- Indian	L) Asian or Asian British - Any other Asian background
SK110	SIKH	INDIA SIKH	283,657	SOUTH ASIA	SIKH	Punjabi	4- Indian	H) Asian or Asian British - Indian
ZU110	UNCLASSIFIED	UNCLASSIFIED	21,826	UNCLASSIFIED	Not Applicable	Not Applicable	9- Unknown	Z) Unknown
ZZ110	VOID	VOID	12,118	UNCLASSIFIED	Not Applicable	Not Applicable	9- Unknown	Z) Unknown
ZZ211	VOID	VOID - SURNAME	819	UNCLASSIFIED	Not Applicable	Not Applicable	9- Unknown	Z) Unknown
ZZ212	VOID	VOID INITIAL	94,621	UNCLASSIFIED	Not Applicable	Not Applicable	9- Unknown	Z) Unknown
ZZ213	VOID	VOID OTHER	56	UNCLASSIFIED	Not Applicable	Not Applicable	9- Unknown	Z) Unknown
ZZ214	VOID	VOID PERSONAL NAME	5,464	UNCLASSIFIED	Not Applicable	Not Applicable	9- Unknown	Z) Unknown
ZZ215	VOID	VOID TITLE	1,858	UNCLASSIFIED	Not Applicable	Not Applicable	9- Unknown	Z) Unknown
ZZ316	VOID	NOT_FOUND	110,049		Not Applicable	Not Applicable	9- Unknown	Z) Unknown

Table 7: Lookup table between CEL Types and various onomastic groups (CEL Groups), languages, religions, geographical regions, and Census categories

9 References

Abbotts J, Williams R, Smith GD. 1999. Association of Medical, Physiological, Behavioural and Socio-Economic Factors with Elevated Mortality in Men of Irish Heritage in West Scotland. *Journal Of Public Health Medicine* 21(1): 46-54

American Council of Learned Societies. 1932. *Report of Committee on Linguistic and National Stocks in the Population of the United States.* Annual Report for the Year 1931. American Historical Association. Washington, D.C.

APHO. 2005. *Ethnicity and Health.* Indications of public health in the English Regions. Rep. 4, Association of Public Health Observatories (APHO).

Aspinall PJ. 2000. The New 2001 Census Question Set on Cultural Characteristics: Is It Useful for the Monitoring of the Health Status of People from Ethnic Groups in Britain? *Ethnicity and Health* 5(1): 33 - 40

Association of Public Health Observatories. 2005. *Ethnicity and Health.* Indications of public health in the English Regions. Rep. 4, (APHO).

Bhopal R. 2004. Glossary of Terms Relating to Ethnicity and Race: For Reflection and Debate. *Journal Of Epidemiology And Community Health* 58(6): 441-445

Bhopal R, Fischbacher C, Steiner M, Chalmers J, Povey C, et al. 2004. *Ethnicity and Health in Scotland: Can We Fill the Information Gap?*, Centre for Public Health and Primary Care Research. University of Edinburg. Available at: http://www.chs.med.ed.ac.uk/phs/research/Retrocoding%20final%20report.pdf. Accessed: 22/11/2005.

Birkin M and Clarke G. 1998. Gis, Geodemographics, and Spatial Modeling in the U.K. Financial Service Industry. *Journal of Housing Research* 9(1): 87-111

Boyle P. 2004. Population Geography: Migration and Inequalities in Mortality and Morbidity. *Progress in Human Geography* 28(6): 767-776

Brubaker R. 2004. *Ethnicity without Groups*. London: Harvard University Press.

Buechley RW. 1967. Characteristic Name Sets of Spanish Populations. *Names* 15: 53-69

Bulmer M. 1996. The Ethnic Group Question in the 1991 Census of Population. In *Ethnicity in the 1991 Census. Volume 1. Demographic Characterisitics of the Ethnic Minority Populations*, Coleman D, Salt J (eds.), Office for National Statistics, HMSO: London: xi -xxix.

Chaudhry S, Fink A, Gelberg L, Brook R. 2003. Utilization of Papanicolaou Smears by South Asian Women Living in the United States. *Journal of General Internal Medicine* 18(5): 377-384

Clark WAV and Morrison PA. 1995. Demographic Foundations of Political Empowerment in Multiminority Cities. *Demography* 32(2): 183-201

Coleman D. 2006. Immigration and Ethnic Change in Low-Fertility Countries: A Third Demographic Transition. *Population and Development Review* 32(3): 401-446

Coleman D and Salt J, eds. 1996. *Ethnicity in the 1991 Census. Volume 1. Demographic Characterisitics of the Ethnic Minority Populations*. Office for National Statistics, HMSO London

Cook D, Hewitt D, Milner J. 1972. Uses of the Surname in Epidemiologic Research. *American Journal of Epidemiology* 95: 38-45

Danmarks Statistik. 2006. *Mest Populære for- Og Efternavne for Alle Danskere.* Available at: http://www.dst.dk/Statistik/Navne/pop.aspx. Accessed: 23/07/2006.

Dorling D and Rees P. 2003. A Nation Still Dividing: The British Census and Social Polarisation 1971 - 2001. *Environment And Planning A* 35(7): 1287-1313

Edina. 2006. *UK Placename Gazetter.* UK Borders. Available at: http://edina.ac.uk/ukborders/. Accessed: 20/03/2006.

Gerrish K. 2000. Researching Ethnic Diversity in the British Nhs: Methodological and Practical Concerns. *Journal of Advanced Nursing* 31: 918-925

Gesellschaft für deutsche Sprache. 2006. *Beliebteste Vornamen.* Available at: http://www.gfds.de/index.php?id=63. Accessed: 23/10/2006.

Gill P, Bhopal R, Wild S, Kai J. 2005. Limitations and Potential of Country of Birth as Proxy for Ethnic Group. *British Medical Journal* 330(7484): 196

Graham D and Waterman S. 2005. Underenumeration of the Jewish Population in the UK 2001 Census. *Population, Space and Place* 11(2): 89-102

Hage BH, Oliver RG, Powles JW, Wahlqvist ML. 1990. Telephone Directory Listings of Presumptive Chinese Surnames: An Appropriate Sampling Frame for a Dispersed Population with Characteristic Surnames. *Epidemiology* 1(5): 405-408

Hanks P. 2003. *Dictionary of American Family Names* New York: Oxford University Press.

Hanks P and Tucker DK. 2000. A Diagnostic Database of American Personal Names. *Names* 48(1): 59-69

Harding S, Dews H, Simpson S. 1999. The Potential to Identify South Asians Using a Computerised Algorithm to Classify Names. *Population Trends* 97: 46-50

Harris R, Sleight P, Webber R. 2005. *Geodemographics: Neighbourhood Targeting and Gis*. Chichester, UK: John Wiley and Sons.

Haut Conseil à l'Integration. 1991. *Pour Un Modèle Francais D'integration.* Premier Rapport Annuel. La Documentation Francaise. Paris.

Himmelfarb HS, Loar RM, Mott SH. 1983. Sampling by Ethnic Surnames: The Case of American Jews. *Public Opinion Quarterly* 47: 247-260

Hinton L, Jenkins CN, McPhee S, Wong C, Lai KQ, et al. 1998. A Survey of Depressive Symptoms among Vietnamese-American Men in Three Locales: Prevalence and Correlates. *The Journal of Nervous and Mental Disease* 186(11): 677-683

Hofstetter CR, Hovell MF, Lee J, Zakarian J, Park H, et al. 2004. Tobacco Use and Acculturation among Californians of Korean Descent: A Behavioral Epidemiological Analysis. *Nicotine and Tobacco Research* 6(3): 481-489

IDESCAT. 2006. Els Noms De La Població De Catalunya Per Nacionalitats. Institut d' Estadistica de Catalunya

Instituto de Estadística de la Comunidad de Madrid. 2006. Guía De Nombres Y Primer Apellido De Los Residentes En La Comunidad De Madrid 1998-2005

Johnston R, Wilson D, Burgess S. 2004. School Segregation in Multiethnic England. *Ethnicities* 4(2): 237-265

Kertzer DI and Arel D. 2002. *Census and Identity. The Politics of Race, Ethnicity, and Language in National Censuses*. Cambridge: Cambridge University Press.

Kitano HH, Lubben JE, Chi I. 1988. Predicting Japanese American Drinking Behavior. *The International Journal of the Addictions* 23(4): 417-428

Lai DW. 2004. Impact of Culture on Depressive Symptoms of Elderly Chinese Immigrants. *Canadian Journal of Psychiatry* 49(12): 820-827

Large P and Ghosh K. 2006. A Methodology for Estimating the Population by Ethnic Group for Areas within England. *Population Trends* 123: 21-31

Lauderdale D and Kestenbaum B. 2000. Asian American Ethnic Identification by Surname. *Population Research and Policy Review* 19(3): 283-300

Leaman AM, Rysdale E, Webber R. 2006. Use of the Emergency Department by Polish Migrant Workers. *Emerg Med J* 23(12): 918-919

Leppard D. 2005. Race Chief Warns of Ghetto Crisis. In *The Sunday Times*

Levitt SD and Dubner SJ. 2005. *Freakonomics : A Rogue Economist Explores the Hidden Side of Everything*. New York: HarperCollins.

London Health Observatory. 2003. *Missing Record: The Case for Recording Ethnicity at Birth and Death Registration.* LHO Reports. Available at: http://www.lho.org.uk/viewResource.aspx?id=7954. Accessed: 01/09/2006.

London Health Observatory. 2005. *Using Routine Data to Measure Ethnic Differentials in Access to Revascularisation in London.* Available at: http://www.lho.org.uk/viewResource.aspx?id=9732. Accessed: 20/07/2006.

Manni F, Toupance B, Sabbagh A, Heyer E. 2005. New Method for Surname Studies of Ancient Patrilineal Population Structures, and Possible Application to Improvement of Y-Chromosome Sampling. *Am J Phys Anthropol* 126(2): 214-228

Marmot M, Adelstein A, Bulusu L. 1984. *Immigrant Mortality in England and Wales 1970-78: Causes of Death by Country of Birth.* Opcs. Her Majesty's Stationery Office. London.

Mason D. 2003. *Explaining Ethnic Differences: Changing Patterns of Disadvantage in Britain*. Bristol: Policy Press.

Mateos P. 2007a. A Review of Name-Based Ethnicity Classification Methods and Their Potential in Population Studies. *Population Space and Place* 13: (in press)

Mateos P. 2007b. Segregación Residencial De Minorías Étnicas Y El Análisis Geográfico Del Origen De Nombres Y Apellidos [Residential Segregation of Ethnic Minorities and Geographic Analysis of Name Origins]. *Cuadernos Geograficos* 40: (in press)

McAuley J, De Souza L, Sharma V, Robinson I, Main CJ, Frank AO. 1996. Self Defined Ethnicity Is Unhelpful. *British Medical Journal* 313(7054): 425b-426

Nanchahal K, Mangtani P, Alston M, dos Santos Silva I. 2001. Development and Validation of a Computerized South Asian Names and Group Recognition Algorithm (Sangra) for Use in British Health-Related Studies. *Journal Of Public Health Medicine* 23(4): 278-285

National Geospatial Agency. 2006. *Geonet Names Server.* Available at: http://earth-info.nga.mil/gns/html/index.html. Accessed: 13/12/2005.

NHS Executive. 1994. *Collection of Ethnic Group Data for Admitted Patients.* EL/94/77. Nhse. Leeds.

Office for National Satistics. 2006. *National Statistics Postcode Directory (Nspd) User Guide.* London. Available at: http://www.statistics.gov.uk/geography/downloads/NSPDUserGuide.pdf. Accessed: 23/11/2006.

Olson S. 2002. *Mapping Human History: Genes, Race, and Our Common Origins*. New York: First Mariner Books.

ONS. 2003. *Ethnic Group Statistics: A Guide for the Collection and Classification of Data.* Statistics Ofn. ONS. London. Available at: http://www.statistics.gov.uk/about/ethnic_group_statistics/downloads/ethnic_group_statistics.pdf

Openshaw S. 1984. *The Modifiable Areal Unit Problem*. Norwich: Geo Books.

Parsons C, Godfrey R, Annan G, Cornwall J, Dussart M, et al. 2004. *Minority Ethnic Exclusions and the Race Relations (Amendment) Act 2000. Research Report Rr616.* Hmso Dfeas. London. Available at: http://www.dfes.gov.uk/exclusions/uploads/RR616.pdf. Accessed: 28/12/2005.

Peach C and Owen D. 2004. *Social Geography of British South Asian Muslim, Sikh and Hindu Sub-Communities.* ESRC End of Project Full Report R-000239765. Available at: http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/ (search for "R-000239765"). Accessed: 15/08/2006.

Rahman MM, Luong NT, Divan HA, Jesser C, Golz SD, et al. 2005. Prevalence and Predictors of Smoking Behavior among Vietnamese Men Living in California. *Nicotine and Tobacco Research* 7(1): 103-109

Rissel C, Ward JE, Jorm L. 1999. Estimates of Smoking and Related Behaviour in an Immigrant Lebanese Community: Does Survey Method Matter? *Australian and New Zealand Journal of Public Health* 23(5): 534-537

Robinson GM. 1998 *Methods and Techniques in Human Geography*. Chichester: John Wiley and Sons.

Schürer KE. 2004. Surnames and the Search for Regions. *Local Population Studies* 72

Silva MJ, Martins B, Chaves M, Afonso AP, Cardoso N. 2006. Adding Geographic Scopes to Web Resources. *Computers, Environment and Urban Systems* 30(4): 378-399

Simpson L. 2004. Statistics of Racial Segregation: Measures, Evidence and Policy. *Urban Studies* 41: 661-681

Skerry P. 2000. *Counting on the Census? Race, Group Identity, and the Evasion of Politics*. Washington: Brookings Institution Press.

Sparks E. 2005. Experian Consumer Dynamics Characteristics.

Statbel. 2006. *Noms De Famille Les Plus Fréquents - Belgique Et Régions.* Available at: http://statbel.fgov.be/figures/d21a_fr.asp. Accessed: 12/09/2006.

Statistics Iceland. 2006. *Forenames in the National Register of Persons* Available at: http://www.statice.is/?PageID=846. Accessed: 13/08/2006.

Stillwell J and Duke-Williams O. 2005. Ethnic Population Distribution, Immigration and Internal Migration in Britain. What Evidence of Linkage at the District Scale. Presented at *British Society for Population Studies Annual Conference*, University of Kent at Canterbury 12-14 September. Available at: http://www.lse.ac.uk/collections/BSPS/pdfs/Stillwell_ethnicpopdist_2005.pdf Accessed: 20/06/2006.

Surname Profiler. 2006.

The Economist. 2005. One Man's Ghetto. In *The Economist*, pp. 16

The Economist. 2006. Hostility at Home.

The Guardian. 2004. Gateway to the Past.

Tu SP, Yasui Y, Kuniyuki A, Schwartz SM, Jackson JC, Taylor VM. 2002. Breast Cancer Screening: Stages of Adoption among Cambodian American Women. *Cancer Detection and Prevention* 26(1): 33-41

Tucker DK. 2003. Surnames, Forenames and Correlations. In *Dictionary of American Family Names* Hanks P (eds.), Oxford University Press: New York: xxiii-xxvii.

Tucker DK. 2005. The Cultural-Ethnic-Language Group Technique as Used in the Dictionary of American Family Names (Dafn). *Onomastica Canadiana* 87(2): 71-84

Tucker DK. 2006. Surname Distribution Prints from the UK 1998 Electoral Roll Compared with Those from Other Distributions *Nomina* in press

US Senate. 1928. *Immigration Quotas on the Basis of National Origin.* Rep. Miscellaneous Documents 8870 vol.1 nr 65, 70th Congress 1st Session. Washington, DC.

Van Ryn M and Fu SS. 2003. Paved with Good Intentions: Do Public Health and Human Service Providers Contribute to Racial/Ethnic Disparities in Health? *American Journal of Public Health* 93(2): 248-255

Wild S and McKeigue P. 1997. Cross Sectional Analysis of Mortality by Country of Birth in England and Wales, 1970-92. *British Medical Journal* 314(7082): 705-

Word DL and Perkins RC. 1996. *Building a Spanish Surname List for the 1990s a New Approach to an Old Problem.* Technical Working Paper 13. US Census Bureau, Population Division. Washington DC. Available at: http://www.census.gov/population/documentation/twpno13.pdf. Accessed: 29/05/2005.