*Paper to be presented at the Social Statistics and Ethnic Diversity conference, Montreal 6-8 December 2007. This is a draft version, please do not cite. Contact the author if you are interested in an updated version of the paper*

# Classifying ethnicity using people's names

Pablo Mateos

University College London, Department of Geography
Gower Street, London, WC1E 6BT, England
p.mateos@ucl.ac.uk

## 1 Introduction

There has been a surge in population studies research on ethnic minorities over the last decade, partly arising from an increase in the availability and comparability of population statistics by ethnic group, especially following the round of censuses at the turn of the Millennium. Although these efforts have helped to broaden our understandings of today's increasingly multicultural societies, use of such data implies a grasp of far-reaching ontological and epistemological issues.

Delineating the ethnicity of a population sub-group is difficult in both conceptual and measurement terms, because ethnicity remains one of the most contested and unstable research concepts in science. Ethnicity is socially constructed and relates to several dimensions of a person's identity, and as such is inherently contextual and transient. Furthermore, the purposes for which ethnicity data are collected, their quality and availability are very varied and are usually based on differently preconceived ontologies of ethnicity. Common consequences of these issues are a lack of consistency between different studies, incompleteness of coverage, and impediments to longitudinal analysis. More fundamentally still, the various socioeconomic, demographic, and cultural correlates of ethnicity cannot be identified accurately.

This paper proposes an alternative approach to researching the ontology of ethnicity based on the origin of surnames and forenames. It is argued that this better reflects the complex dimensions of ethnicity and can be more flexibly adapted to the particularity of each study, and as such that it presents a way of better managing the effect of some of the problems associated with self-assigned ethnicity, such us the volatility of the ontologies and categories of ethnicity, and its lack of availability and comparability across datasets.

The alternative methodology developed to ascribe population ethnicity, uses a tool to assign an individual's forename and surname to one of 185 Cultural Ethnic and Linguistic groups (CEL), which are weighted according to name scores, in order to assign the most probable CEL allocated to each individual at very fine geographical levels. This classification can be continuously updated using a basic name and address register.

Such method has been applied to the UK Electoral Roll and European telephone directories as well as to several health registers in London. The accuracy of the method has been evaluated using separate datasets where the self-reported ethnicity of individuals has been recorded.

The outcome of the research is an improved methodology for classifying population registers, as well as small areas (typically postcodes), into cultural, ethnic and linguistic groups, that makes possible the creation of much more detailed, frequently updated, representations of the multicultural geographies of contemporary cities.

# 2 The need for ethnicity classifications

## 2.1 Background

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.

## 2.2 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 (2007). 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).

## 3 The CEL Taxonomy

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.

## 3.1 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 (contact the author for a full list with the correspondence between CELs and the different aggregations proposed).

| 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**

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 Mateos et al (2007) 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 the Appendix of Mateos et al (2007).

# 4 Preliminary Steps to Creating a Name to Ethnicity Classification

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.

## 4.1 Reference and Target Populations

The literature review conducted by Mateos (2007) 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 (2007), 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.

## 4.2 Steps to build a forenames seed list

The seed list of non-British forenames was built using the information already gathered through the heuristic approach, as justified above. The objective of this phase was to extract the most representative non-British forenames, re-classify them using a the coarser taxonomy of CEL Subgroups, and rate each forename-to-CEL assignment with a probability rate optimised for the Forename-Surname Clustering (FSC) technique (in Mateos et al (2007) FSC was referred to as *triage* technique).

This was achieved through the series of steps summarised below, as a sequence of the ten steps taken to obtain a forename seed list, which is itself the main ingredient in the Forename-Surname Clustering (FSC) carried out in the following section.

1) *Compiling a forename type frequency list*. Using the GB04 register, a list of the frequencies of British forename types was produced, initially including 437,639 forename types. However, of that figure 280,214 forename types occur only once while the remaining 157,424 forename types have a frequency greater than one.
2) *Selecting the most frequent non-British forename types*. The above list was reduced by subtracting the 22,078 British forename types compiled in Mateos et al (2007), as well as any other forename type with a frequency (tokens) of 10 or lower. After this filtering, the size of the seed list at this stage is 24,200 forename types.
3) *Removing highly popular forenames*. A further group of forenames was removed at the other end of the frequency distribution, this time the most frequent, in order to reduce the risk of misclassifications by highly popular forenames usually found in a number of CELs, such as Maria, Ana, Natasha, Mohammed, Ahmed, etc. After a few tests of cross-occurrences of these highly popular forenames in the database, a threshold was adopted to exclude those forename types with a frequency of 4,000 tokens or more, which removed 273 forename types from the list. At this stage the size of the seed list was 23,927 forename types.
4) *Removing short character forenames*. Another potential source of misclassification experienced in the heuristic approach was derived from short forenames or surnames,

such as Lee, Jay, Bob, Van, Isa, Che, etc that can be assigned to different CELs. Therefore, forenames with a character length of three or less were removed from the seed list, including a total of 1,194 forename types of which a high proportion were also initials or honorifics (eg. Mr.). The forename seed list at stage had 22,733 types.

5) *Forename-Surname-CEL linkage*. The interim forename seed list was linked to the complete 2004 Electoral Register (GB04), through the forename of each individual. Those same individuals were further linked to the surname-to-CEL table classified in the heuristic approach through the individual's surname. Therefore, at this stage the linkage of the three tables had the following schema:

   *(A) non-British forename list* => *(B) GB04* Electoral Register <= *(C) Surname-to-CEL*

   This is read as three tables labelled (A), (B) and (C) linked by 'one-to-many' database relationships of the type 'left join' ( =>) and 'right join' (<=)

6) *Computation of CEL percentages by forename*. For each forename in table (A) above was calculated the count of people (tokens) in table (B) that had a surname associated with a particular CEL Subgroup in table (C). British CEL Subgroups were ignored for the purpose of this calculation. This resulted in a summary table, as per the example given in Table 3, in which the rows were any combination of a forename type with a CEL Subgroup, reporting the count of surname tokens and the percentage of the total forename tokens. This table will be referred here as table (D).

| Forename | CEL Subgroup | Surname tokens | % of total tokens |
|---|---|---:|---:|
| AAMIR | BANGLADESHI | 7 | 2.8% |
| AAMIR | HINDI INDIAN | 5 | 2.0% |
| AAMIR | INDIA NORTH | 1 | 0.4% |
| AAMIR | MUSLIM MIDDLE EAST | 6 | 2.4% |
| AAMIR | MUSLIM NORTHAFRICAN | 1 | 0.4% |
| AAMIR | MUSLIM SOUTH ASIAN | 6 | 2.4% |
| AAMIR | PAKISTANI | 198 | 79.8% |
| AAMIR | PAKISTANI KASHMIR | 19 | 7.7% |
| AAMIR | SIKH | 3 | 1.2% |
| AAMIR | SPANISH | 1 | 0.4% |
| AAMIR | VOID | 1 | 0.4% |
|  | **TOTAL** | **248** | **100.0%** |

**Table 3: Example of calculation of CEL percentage per forename (excluding British CEL Subgroups)**

7) *Selection of the CEL Subgroup with the highest percentage*. For each forename in table (D) above, the CEL Subgroup with the highest percentage of surname tokens was selected as the most representative CEL of that forename. In the example given in Table 6.2 this resulted in the classification of the forename ‘Aamir’ in the ‘Pakistani’ CEL Subgroup with 79.8% of the surname tokens for that forename type. In instances where the highest percentage was shared by more than one CEL Subgroup, it was decided to eliminate the forename from the forename seed list (1,794 forename types), since this situation would lead into potential further misclassifications when using the FSC technique. This resulted in a forename seed list size of 20,939 forename types, with their assigned CEL Subgroup and the corresponding highest percentage. This list is termed here table (E), an example of which is given in Table 4.

| Forename | CEL Subgroup | % of total surname tokens within the selected CEL Subgroup |
|---|---|---|
| AAMIR | PAKISTANI | 79.8% |

**Table 4: Example of the final selected CEL Subgroup for a forename and percentage of surname tokens**

At the end of these seven steps a new forename seed list was available including 20,939 forename types. However, even when the raw percentages of surname tokens within the selected CEL Subgroup associated with each forename were a good indicator of how well the forename represents its allocated CEL Subgroup (literally the percentage of surname tokens that are also from the same CEL Subgroup), they could not be compared on equal terms across CEL Subgroups. This is because in some CEL Subgroups which are more integrated into the host society or whose forenames overlap with another CEL Subgroup, a low value of the percentage for a forename, for example 34%, might nevertheless present a strong indicator that the forename belongs to that CEL Subgroup. On the contrary, in more isolated groups, such as the Japanese CEL, a higher value, for example 50%, might mean a low association with the CEL Subgroup. Therefore, these percentages needed to be standardised into a common scale that takes into account the context of the CEL Subgroup’s percentage values distribution, in order to facilitate direct comparison of values across CEL Subgroups.

8) *Standardization of percentage values.* Several methods of standardisation were tested, and z-scores were finally selected since they were the most commonly used and were appropriate for this context. z-scores measure how many standard deviations an observed value is away from the mean of a full range of values, giving a positive figure if it is above the mean and a negative below it (Robinson, 1998). Calculation of z-scores is based upon the mean and the standard deviation of the full range of values. In this calculation the range of values is given by the distribution of percentages for each forename within a CEL Subgroup that appear in table (E) above. The calculation of those percentages was explained in the previous step an example of which appears in Table 4 (79.8% for the forename 'AAMIR'). The z-score is calculated individually for each forename in Table (E), as per the following calculation:

$$z = \frac{X - \mu}{\sigma}$$

Where $z$ is the z-score, $X$ is the percentage associated with a forename type (in the example of the forename 'AAMIR' given above this percentage is 79.8% of surnames associated with the Pakistani CEL Subgroup); $\mu$ and $\sigma$ are respectively the mean and the standard deviation of the distribution of percentages within the same CEL Subgroup ('Pakistani' in the example of 'AAMIR' above). This is calculated for all forenames relative to their allocated CEL Subgroup.

The resulting z-scores for each CEL Subgroup are normally distributed about the mean value of zero with a range determined by the maximum number of standard deviations recorded by the most extreme values. When all the z-scores for all the 20,939 forename types – each forename being z-standarised within their CEL Subgroup- are aggregated, the distribution of the combined z-scores is not normally distributed, although it is near-normal, with extreme values of -3.28 to 6.86 and a mean of zero. The histogram of the frequency distributions of the z-scores for the 20,939 forename types is shown in Figure 1, showing the advantage of the standardisation between CEL Subgroups.

If only surname tokens were taken into account to ascribe a forename to a CEL, one surname type with five tokens, for example, would have the same weight as five surname types with one token each all associated with the same CEL, when it is intuitively known that the latter shows a stronger correlation with a CEL than the former. Therefore, a surname token–only approach to build the automated CEL classification is discouraged, because of its high sensitivity to potential incorrect CEL allocations of surname types with large numbers of tokens.

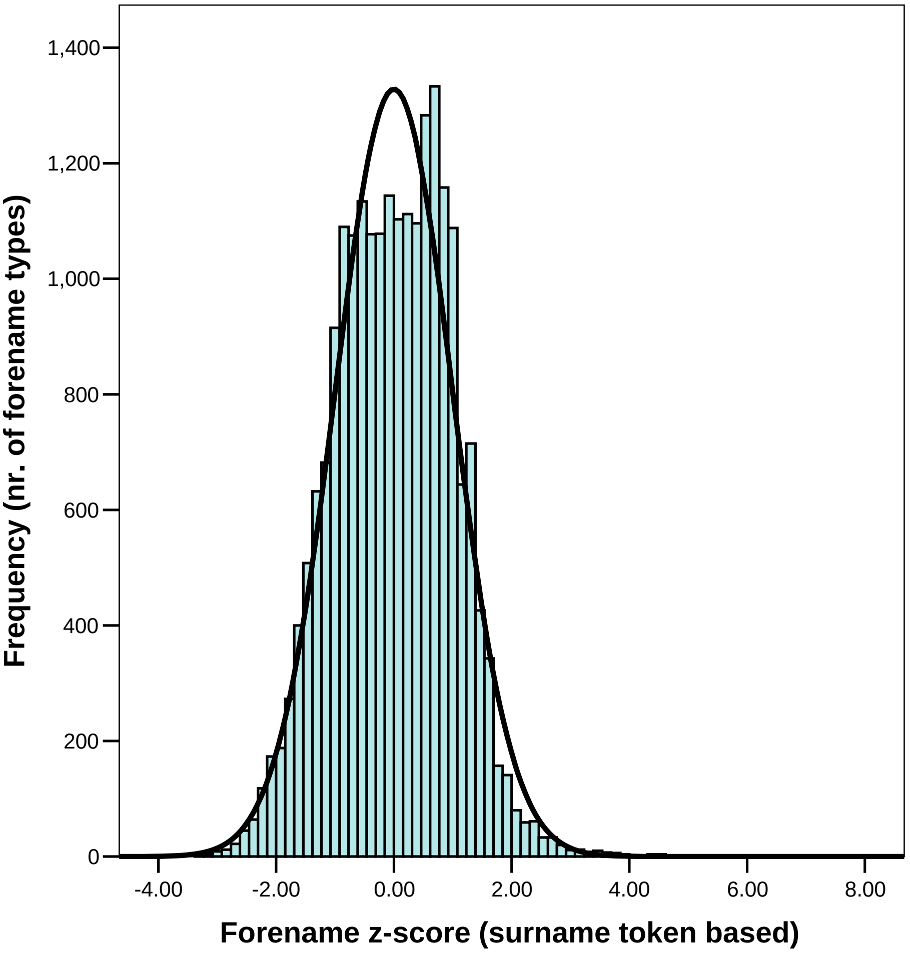

**Figure 1: Histogram of forename z-scores distribution (based on surname tokens)**

This histogram shows the forename type frequency distribution of the z-score value for each forename type (n=20,939). Those z-scores standardised by CEL Subgroup the percentage of surname tokens per forename type (see steps 6, 7 and 8 in the text and Table 4). The Normal curve is also shown for comparison with the histogram which shows a slight negative skewness.

Even so, surname tokens cannot be discarded altogether. It could be argued that the objects being classified are populations of individuals, and thus the CEL system is classifying people or tokens of names. Furthermore, if a surname type-only approach is followed, the number of objects to cluster in FSC would be significantly reduced, removing its classificatory power. For example, in some instances for one forename there would just be three surname types with no weight information and three CEL Subgroups to choose from. But if it is known that one of the surname types has ten times as many tokens as the other two the decision on the

CEL Subgroup is much clearer. Therefore, a mixed approach using both surname tokens and types is advised and is used here.

9) *Average of surname token and type-derived z-scores*. Steps 6, 7 and 8 were repeated once again, but this time taking into account the number of surname types associated with each forename, rather than surname tokens, previously described in steps 6, 7 and 8 above. At the end of this second round, a second z-score value based on percentages of surname types (z_typ) was computed. Therefore two z-scores were obtained for each of the 20,939 forename types, one calculated using the percentage of surname tokens (z_tok) and another one using surname types (z_typ). Finally, a mixed approach was taken, by taking the average of these two z-scores (z_tok and z_typ), since the interest here was in deriving a synthetic indicator of how well a forename represents a CEL Subgroup and there is not a good reason to give one more weight over the other.

$$Avg(z\_score) = \frac{\mathrm{z\_tok} + \mathrm{z\_type}}{2}$$

The advantage of using the average of the z-scores is that it smoothes out any major bias introduced by large or rare surnames used in the calculation, as clearly suggested by Figure 2. The graph shows the z-score values of each of the 20,939 forenames in the seed list, calculated for surname types (z_typ) and surname tokens (z_tok) with their arithmetic average superimposed (avg(z-scores)), ordered by the latter along the x-axis.

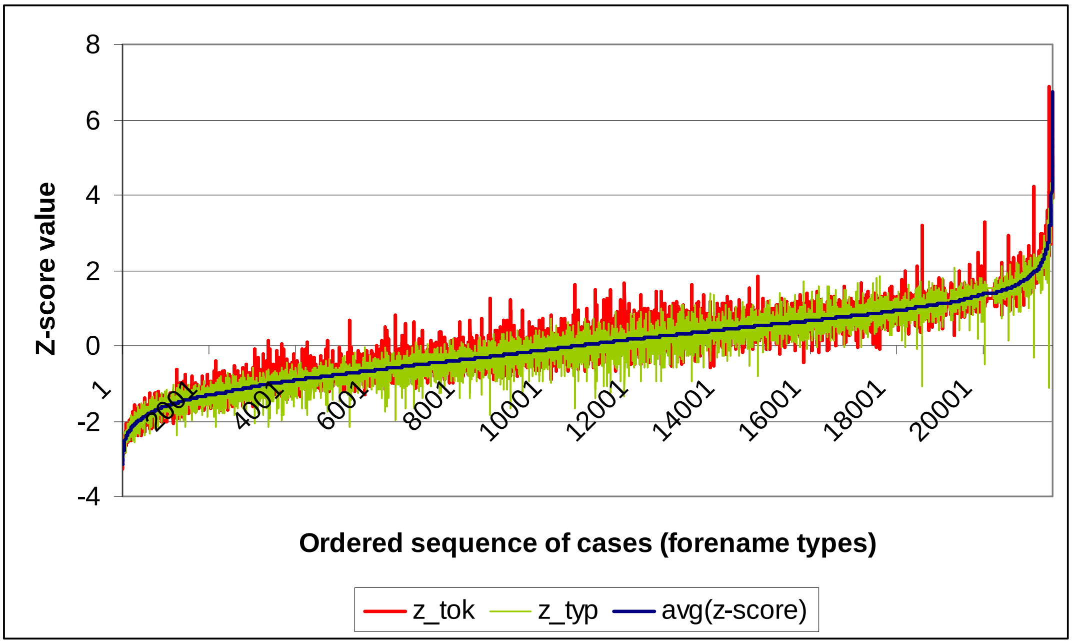

**Figure 2: Distribution of forename z-score values based on surname tokens, types and their average**

This graph shows the z-score values on the y-axis and the sequence of each of the 20,939 forename types on the x-axis ordered by ascending average z-score. The three lines show; z-score values based on surname tokens ('z_tok' in red), based on surname types ('z_typ' in green), and the average between the two ('avg (z-score)' in dark blue).

10) *Selection of higher z-scores and transformation to a final scale.* Finally, after an exploratory analysis of the distribution of z-scores obtained for all seed forenames, it became obvious that the forenames with the lowest z-scores were not at all representative of any CEL Subgroup. These were forenames that had a low percentage across all CEL Subgroups in step 6 above, of which the highest was picked up in step 7, but not necessarily meaning that the forename represented that CEL Subgroup. This was common amongst the rarer forenames. A visual evaluation of some CEL Subgroups by expert collaborators familiar with the names in those CELs, suggested that the forenames with z-score values below -1 were either wrongly assigned to one of the CEL Subgroups or that they were bad indicators of such Subgroups. This made sense because they were more than one standard deviation below the average percentage of the CEL Subgroup. Therefore it was decided to eliminate all forenames with z-scores below -1 from the seed list. This further removed 2,814 forename types leaving the final list size in 18,125 forename types.

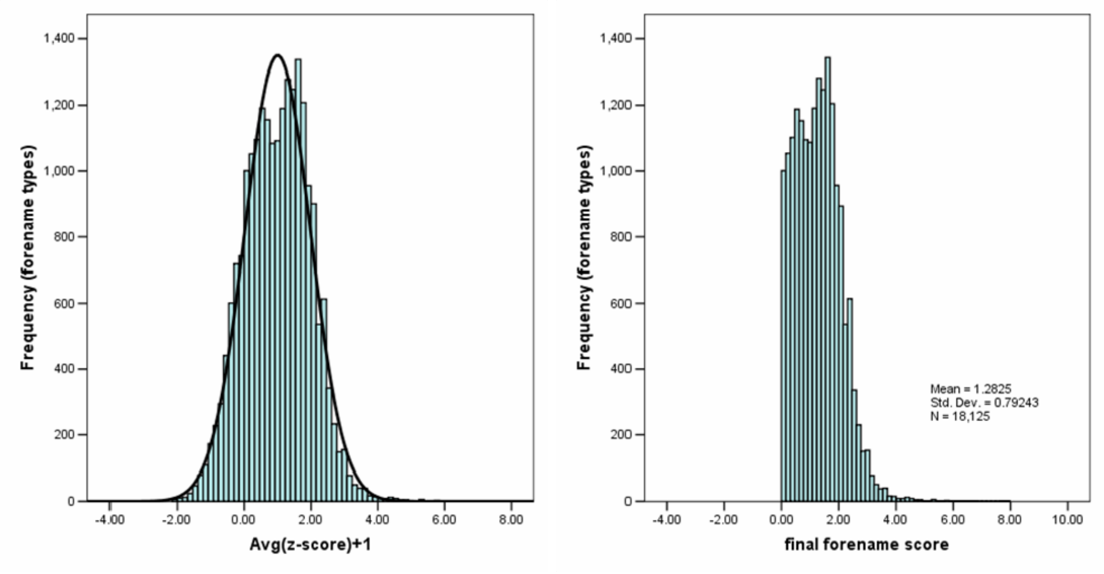

**Figure 3: Histograms of average z-score+1 (left) and truncated final forename score (right)**

Theses two histograms show the effect of the transformation introduced in step 10. In the histogram on the left, the effect of adding a value of one in the whole distribution being shifted to the right can be clearly seen (the mean value now becomes 1, while z-scores have always a mean of 0). The histogram on the right, built to match the same scale and bin size as the previous one, shows the effect of truncating any negative values, resulting in 18,125 positive scores.

> In order to make all the scores positive, and since the distribution was now truncated in the negative values side, and thus ranging from -1 to 6.86, a very simple transformation was applied by adding a value of 1 to all z-scores, resulting in a positive scale starting at 0 and in this case a maximum value of 7.86. This final transformed and truncated z-score will be termed hereinafter the *forename score*. Figure 3 shows this process in the reverse order, so that the effect of adding a value of one in the whole distribution being shifted to the right can be clearly seen first (left histogram, with mean of 1) and then the effect of truncating any negative values after that transformation (right histogram).

Through the ten steps process described here, a final forename seed list was produced, comprised of 18,125 non-British forename types classified by CEL Subgroup and each with an assigned score. A total of 62 CEL Subgroups were represented in the list. A summary of its contents is provided in Table 5, where the number of forename types, the average forename score and standard deviation is shown for each of the CEL Subgroups arranged by CEL Group.

These same ten steps were then repeated for the list of 'British Isles' CEL Subgroups forenames (English, Cornish, Welsh, Scottish and Irish), initially removed in step 2 above as to facilitate the FSC technique for non-British forenames. This second run of the ten steps produced an additional list of 23,419 British and Irish forename types, which after appending to the non-British CEL seed list, produced a final forename seed list of 41,544 forename types. This list will be hereinafter called the ‘seed list’, and it had three fields; ‘forename’, ‘CEL Subgroup’, and ‘score’.

Attempts were made to externally validate this seed list with onomastic experts who could judge its completeness and accuracy, but they were not successful. Automatic validation using the 80,000 diagnostic forenames list from the DAFN seemed to be the best way of achieving this, but because of copyright issues with Oxford University Press, the publisher of the dictionary, this was not possible. The evaluation of the seed list will become part of the evaluation of the whole methodology described in the next chapter.

# 5 Developing an Automatic Classification of Names

The process of building an automated classification of names in CELs had two phases. The first phase entailed building a forenames seed list, which was used in a second phase to classify the surnames and a larger number of forenames into CELs. The previous section dealt with the first phase, while this section will describe the processes involved in the second phase. This second phase started with a forename seed list that served as the main input to ‘fuel’ the FSC triangulation engine described in this section. The triangulation was performed in a series of repetitive cycles, of which only the first two are described in this section.

| CEL Group | CEL Subgroup | Forname types | Avg of final score | Std. dev of final score |
|---|---|---|---|---|
| AFRICAN | AFRICAN | 19 | 1.08 | 0.87 |
| AFRICAN | BLACK SOUTHERN | 45 | 1.31 | 0.79 |
| AFRICAN | CONGOLESE | 20 | 1.30 | 0.84 |
| AFRICAN | ETHIOPIAN | 17 | 1.25 | 0.91 |
| AFRICAN | GHANAIAN | 114 | 1.32 | 0.86 |
| AFRICAN | NIGERIAN | 776 | 1.38 | 0.58 |
| AFRICAN | SIERRA LEONIAN | 75 | 1.29 | 0.84 |
| AFRICAN | UGANDAN | 1 | 1.00 | 0.00 |
| EAST ASIAN | CHINESE | 49 | 1.24 | 0.82 |
| EAST ASIAN | EAST ASIAN | 15 | 1.32 | 0.75 |
| EAST ASIAN | HONG KONGESE | 307 | 1.33 | 0.78 |
| EAST ASIAN | KOREAN | 9 | 1.27 | 0.45 |
| EAST ASIAN | MALAYSIA | 4 | 1.00 | 1.00 |
| EAST ASIAN | VIETNAMESE | 114 | 1.29 | 0.66 |
| ENGLISH | BLACK CARIBBEAN | 1 | 1.00 | 0.00 |
| EUROPEAN | AFRIKAANS | 41 | 1.18 | 0.90 |
| EUROPEAN | ALBANIA | 6 | 1.23 | 0.91 |
| EUROPEAN | BALKAN | 301 | 1.33 | 0.75 |
| EUROPEAN | BALTIC | 80 | 1.18 | 0.96 |
| EUROPEAN | CZECH & SLOVAKIAN | 24 | 1.21 | 0.85 |
| EUROPEAN | DUTCH | 104 | 1.16 | 0.95 |
| EUROPEAN | EUROPEAN OTHER | 12 | 1.14 | 0.54 |
| EUROPEAN | FRENCH | 186 | 1.26 | 0.88 |
| EUROPEAN | GERMAN | 282 | 1.20 | 0.93 |
| EUROPEAN | HUNGARIAN | 61 | 1.28 | 0.82 |
| EUROPEAN | ITALIAN | 699 | 1.34 | 0.83 |
| EUROPEAN | POLISH | 347 | 1.26 | 0.89 |
| EUROPEAN | ROMANIAN | 39 | 1.06 | 0.99 |
| EUROPEAN | RUSSIAN | 93 | 1.31 | 0.82 |
| EUROPEAN | UKRANIAN | 49 | 1.38 | 0.53 |
| GREEK | GREEK | 653 | 1.37 | 0.78 |
| HISPANIC | PORTUGUESE | 221 | 1.30 | 0.85 |
| HISPANIC | SPANISH | 469 | 1.25 | 0.89 |
| INTERNATIONAL | INTERNATIONAL | 10 | 1.00 | 0.96 |
| JAPANESE | JAPANESE | 148 | 1.20 | 0.89 |
| JEWISH & ARMENIAN | ARMENIAN | 53 | 1.24 | 0.90 |
| JEWISH & ARMENIAN | JEWISH | 244 | 1.21 | 0.92 |
| MUSLIM | BANGLADESHI | 1351 | 1.31 | 0.75 |
| MUSLIM | ERITREAN | 18 | 1.24 | 0.88 |
| MUSLIM | IRANIAN | 101 | 1.14 | 0.94 |
| MUSLIM | LEBANESE | 2 | 1.58 | 0.00 |
| MUSLIM | MUSLIM | 12 | 1.00 | 0.99 |
| MUSLIM | MUSLIM MIDDLE EAST | 676 | 1.17 | 0.90 |
| MUSLIM | MUSLIM NORTHAFRICAN | 7 | 1.00 | 0.97 |
| MUSLIM | MUSLIM SOUTH ASIAN | 15 | 1.09 | 0.87 |
| MUSLIM | PAKISTANI | 3326 | 1.33 | 0.72 |
| MUSLIM | PAKISTANI KASHMIR | 165 | 1.16 | 0.80 |
| MUSLIM | SOMALIAN | 45 | 1.31 | 0.80 |
| MUSLIM | TURKISH | 757 | 1.26 | 0.83 |
| NORDIC | DANISH | 86 | 1.09 | 0.96 |
| NORDIC | FINNISH | 112 | 1.32 | 0.83 |
| NORDIC | NORDIC | 3 | 1.00 | 1.00 |
| NORDIC | NORWEGIAN | 23 | 1.12 | 0.93 |
| NORDIC | SWEDISH | 72 | 1.13 | 0.95 |
| SIKH | SIKH | 1445 | 1.39 | 0.48 |
| SOUTH ASIAN | HINDI INDIAN | 2385 | 1.41 | 0.65 |
| SOUTH ASIAN | HINDI NOT INDIAN | 34 | 1.13 | 0.92 |
| SOUTH ASIAN | INDIA NORTH | 324 | 1.22 | 0.91 |
| SOUTH ASIAN | SOUTH ASIAN OTHER | 4 | 1.34 | 0.35 |
| SOUTH ASIAN | SRI LANKAN | 807 | 1.30 | 0.79 |
| UNCLASSIFIED | VOID | 334 | 1.19 | 0.90 |
| | | | | |
| **TOTAL** | | **18125** | **1.28** | **0.79** |

**Table 5: Summary of the contents of the final non-British forename seed list.**

## 5.1 Cycle 1; forename seed list and surname clustering

Cycle 1 of the FSC triangulation started with the forename seed list of 41,544 forename types, which as stated above, contained the forename, CEL Subgroup, and score. This forename seed list was used to search in the Electoral Register for surnames associated with them, and thus calculate the CEL Subgroup composition of each surname. This was achieved in a series of steps.

1) *Table linking*. The GB Electoral Register (GB04 file) was linked to two tables: to the forename seed list through the forename field; and to a new forename-to-gender table also through the forename field. Figure 4 shows the relationships between these tables. The forename-to-gender table was created associating each forename in the GB Electoral Register with a gender by aggregating the gender field for each elector in the Electoral Register by forename. As a result, each forename ended up with one of four possible 'gender values'; Female, Male, Both, or Unknown. 'Both' referred to forename types where both sexes represented at least 10% of the total forename tokens, while 'unknown' were cases of forename types with no gender reported in the Electoral Roll.

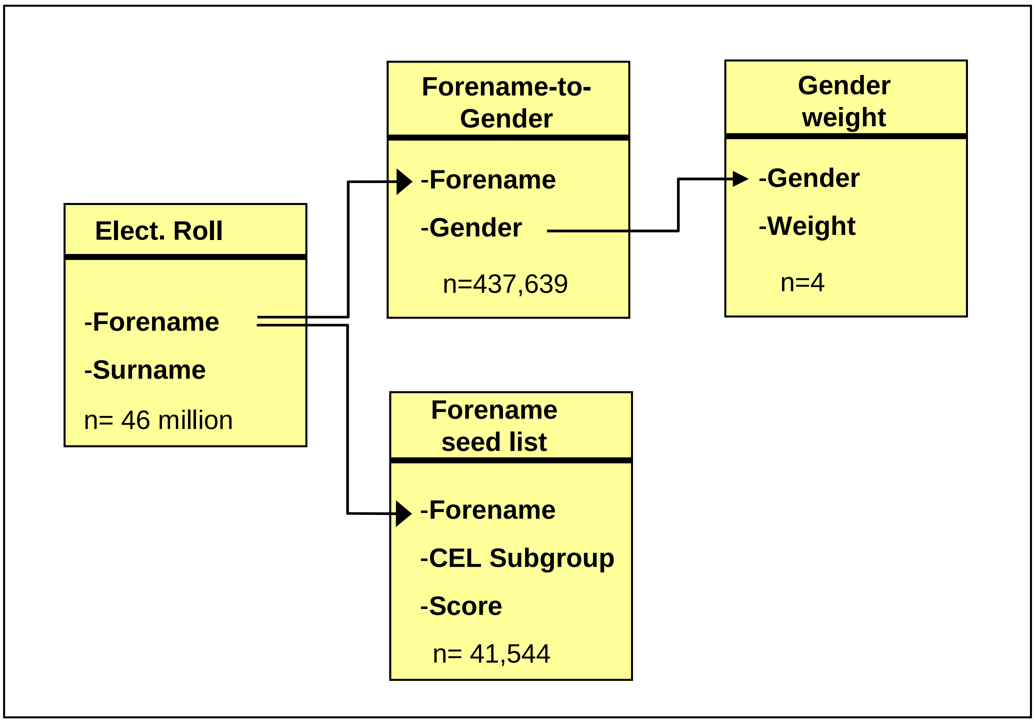

**Figure 4: Tables relationships in cycle 1 step 1**

Diagram of the relational database structure between the tables in cycle 1 step 1. The direction of the arrows represent a 'many to one' type of relationship between the tables (from arrow stem to arrow pointer). n=number of records in each table. The 'gender weight' table has 4 records (Male, Female, Both, Unknown).

2) *Gender weighting*. In order to improve the effectiveness of the FSC technique, Tucker (2005) proposed to weight down those forenames known to be of female gender to reduce the adverse impact of multicultural marriages in which the woman takes the husband's surname, thereby introducing what he describes as an 'artificial relationship' (perhaps more correctly an 'ambiguous relationship') between a forename's and a surname's CEL. When applying his version of FSC, Tucker (2007) used a weight of 0.8 for every female forename in the database, while male or unisex forenames were assigned a weigh of 1. It was decided to adopt the same successful strategy in this automated classification. A new table of gender weightings (Female = 0.8, Male = 1, Both/Unknown = 1) was created and linked to the forename-to-gender table as shown in Figure 4. A summary of the gender distribution of the GB 04 Electoral Register in terms of the number of forename tokens and types is shown in Table 6.

| Gender | Forename Tokens | | Forename Types | |
|---|---|---|---|---|
| Female | 24,702,133 | 53.3% | 180,497 | 41.2% |
| Male | 21,446,125 | 46.3% | 152,736 | 34.9% |
| Both or Unknown | 158,064 | 0.3% | 104,406 | 23.9% |
| **Total** | **46,306,322** | 100.0% | **437,639** | 100.0% |

**Table 6: Summary of the total number of forename tokens and types per gender in GB 04 Electoral Register**

3) *Calculation of personal weight score*. A query was performed on the tables shown in Figure 4, for every person whose forename was found in the forename seed list, producing a record including; forename, surname, forename CEL Subgroup, forename score, gender weight, and a 'weighted personal score' (calculated by multiplying the forename score by the gender weighting). These records were stored in an interim table termed here table (A).

4) *Calculation of surname to CEL Subgroup frequencies and cumulative score*. For every surname type and CEL Subgroup combination in table (A), a calculation was made summing up the weighted personal scores (creating a 'cumulative personal score'), and counting the frequency of forename tokens and forename types, calculating the relative frequency (in percentage) of both forename tokens and types over the total for that surname. If the percentage of forename tokens of the British and Irish CEL

Subgroups was below 95%, then these CEL Subgroups and their associated frequencies were removed from the calculation. This threshold was selected as a result of the classificatory experience in the heuristic approach, and it is related to the overall size of the non-British or Irish minorities in the UK. The percentage of non-‘White British/White Irish’ groups in the 2001 UK Census is 10.8%. However, because of the abundance of British forenames amongst second generation ethnic minorities, and the number of multicultural marriages involving surname change, when calculating the expected percentage of non-British forename tokens in the population, the final figure should be much lower than 10.8%. The exploratory analysis developed in the heuristic approach indicated that the real threshold should be 5% of the overall forename tokens. This is the threshold used here, assuming any surname with less than 95% of its forename tokens as British or Irish should be taken as most likely of ‘foreign’ origin. The results of the calculation described in this step were stored in an interim table (B), an example of which is provided in Table 7.

| | | | | **Cumulative personal score** | |
|---|---|---|---|---|---|
| **Surname** | **CEL Subgroup** | **Forename Tokens** | **Forename Types** | **Value** | **Percentage** |
| CARVALHO | SPANISH | **61** | 22 | **62.47** | **38.89%** |
| CARVALHO | GHANAIAN | 1 | 1 | 0.13 | 0.08% |
| CARVALHO | NIGERIAN | 1 | 1 | 1.14 | 0.71% |
| CARVALHO | HINDI INDIAN | 1 | 1 | 0.05 | 0.03% |
| CARVALHO | PORTUGUESE | **50** | 32 | **76.92** | **47.88%** |
| CARVALHO | ITALIAN | 20 | 14 | 19.94 | 12.41% |
| **TOTAL** | | **134** | **71** | **160.65** | **100.00%** |

**Table 7: Example of the different CEL Subgroups associated with a surname type as calculated in step 4**

5) *Selection of the CEL Subgroup with the highest cumulative personal score*. For each surname type in table (B) (see the example given in Table 7), the CEL Subgroup with the highest cumulative personal score was selected as the most representative CEL Subgroup of that surname. In the example given in Table 7 this resulted in the classification of the forename ‘Carvalho’ to the ‘Portuguese’ CEL Subgroup, with a total cumulative score of 76.92. This example is very interesting to demonstrate the value of using the highest cumulative score as opposed to just the highest counts of forename tokens or types. ‘Carvalho’ actually had more forename tokens associated

with the ‘Spanish’ CEL Subgroup, but the Portuguese ones had much higher scores and were weighted more in the final allocation (see figures highlighted in bold in Table 7). This is because of a historic overlap between Portuguese and Spanish forenames (which are derived from the same catholic religious figures written exactly in the same way in both languages), an example of a problem that can be overcome by using the scores in the forename seed list as described in Section 4.2.

6) *Creation of a new surname-to-CEL table.* A new interim table (C) was created using the result of the previous step. It was then filtered to remove any surname types with a total frequency of less than 10 tokens, in order to avoid potential future misallocations of CELs through further iterations of FSC because of rare surnames. This final table was termed the ‘surname-to-CEL table’ and included the following fields:
    - Surname type
    - CEL Subgroup (selected in step 5)
    - Average personal score (see below)

    The ‘average personal score’ was calculated by dividing the ‘cumulative personal score’ of the selected CEL Subgroup by the number of forename tokens of that CEL Subgroup. In the example given in Table 7 this was 76.92 / 50 = 1.54, meaning that the surname ‘Carvalho’ is associated with Portuguese forenames in the seed list that taken together have a gender and population weighted average score of 1.54. At this stage the new ‘surname-to-CEL table’ had 90,729 surname types.

7) *z-score standardisation and final score selection.* The average personal score calculated above for the ‘surname-to-CEL table’ was standardised using z-scores, using the mean and the standard deviation of the average personal score values within each CEL Subgroup. The z-scores were calculated in exactly the same way as shown in Section 4.2 (step 8) above. As pointed out in that section, the result of this standardisation was a positive or negative value distributed around zero and with a range determined by the number of standard deviations away from the mean that bounded the most extreme values. Those surnames with an average z-score of less than -1 were deleted from the surname-to-CEL table, since they were deemed to not be representative of the CEL Subgroup in this first cycle. Finally, the average z-score was transformed by adding to it a value of ‘1’ resulting in a final ‘surname score’. The range of surname scores in this case was between 0 and 0.83. The resulting surname

score was added to the final version of the ‘surname-to-CEL’ table which at this stage had 72,884 surname types.

As a result of these seven steps in cycle 1 a new ‘surname-to-CEL’ table with 72,884 surname types was created, with just three fields; ‘surname’, ‘CEL Subgroup’, and ‘score’. This is the first version of this table, which after subsequent iterations of cycles 1 and 2 was expanded with more surnames as will be explained at the end of this section.

### 5.2 Cycle 2: surname-to-CEL table and forename clustering

Cycle 2 of the automated approach used the surnames-to-CEL table to classify further forenames by CEL Subgroup. Repetition of the descriptions of calculations performed which were identical to cycle 1 will be avoided here, and reference will be made to the detailed explanation in subsection 5.1. The purpose of this subsection will be to highlight any differences in the approach. The terminology of SCEL and FCEL used in Chapter 5 will be used here again. SCEL refers to the CEL Subgroup assigned to a surname and FCEL to that assigned to a forename.

Therefore, the objective of cycle 2 was to classify a large number of forename types into CEL Subgroups, beyond the original 18,125 forename types included in the forename seed list. This was achieved through the following steps, mirroring those described in cycle 1.

1) *Table linking*. The GB Electoral Register (GB04 file) was linked to three tables; to the surname-to-CEL table, developed in the previous subsection, through the surname field, to the forename-to-gender tables through the forename field, and through the latter to the gender weighting table. Figure 5 shows the relationships between these three tables, which is very similar to Figure 4, differing only in bottom-middle table.

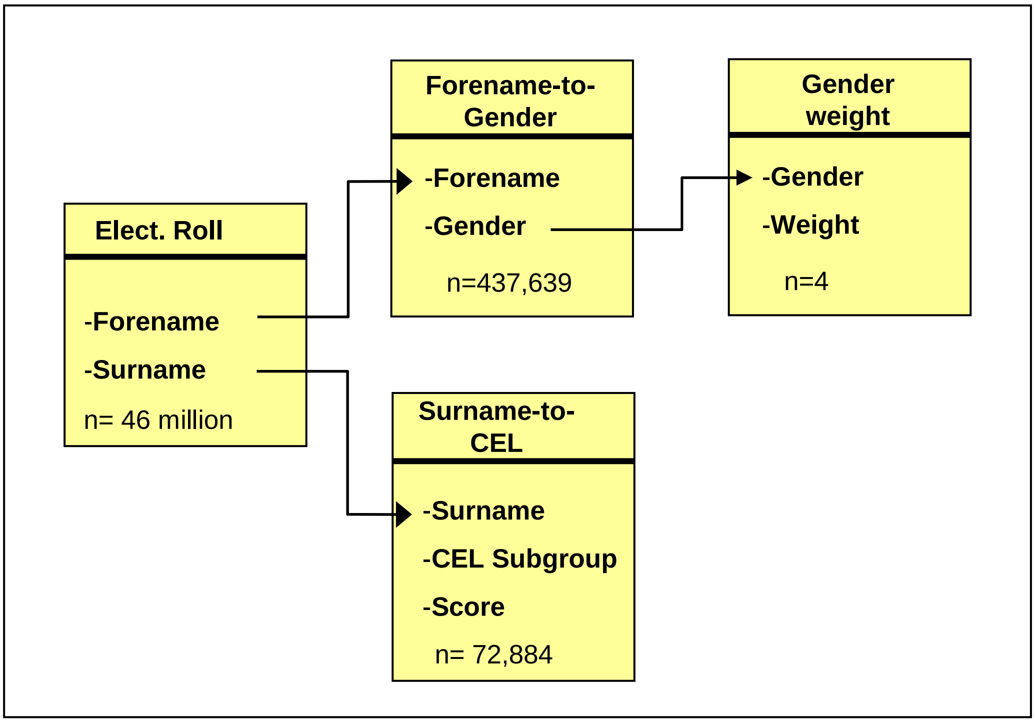

**Figure 5: Tables relationships in cycle 2 step 1**

Diagram of the relational database structure between the tables in cycle 2 step 1. The direction of the arrows represent a 'many to one' type of relationship between the tables (from arrow stem to arrow pointer). n=number of records in each table. The 'gender weight' table has 4 records (Male, Female, Both, Unknown).

2) *Gender weighting*. The same type of gender weighting was applied in this iteration as in step 2 of cycle 1.

3) *Calculation of personal weight score*. A query was performed on the tables shown in Figure 5, for every person whose surname was found in the surname-to-CEL table, producing a record including;

   - Forename
   - Surname
   - SCEL (the CEL Subgroup from the surname-to-CEL table)
   - Surname score (from the surname-to-CEL table)
   - Gender weight
   - 'Weighted personal score' (calculated by multiplying the surname score by the gender weighting).

   All of these records were stored in an interim table termed here table (A).

4) *Calculation of forename to CEL Subgroup frequencies and cumulative scores*. For every forename type in table (A) and CEL Subgroup combination, the same calculation applied in cycle 1 step 4 was performed here, including the removal of

British and Irish CELs if the percentage of surname tokens was below the 95% threshold. The results of this calculation were stored in table (B).

5) *Selection of the CEL Subgroup with the highest cumulative personal score*. For each surname type in table (B) above, the CEL Subgroup with the highest cumulative personal score was selected as the most representative CEL Subgroup of that forename type.

6) *Creation of a new forename-to-CEL table*. A new interim table (C) was created with the result of the previous step. It was then filtered to remove any forename types with a total frequency of less than 5 tokens, in order to avoid potential future misallocations of CELs through further iterations of FSC because of rare surnames. This final table was termed the 'forename-to-CEL table' and included the following fields:
   - Forename type
   - CEL Subgroup (selected in step 5)
   - Average personal score (as per cycle 1 step 6)

   At this stage the new 'forename-to-CEL table' had 89,211 forename types

8) *z-score standardisation and final score selection.* The average personal score calculated above for the 'forename-to-CEL table' was standardised using z-scores, in exactly the same way as shown in cycle 2, step 8 above, including the truncation of values below -1 and the transformation by adding to it a value of '1'. This resulted in the final 'forename score'. The range of surname scores in this case was between 0 and 0.72. The resulting forename score was added to the final version of the 'forename-to-CEL' table which at this stage had 81,653 forename types.

9) *CEL Subgroup consistency check.* A final check was performed on this new 'forename-to-CEL table' for those forename types that already existed in the 'forename seed list' comparing the attributes of both tables. If there was a mismatch between the two CEL Subgroups independently assigned in each of these two tables, the forename was finally allocated to the CEL Subgroup with the highest score.

## 5.3 Subsequent cycles of forename-surname clustering (FSC)

Cycles 1 and 2 described in this section were essentially two iterations of the same process. Subsequent iterations of these cycles were further continued into cycles 3, 4 and beyond, comprising a true automated approach. This increased the number of surnames and forenames that were classified with a CEL bringing it as close as possible to the objective of classifying all forename types and surname types with a frequency of 3 tokens or more in the GB Electoral Register. Figure 6 shows this iteration of cycles diagrammatically, with cycle 1 starting with a forename seed list to produce a surname-to-CEL table which is in turn used to produce a forename-to-CEL table in cycle 2 and both expanded through subsequent cycles 3 and beyond.

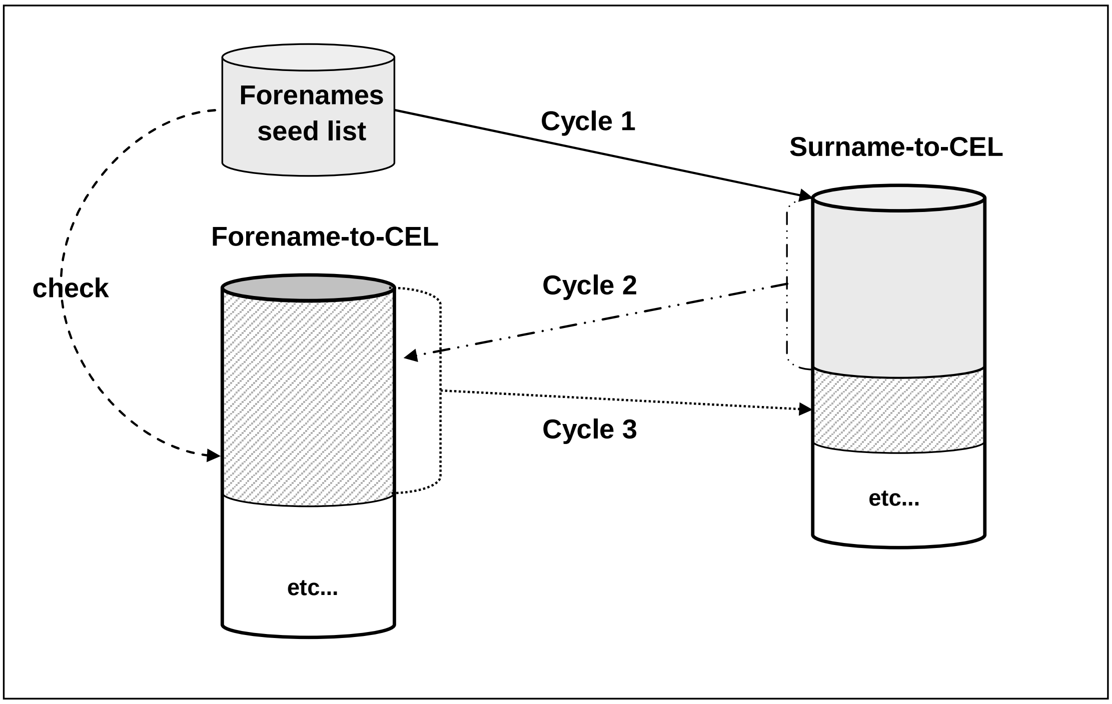

**Figure 6: Cycles in the automated classification**

This diagram shows the process flow described in Section 6.4, starting with a forenames seed list used in cycle 1 to produce a surname-to-CEL table, which is then in turn used in cycle 2 to produce a forename-to-CEL table, and so on. Only cycles 1, 2 and 3 are shown in the diagram.

The final surname-to-CEL table had 225,576 surname types, and the final forename-to-CEL table had 98,624 forename types.

# 6 Validating the CEL Name Classification

In order to demonstrate the usefulness of this CEL classification for applications to classify a population into cultural, ethnic and linguistic groups, and to measure its classificatory effectiveness, it first needs to be validated against some populations for which ethnicity is already known through an independent source (i.e. not based on names). Thus this section's objective is to validate the effectiveness of the CEL classification proposed in this paper.

Hereinafter ‘CEL classification’ will be used as a shorthand name for the automated classification of names into CEL Subgroups presented in the previous chapter. When the term CEL is used, it will refer to any or all levels in the CEL taxonomy, although it will usually refer to the CEL Subgroups.

In order to evaluate the CEL classification a preliminary step is to apply the separate forename-to-CEL and surname-to-CEL lists developed in the previous chapter in order to classify individuals into a single CEL. If both FCEL and SCEL are identical the solution is straightforward, but some type of arbitration is required when there is a conflict between the two. This is where the scores developed in Chapter 6 prove very useful. The first section of this chapter will deal with the rules developed to assign a person with a CEL, using what will be termed the PCEL (Person Cultural, Ethnic and Linguistic group). Once a PCEL has been assigned to an individual, validation of the classification can take place.

## 6.1 Person Level CEL Allocation Algorithm

This section explains the process by which the CEL classification 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 classification process described in chapters 5 and 6 assigns a categorical SCEL classification to each surname and an FCEL to each forename. Such a categorical assignment is necessary in order to maximise the accuracy of the FSC technique in the assignment of a CEL Subgroup to forenames and surnames. However, once the process of FSC assignment has been finalised resulting in the categorical assignment of names to SCELs and FCELs, the use of a proportional assignment for each person CEL (PCEL) was also developed.

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 if assigned only to a single CEL. Proportional assignment is also useful in instances where the actual boundary between different CEL categories is imprecise, whether geographical, linguistic, religious, or cultural. An example of geographical boundary imprecision is for instance between the Netherlands and Germany where many names are common in both cultures. Proportional assignment will also be useful in the future in situations where other multiple sources of information could be used in combination with a name's ethnicity, such as for example the postcode of a person's residence or his or her place of birth. These aspects of the application of proportional assignment to the classification of actual people by CEL (as opposed to the classification of particular name types) are described in this section.

In order to facilitate the process of proportional assignment of CELs to a person, the name-to-CEL scores created in Chapter 6 will be used. These scores represent the degree to which a CEL allocated to a name type 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 if the FCEL of the person is not of South Asian origin, it can easily override the SCEL and the person be finally assigned with the FCEL.

The two name-to-CEL tables explained in the Chapter 6 are 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:
(The algorithm is presented as pseudo-code, with comments tagged as '##' and in italics)

```
## Evaluate if both CEL Subgroups are the same
CASE1         SCEL Subgroup = FCEL Subgroup, then:
        ## Assign PCEL
        PCEL = SCEL Subgroup = FCEL Subgroup
## Evaluate if CEL Groups are the same and if so assign that CEL Group
CASE2         SCEL Group = FCEL Group, then
        PCEL= SCEL Group= FCEL Group
## If the absolute difference between scores is small then assign PCEL to the CEL Group with the highest score
CASE3         |SCEL Subgroup score - FCEL Subgroup score| < 0.05, then
        PCEL= MAX(SCEL or FCEL Group score)
## Evaluate if both SCEL and FCEL exist for that person and assign PCEL to the CEL Subgroup with the highest score
```

```
CASE4       SCEL AND FCEL exist, then
            PCEL= MAX(SCEL or FCEL Subgroup 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 FCEL 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 Subgroup 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 future user of this classification 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 score, 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 the PCEL was assigned:

```
## 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 Subgroup score - FCEL Subgroup 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'.

## 6.2 Inherent Difficulties of External Validation of the Classification

The evaluation of the power of name classifications to stratify a population of individuals into ethnic groups has been a recurrent theme in the public health literature for over the last half a century. A full review of this history and the features of the main studies is offered in Chapter 3 and will not be repeated here. The general pattern of these studies is that they first develop a name-to-ethnicity reference list, based on a reference population, which is later applied to classify a second independent names list, termed target population, for which its ethnicity is previously known through an independent method (i.e. not based on names). As it was discussed in Chapters 3 and 4, this paper research did not follow this route because the objective of classifying the entire population of Great Britain into all of the possible ethnic groups present was at odds with the possibility of obtaining such reference and target populations with ethnicity information for the whole country. Therefore, because of this lack of availability of extensive ethnicity data, the validation of the classification for the entire population cannot be done in the same way in these studies. However, appropriate ethnicity and name data were obtained for a fraction of the population in London, and it is using these data that one aspect of the validation will be based.

Furthermore, another major issue with validation of name-based ethnicity classifications is related to the difference in the ontological nature of the qualities to be compared and evaluated. This is linked to problems of defining and measuring ethnicity, reviewed in Chapter 2. As discussed in that chapter, ethnicity is a socially constructed concept and ethnic self-identification is a subjective decision of the individual that can change through time, with the method of data collection, type of question asked and the group categorization offered. On the other hand, the concept of cultural, ethnic and linguistic groups developed in this paper extending from the previous onomastics literature, relate to the independent measurement of differences in the naming practices, geographies and histories of human groups. As such, self-reported ethnicity and automatically assigned name-based cultural, ethnic and linguistic groups are two constructs that differ substantially in nature and definition and hence the validation of the latter using the former presents inherent difficulties.

One example of the ontological problems found in the research reported in this chapter, is the high mismatch between the proportion of persons with Irish names in Britain and the proportion of persons who define themselves as of ‘White Irish’ ethnicity in the UK 2001 Census, the latter being usually much smaller than the former. This is of course because of the long history of Irish migration in Britain, the different perceptions of Irish identity that people in Britain with names originating in Ireland have, the time that has elapsed since migration and number of generations passed, as well as engagement with aspects of identity politics, religion and nationalism of a very subjective nature. On the other hand, rule-based name to ethnicity classifications are blind to these aspects and as expected classify all people with identical names in the same way.

This difference of nature between self-identified ethnicity and name-based cultural, ethnic and linguistic groups should not be seen as inherently disadvantageous. Indeed the blindness of name-based CEL classification can be used to identify the heavy baggage that is sometimes attached to self-assigned ethnicity classifications and to present a picture that is unaffected by the design of the data collection method and or changing public perceptions of identity. Therefore, such ontological distinctions should be taken as potential caveats when interpreting or validating name-based ethnicity classifications using self-reported ethnicity data, since they can never be identical.

Other ways of validating the CEL classification might have entailed manual checking of two types. One option would have been to check the CEL classification of the individual names against the onomastic or linguistic origin of a name using several name dictionaries. Clearly, this would have been very time consuming since several dictionaries would need to be used, sometimes there are several entries for each name to choose from, and the coverage of surname dictionaries is very poor. A second option would have been to ask volunteers from different countries to evaluate their opinions of the CEL classifications, but this would have been very subjective and prone to error. Preliminary manual checks performed as part of the heuristic phase of the classification indicated that there is a tendency towards high overlap between different people classifying names from close cultures or languages - such as Portuguese, Spanish and Italian names, or Bangladeshi and Pakistani - who tend to classify a large number of names as from ‘their own’ culture. This problem of high overlap between manual coders has also been reported in the literature (Martineau and White, 1998).

As a result, self-reported ethnicity is by far the best data source available for the purpose of validating the CEL classification, and hence it will be used in this chapter in two different types of validation. The results nevertheless need to be interpreted in the light of the necessary caveats discussed in this section.

### 6.3 Validation against Hospital Admission Ethnicity Data

Following a Public Health tradition in name-based ethnicity studies, the CEL name classification was validated against self-reported ethnicity recorded in hospital admissions data. As part of the *Knowledge Transfer Partnership* between University College London and Camden Primary Care Trust that funded this paper research, access to an extensive list of individuals admitted to hospital was obtained from Camden and Islington Primary Care Trusts (PCT) for research purposes. Permission for access to this dataset was requested by the author and approved by Camden PCT Research Ethics Committee, Islington PCT Caldicott Guardian, and North Central London NHS Research Consortium.

The ethnic composition of the population of Camden and Islington is very diverse, as it can be seen in Figure 7, which lists the percentage of the total population that each ethnic minority represents in the London Boroughs of Camden and Islington, compared with the equivalent shares for London and England. From those relative differences it can be clearly appreciated that the Bangladeshi group is the largest minority group, followed by individual ethnicities that comprise the 'White Other' and 'Black African' categories (such as Somalis, Greeks, Kosovans, or Congolese). In Camden schools alone, there are 3,100 speakers of Bengali/Sylheti, over 1,100 of Somali, and more than 200 speakers of each of the following languages; Albanian, Arabic, French, Spanish, Portuguese and Lingala (London Borough of Camden, 2007).

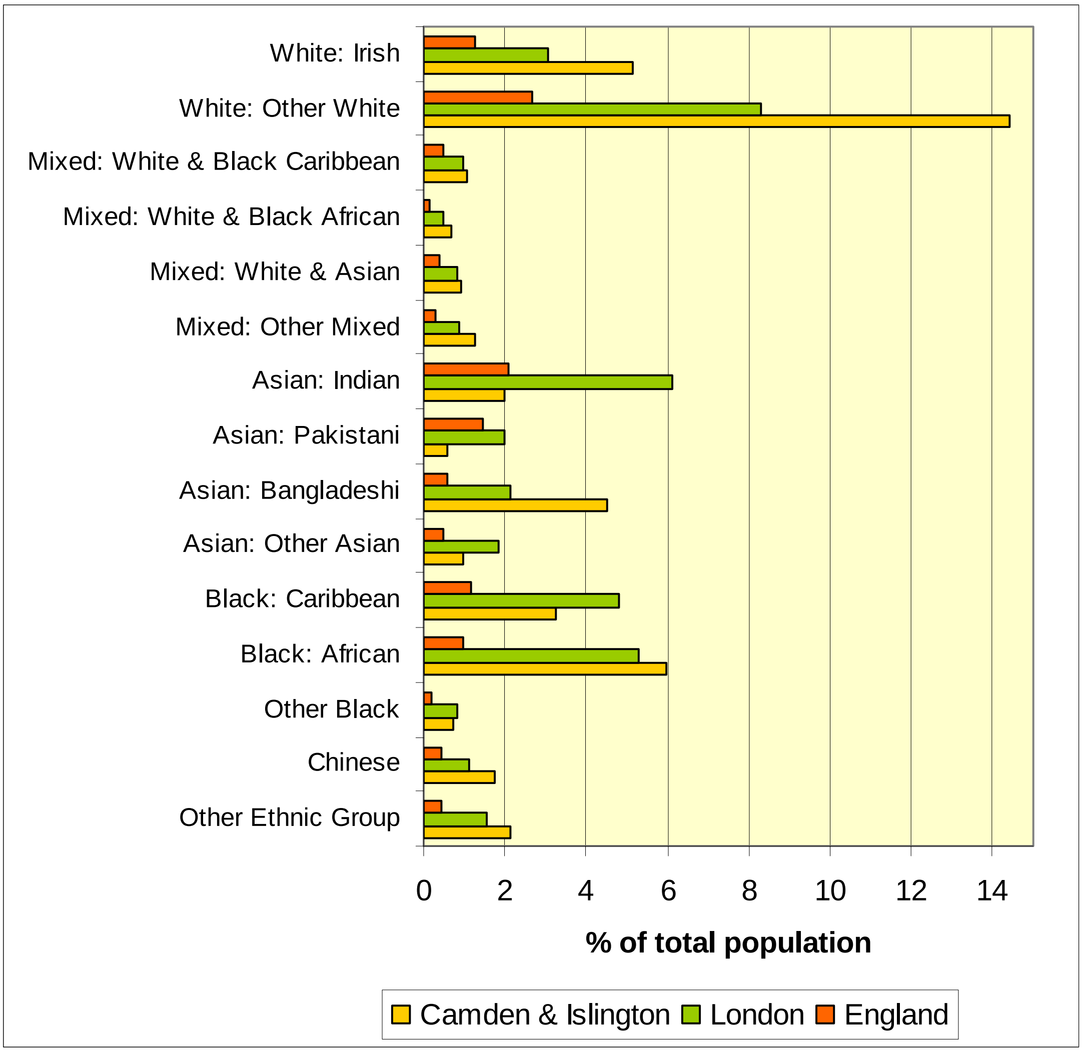

**Figure 7: Population by ethnic minority in the London Boroughs of Camden and Islington**

The chart shows the percentage of the total population that each ethnic minority represents in the London Boroughs of Camden and Islington (dark yellow), compared with the percentages for London (green) and England (dark orange). Source: ONS 2001 Census.

## 6.4 Hospital Episode Statistics data sources and preparation

The dataset accessed is termed Hospital Episode Statistics (HES) within the NHS, and hereinafter is referred to as 'HES'. It includes an entry for every hospital admission of both inpatients and outpatients, although only inpatients are used for this study. For every admission, a set of general information about the person admitted, the medical condition and various other hospital administrative transactions is recorded. For the purposes of this validation only the information about the person admitted was obtained. The relevant fields of patient general information actually used are the following:

- NHS Number (a unique ID for every patient in the National Health Service)

- Patient Forename
- Patient Surname
- Patient Sex
- Patient Unit Postcode
- Patient Date of Birth
- Patient ethnic group

The time period available was 8 years worth of data, from April 1998 to March 2006, for patients registered in the London Boroughs of Camden or Islington, which gave a total number of 835,144 hospital admissions, belonging to an approximate number of 343,068 unique patients. For reference purposes, the total population of these two London Boroughs was 373,817 people in the 2001 Census, although, being situated in inner London they have a high population turnover estimated in 20% a year (London Borough of Camden, 2007). Assuming this turnover rate is of people who move in and out these two Boroughs (rather than within), and never come back during the 8 years of HES data, this would produce total number of 523,343 people moving out of the area (373,817 x 20% x 7 interannual periods), which in addition to the permanent 'stock of residents' would give a total 897,160 Camden and Islington 'accumulated residents' over the 8 year period (523,343+373,817). This means that the HES data represents the 38% of the likely total potential population that was admitted to hospital during that period. This comprises a large proportion of the total population, although it may be biased in its characteristics in terms of age and ethnicity. It is widely known that elderly people are much more likely to be admitted to hospital than younger cohorts, and in London it has also been proved that hospital admission rates are associated with the general prevalence of chronic illness and deprivation in a local population (Majeed et al, 2000). Therefore, the population represented by the HES dataset is expected to be more weighted towards older groups and socio-economically deprived groups.

However, the completeness and quality of the HES dataset is very poor, especially regarding patient ethnicity data. Frequent problems include: inconsistent ethnic group coding, sometimes even for the same patient; mixing of the 1991 and 2001 Census ethnicity classifications; or use of the catch all 'Unknown Ethnicity' category. Ethnicity coding in Hospital Admissions has been mandatory in the UK since 1994 (NHS Executive, 1994), but it has taken a long time to reach nearly full coverage and a consistent coding framework. This poor quality has been widely denounced in the literature (Aspinall and Jacobson, 2004;

Association of Public Health Observatories, 2005), and although specific nationwide guidelines have been issued to improve the situation (Department of Health, 2005) the percentage of hospital admissions being correctly coded by ethnicity has been estimated to be 75% in 2005 (London Health Observatory, 2005). This makes a strong case for the use of name-based ethnicity classifications to audit and complete routinely collected self-reported ethnicity data.

Moreover, another problem with the dataset was missing information, crucially the NHS number with a high proportion of patient admissions having missing (27%), or incomplete or wrong NHS numbers (2% for the two errors combined). This had important implications since the HES dataset is a register of hospital transactions, and not a register of people. That is, when the same person is admitted several times to hospital, a new independent record is generated. In order to be able to study individuals in a 'hospital population', independently of how many times they have been admitted, aggregation of all admissions of each individual person is required. When the same NHS number is not correctly recorded in repeated admissions, there is a high risk of the same person being included several times in the population register.

### 6.4.1 Data preparation: Hospital Episode Statistics

As a result of these important problems of data quality, it was necessary to cleanse the HES data before actually performing the analysis. The steps taken are summarised as follows:

a) *Individual admissions in HES are aggregated by person:*

In this step individual admissions to hospital throughout the 8 years were aggregated by person to create a unique person entry. There were two possible cases followed by a specific action:

Case1: The NHS number is present and complete (71% of HES), and records were aggregated by person through their NHS number.

Case 2: Where no NHS number was present, or it was incorrect, the admissions were aggregated in two steps:

- Firstly, aggregation by date of birth and postcode (which is deemed in the literature to represent unique patients in HES)

- Secondly, aggregation of the above again by date of birth and surname (to avoid duplications of people who have had different addresses)

A final unique ID number was assigned to every person (343,068 people), and traced back to every HES admission.

b) *Ethnic group codes are cleansed*

A range of 220 different ethnic group codes were found in the dataset. However, most admissions had been assigned to the 40 most common codes. A mapping exercise was performed between these 220 codes and the official ethnic group classifications valid during the 8 year period, the 1991 Census ethnicity classification and its 2001 successor, with the help of published references on how hospital admissions should be coded in NHS systems (Department of Health, 2005; NHS Information Authority, 2001).

Finally, in order to be able to compare all HES admissions using the same ethnic group categories, a further lookup table was created, mapping the 2001 Census ethnic groups to the 1991 ones, according to criteria proposed by Platt *et al* (2005).

c) *Individual person ethnicity is assigned*

The aggregated data by person created in step a) were linked to the ethnic group code of each patient, using the 1991 Census categories for all patients, and computed as follows:

- If all HES admissions for the same person contained a unique ethnic group code, the person was assigned with that code (89.8% of the patients).
- For the rest of cases, if after removing the ethnic group categories 'other' or 'unknown' ('8' and '9' in the 1991 Census) the rest of admissions included a consistent code, then the person was assigned with that code (9.8%).
- Otherwise, the person was assigned with a special code for a 'conflicting ethnic group flag' and left outside the analysis (0.4%).

The same process was repeated for the 2001 Census categories, but only for those patients for whom this information was available (178,623 patients or 52% of the total), with the three previous steps results being respectively; 70.1% (unique code), 27.3% (unique after 'S-Other' and 'Z-Unclassified'), and 2.6% (conflicting).

d) *Creation of a final table of individual patients (HES_Person)*

A final table of 343,068 people was generated including the following fields:

- Person ID Number (internally generated)
- NHS Number (if known)

- Person Forename
- Person Surname
- Person Sex
- Person Date of Birth
- Person 1991 ethnic group
- Person 2001 ethnic group (if reported)

This last table, termed *HES_Person*, then formed the basis to subsequent analysis

### 6.4.2 Data preparation: CEL name classification

In order to be able to compare the ethnic group categories reported in HES, that is, the 1991 and 2001 Census ones (10 and 16 groups respectively), with the 66 CEL taxonomy described in Chapter 6, a lookup table between the two needs to be created. This was done by analysing the characteristics of each of the 66 CEL Subgroups and the metadata gathered about them which are presented in Appendix 3. These follow the guidelines established by the Office of National Statistics (ONS) when ascribing individual responses in the Census to one of the pre-set ethnic groups (Office for National Statistics, 2003). These decisions were taken based on the strongest component describing the CEL, be it geographic location, religion or language, and its corresponding allocation in the ONS categorization. As a result, a lookup table between each CEL Subgroup and both a 1991 and a 2001 Census category was established as presented in Appendix 3.

## 6.5 Data Analysis

### 6.5.1 Data analysis: comparing CEL with HES ethnicity

The 343,068 people in *HES_Person* table were assigned to CEL Subgroups using their forenames and surnames and applying the name-to-CEL tables and the personal allocation algorithms described in Section 6.1. A summary of the results obtained at CEL Group level is presented in Table 8, although the individual allocations were made at the CEL Subgroup level. The coverage of names classified was 96.4%, meeting already one of the primary aims of this research: to classify populations into all of the potential ethnic groups present in a society, recognising the majority of names in the local population of Camden and Islington.

| CEL SUBGROUP | PEOPLE | % |
|---|---:|---:|
| ENGLISH | 144,875 | 42.2% |
| CELTIC | 68,682 | 20.0% |

| | | |
|---|---|---|
| MUSLIM | 47,602 | 13.9% |
| EUROPEAN | 23,692 | 6.9% |
| HISPANIC | 10,691 | 3.1% |
| AFRICAN | 8,862 | 2.6% |
| SOUTH ASIAN | 7,158 | 2.1% |
| GREEK | 6,763 | 2.0% |
| EAST ASIAN | 4,481 | 1.3% |
| JEWISH& ARMENIAN | 3,022 | 0.9% |
| NORDIC | 2,938 | 0.9% |
| SIKH | 1,215 | 0.4% |
| JAPANESE | 622 | 0.2% |
| INTERNATIONAL | 50 | 0.0% |
| VOID | 10,367 | 3.0% |
| UNCLASSIFIED | 2,048 | 0.6% |
| **TOTAL** | **343,068** | **100.0%** |
| Total valid CELs | 330,603 | 96.4% |
| Total non-valid CELs | 12,465 | 3.6% |

**Table 8: Results of classifying the HES_Person table using the CEL name classification summarised at CELGroup level**

The CEL Subgroup assigned to each person, was then re-computed into their corresponding 1991 and 2001 Census ethnic group code, using the lookup table described in section 6.4.2. At this stage a database query was created to compare the ethnic code in *HES_Person* table with that derived using the CEL name classification (converted to Census categories). The query generated a matrix comparing the results of both classifications over the same people in *HES_Person*, using the 1991 Census categories for all persons, and a separate matrix with the 2001 Census categories only for those patients for which they had been originally reported at this level (52% of the total patients).

| | | **Actual Ethnicity from HES data** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Predicted by CEL** | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **Total** |
| 0 | White | **150,574** | 7,971 | 4,468 | 2,535 | 595 | 68 | 160 | 488 | 17,383 | 73,920 | 258,162 |
| 1 | Black - Caribbean | 92 | **226** | 21 | 32 | 3 | | | | 69 | 197 | 640 |
| 2 | Black - African | 857 | 283 | **5,996** | 698 | 53 | 14 | 41 | 23 | 1,695 | 4,716 | 14,376 |
| 3 | Black - Other | | | | | | | | | | | 0 |
| 4 | Indian | 1,066 | 96 | 562 | 125 | **2,184** | 85 | 171 | 30 | 1,679 | 3,503 | 9,501 |
| 5 | Pakistani | 856 | 60 | 1,736 | 306 | 690 | **861** | 2,390 | 17 | 2,507 | 4,625 | 14,048 |
| 6 | Bangladeshi | 284 | 30 | 373 | 122 | 687 | 194 | **6,086** | 5 | 1,174 | 3,777 | 12,732 |
| 7 | Chinese | 227 | 39 | 72 | 21 | 11 | 2 | 7 | **1,473** | 531 | 1,088 | 3,471 |
| 8 | Other ethnic group | 3,811 | 111 | 990 | 228 | 202 | 112 | 280 | 358 | **5,858** | 5,747 | 17,697 |
| *9* | *Not Given /Unclassified* | 3,364 | 328 | 1,706 | 322 | 164 | 32 | 107 | 47 | 2,199 | **4,079** | 12,348 |
| | **Total** | 161,131 | 9,144 | 15,924 | 4,389 | 4,589 | 1,368 | 9,242 | 2,441 | 33,095 | 101,652 | **342,975** |

**Table 9: Matrix comparing number of persons by CEL vs. HES Ethnicity using 1991 Census ethnic groups**

Table 9 shows an example of the results based on the 1991 Census classification, as a 9 rows x 9 columns matrix, including the ethnicity predicted by the CEL name classification, as rows, against the actual ethnicity reported in the HES data for that same people, as columns. The over-all prediction success was 51.7%, calculated by summing the elements on the principal diagonal divided by the total number of persons.

As can be appreciated in Table 9, ethnic group '3- Black Other' cannot be estimated using name analysis and hence the line is blank. Furthermore, the ethnic group '9-Unclassified' includes all void and unclassified names in the CEL prediction, whereas the HES data include people who did not report their ethnicity or for whom recording was subject to the data errors discussed above. Therefore, both classifications cannot be considered on a like for like basis, because they measure different things. Thus the column '9- Not Stated' from the HES data was removed from further analyses, since it did not provide sufficient relevant information. However, row '9-Unclassified' was left in the analyses, since it was an output from the CEL classification. Despite this, it is interesting to see that out of the total 101,652 patients with an '9-Unclassified' code in HES the CEL classification is able to identify 96% of them with a likely CEL, most of them as 'White' (73%), which in itself provides another advantage of the CEL classification method in improving poor quality HES data.

In order to evaluate these results, the aim of the name classification should be to maximise the number of cases along the principal diagonal of the matrix in Table 9, and minimise the cases elsewhere in the matrix. In the public health literature, binary classifications of individuals, represented in similar 'confusion matrices', are evaluated according to a set of four widely accepted measures, which are also used in computer science to evaluate any binary classifier. These four measures are known as *sensitivity*, *specificity*, *positive predictive value* (PPV), and *negative predicted value* (NPV), and they were described in Chapter 3 (see Table 3.6). When applied to the validation carried out in this research, Table 3.6 should be read as follows; *Sensitivity* refers to the proportion of members of 'ethnic group X' (gold standard) who were correctly classified as such; *specificity* to the proportion of members of the 'rest of ethnic groups'(gold standard) who were correctly classified as such; *Positive Predictive Value* (PPV) is the proportion of persons classified as 'ethnic group X' (predicted) who were actually from ‘ethnic group X’; *Negative Predictive Value* (NPV), is the proportion of persons classified as the ‘Rest of ethnic groups’ (predicted) who were actually from the ‘Rest of ethnic groups’. These measures are all usually represented as proportions between 0 and 1, and calculated as explained in Mateos (2007).

These classification evaluation measures were calculated for the matrix shown in Table 9, removing the column ‘9- Not Stated’ for the reasons explained above. This offered the base values for sensitivity, specificity, PPV and NPV, which are shown as the minimum values in each range reported in Table 10 (the value on the left of each pair). However, in order to obtain a full range of possible values under different assumptions, further calculations were carried out to assess the effect in the overall measures. The same calculation for the four measures was repeated but now removing the column ‘8 - Other ethnic groups’ from the HES dataset, since this is a ‘catch-all’ category and is also deemed to contain a lot of data entry errors in the hospital admission process (London Health Observatory, 2005). This result is not shown here but lies within the range of values shown in Table 10, and its overall prediction success is 85.4%.

| ***1991 Census Categories*** | | ***Sensitivity*** | ***Specificity*** | ***PPV*** | ***NPV*** |
|---|---|---|---|---|---|
| 0 | White | **0.93 - 0.98** | 0.58 - 0.62 | **0.82 - 0.90** | **0.82 - 0.89** |
| 1 | Black - Caribbean | 0.02 - 0.03 | **1.00 - 1.00** | 0.51 - 0.62 | **0.96 - 0.96** |
| 2 | Black - African | 0.38 - 0.45 | **0.98 - 0.99** | 0.62 - **0.76** | **0.96 - 0.96** |
| 3 | Black - Other | n/a | n/a | n/a | n/a |
| 4 | Indian | 0.48 - 0.52 | **0.98 - 0.99** | 0.36 - 0.50 | **0.99 - 0.99** |
| 5 | Pakistani | 0.63 - **0.70** | **0.96 - 0.97** | 0.09 - 0.12 | **1.00 - 1.00** |
| 6 | Bangladeshi | 0.66 - 0.69 | **0.99 - 0.99** | 0.68 - **0.79** | **0.99 - 0.98** |
| 7 | Chinese | 0.60 - **0.73** | **1.00 - 1.00** | 0.62 - **0.80** | **1.00 - 1.00** |
| 8 | Any other ethnic group | 0.18 | **0.97** | 0.49 | **0.88** |
| *9* | Not Given | n/a | n/a | n/a | n/a |

**Table 10: Sensitivity, Specificity, PPV and NPV of the CEL classification based on 1991 Census Categories**

Ranges represent the minimum values, taking into account all the HES_person records, and the maximum one, removing cases with conflicting ethnicities including ‘8- Other’ or ‘9-Not Stated’. Values highlighted in bold are ≥0.7, and represent stronger classificatory power.

In a third scenario, a new matrix similar to Table 9 was generated but in this case removing those patients whose ethnic codes in HES did not originally match, but for whom the conflicting codes disappear after removing the codes as ‘8- Other’ or ‘9-Not Stated’ (as pointed out in Section 6.4 paragraph c). This subset of patients accounts for 9.8% of the total, and is also deemed to have an assigned ethnic group of dubious quality in HES. Therefore, the new matrix only includes patients for whom 100% of the ethnic codes through various admissions matched, that is, 207,538 patients (60.5% of the total). The four evaluation measures were calculated again for this new matrix, offering improved results. Finally, an additional calculation was re-run by removing from this second matrix columns ‘8’ and ‘9’ (not rows) as done with the previous matrix, which left a population of 177,419 patients (51.7%), whose ethnic group codes in HES matched and who were not coded ‘8’ or ‘9’, hence only ‘0’ to ‘7’. Therefore, this is the dataset with the highest quality in HES, and will provide the maximum value of the evaluation measures, which are reported as the top of the ranges in Table 10. The results of this table as well as of the next tables are discussed in Section 6.6.

As mentioned before, although hospitals have been required to code the ethnicity of patients following the 2001 Census classification into 16 ethnic groups since April 2001 (NHS Information Authority, 2001), this has actually taken several years to implement (London Health Observatory, 2005). During this time a combination of both the 1991 and 2001 Censuses ethnicity classifications have been used. However, in the case of the Camden and Islington HES dataset, where a 2001 Census ethnic category was available (178,623 patients or 52% of the total), the process described in this section of generating two binary classification matrices and calculating the four evaluation measures was repeated. As a result, a new set of value ranges of sensitivity, specificity, PPV and NPV was calculated for the 16 ethnic groups and is summarised in Table 11.

| | ***2001 Census Categories*** | ***Sensitivity*** | ***Specificity*** | ***PPV*** | ***NPV*** |
|---|---|---|---|---|---|
| A | White - British | **0.77** | **0.71** | **0.70 - 0.74** | **0.75 - 0.78** |
| B | White - Irish | 0.60 - 0.61 | **0.92** | 0.22 - 0.23 | **0.98** |
| C | White - Any other White | 0.43 | **0.91 - 0.93** | 0.41 - 0.51 | **0.91** |
| D | Mixed - White and Black | n/a | n/a | n/a | n/a |
| E | Mixed- White and Black | n/a | n/a | n/a | n/a |
| F | Mixed- White and Asian | n/a | n/a | n/a | n/a |
| G | Mixed- Other Mixed | n/a | n/a | n/a | n/a |

| | | | | | |
|---|---|---|---|---|---|
| H | Asian - Indian | 0.53 | **0.98 - 0.99** | 0.39 - 0.46 | **0.99** |
| J | Asian - Pakistani | 0.65 | **0.96** | 0.09 - 0.11 | **1.00** |
| K | Asian - Bangladeshi | 0.66 | **0.99** | **0.72 - 0.77** | **0.98** |
| L | Asian - Any other Asian | 0.0004 | **1.00** | **1.00** | **0.98** |
| M | Black - Caribbean | 0.03 | **1.00** | 0.52 - 0.56 | **0.96** |
| N | Black - African | 0.38 | **0.98 - 0.99** | 0.66 - **0.74** | **0.94 - 0.95** |
| P | Black - Any other Black | n/a | n/a | n/a | n/a |
| R | Chinese | 0.63 | **1.00** | 0.63 - **0.72** | **1.00** |
| S | Any other ethnic group | 0.20 | **0.96** | 0.36 | **0.92** |
| Z | Not Stated | n/a | n/a | n/a | n/a |

**Table 11: Sensitivity, Specificity, PPV and NPV of the CEL classification based on 2001 Census Categories**

Ranges represent the minimum values, taking into account all the HES_person records, and the maximum one, removing cases with conflicting ethnicities including '8- Other' or '9-Not Stated'. Values highlighted in bold are ≥0.7, and represent stronger classificatory power.

### 6.5.2 Data Analysis: evaluating differences in the CEL classification by gender

Another aspect of the CEL name classification that was evaluated was the degree to which its classificatory power diminished when applied to names of females, since many women change their surname after marriage and this is one of the critiques often made of name origin techniques. In a study of Chinese names, Quan *et al* (2006) found that the overall population PPV of 80.5%, decreased to 78.9% for married women. In order to assess the differential ability of the CEL name classification to correctly identify ethnicity by gender, the same exercise described in the previous Section 6.5.1 was repeated separately but only for the male population.

The hypothesis to test is that if the CEL classification is very sensitive to the gender of the population classified, then if it is only applied to the male population the classification ability to correctly assign ethnicity should significantly improve compared with the total population. However, the results proved that when only applied to men the classification showed similar values of sensitivity, specificity, PPV and NPV than for the overall population, especially using the 1991 Census classification, with small differences between -0.03 to 0.03 absolute points (in the 0 to 1 scale) and showing no particular direction. In the case of the 2001 Census classification, differences in the four measures were between -0.05 and 0.05 absolute points, except for PPV where they were between 0 and 0.11 positive points.

However, a differential performance of the CEL classification by ethnic group is also observed in the 2001 male dataset, with the three White groups (A, B, C), Indian (H), and Chinese (R) groups, showing substantially higher values of increase in PPV for males between 0.06 to 0.11 absolute points, when compared with the overall population. The causes that might explain these differences include: a differential behaviour of ethnicity reporting by gender in HES; the problem of small numbers when taking only 2001 Census male patients in HES subdivided by ethnic groups (giving sizes between 200 and 3,000 people per group); the specific gender and ethnic group composition of the population in Camden and Islington; and a component of classification errors with names of women in mixed ethnicity marriages, which are deemed to be higher amongst the five ethnic groups aforementioned.

## 6.6 Discussion of evaluation results

The sections above have described the process carried out to validate the CEL name classification by applying it to a population of 343,068 people admitted to hospital over 8 years in Camden and Islington, and comparing it with the patient reported ethnicity. The results of this validation are summarised in Table 10 and Table 11, where the CEL classification is compared with the actual reported ethnicity using either the nine 1991 Census categories for all the patients, or the sixteen 2001 Census categories for a subset of them (52%), giving a range of values obtained for the measures of sensitivity, specificity, PPV and NPV under different scenarios.

Sensitivity and positive predictive value (PPV) are two statistical measures of how well a binary classification test correctly classifies (sensitivity) or predicts (PPV) cases belonging to their class, in this case an ethnic group, while specificity and negative predictive value (NPV) measure its inverse, that is, cases *not* belonging to that class (Altman and Bland, 1994a; Altman and Bland, 1994b). Table 10 and Table 11 show that the validation of the CEL name classification achieves very high values of specificity and NPV (most of the ethnic groups with values above 0.90), while its sensitivity and PPV present varied results by ethnic group (between 0.50 and 0.90). This result is a direct consequence of one of the main aims of this paper research: ‘to classify entire populations into all of the potential ethnic groups present in a society’, that imposes an objective of maximising population coverage at the cost of increasing errors in the classification. An alternative would have been to just classify the few thousand names that most accurately represent a cultural ethnic or linguistic group, leaving all the other names unclassified. This would have maximised precision at the cost of a low coverage.

However, reflecting on the results shown on Table 10 and Table 11, it can be noticed that the CEL classification achieves an overall high accuracy in the 'White-British', 'Pakistani', 'Bangladeshi', 'Black African' and 'Chinese' groups, with values over 0.7 or 70%, which are all groups well represented in the study area. On the other hand, the ethnic groups where the CEL classification proves less effective are 'Black Caribbean' (because the majority of the names are of British origin), 'Any other ethnic group' (a 'catch-all' category of dubious value) and 'White-Other' (a mix-match category embracing half of the world: (Connolly and Gardener, 2005). Furthermore, as is obvious from the conception of the CEL classification, the name approach is not able to identify persons who assign themselves to one of the four mixed ethnicity groups (D-G) of the 2001 Census categories, or the vague 'Any other Black background' (P).

Amongst the main factors that explain the results obtained, the following can be mentioned:

- The gold standard for ethnicity used, the 'self-reported' ethnicity of hospital admission records, is of very low quality (Aspinall and Jacobson, 2004), and even after the efforts made here to remove the more dubious cases (i.e. conflicting ethnic groups for the same person), there are known cases of ethnicity being assigned by nurses or administrative staff without direct patient consultation.
- The CEL name classification does not identify 'mixed ethnicity' through names, and hence the CEL allocation of persons who have reported mixed ethnicity, as well as the 'Other' categories using the Census classification in HES, cannot be compared with the gold standard like with like. However, the CEL classification does provide much finer detail by cultural, ethnic and linguistic group not present in the Census categories, as shown in Table 9, column '9', where it improves the 'Not Given' responses in the HES dataset, identifying 96% of them.
- The mapping between the 66 CEL Subgroups and the 9 or 16 Census ethnic groups is not perfect, since the essence of each of the classes is radically different and hence leads to different ontologies of ethnicity. This in turn has an impact in the evaluation comparing the two.

- The CEL name classification has been built to maximise population coverage at a UK National level, while the HES dataset represents a very specific spatio-temporal section of the UK population; people admitted to hospital in Camden and Islington between 1998 and 2006. The opposite problem has been reported in the *Nam Pechan* name classification, which was built for an area in Bradford and when applied to other areas in the country proved less effective (Cummins et al, 1999). In the validation presented here, the situation is the opposite, a nationally designed classification that although having been just validated on the particular hospital population of Camden and Islington, ought to perform well when applied to other regions.
- Finally, errors that remain after controlling for the above would have to be explained by the differential ability of people's names origins to manifest current conceptions of ethnic groups, as applied through the methodology presented in this paper.

# 7 Conclusion

The subdivision of populations according to ethnicity and geography has allowed social scientists to gain better understandings of contemporary society and neighbourhoods, as populations and cities have become increasingly multi-culturally diverse and globally connected. However, there is a desperate requirement to improve the depth of such understandings, especially the complex processes of population composition and change by ethnic group and small area. New methods are required which might be adapted to rapid changes in international migration and ethnic group formation processes. Such improved methods will prove key in informing policy to reduce ethnic inequalities, produce accurate population statistics and plan for the future complex needs of our societies and cities.

This paper has sought to make a contribution to these methodological requirements, by developing an ontology of ethnicity based on the classification of names according to their cultural, ethnic and linguistics groups. This has been termed the CEL classification. The steps to develop and validate this methodology have been fully described in a robust and transparent manner, and its results are available for other researchers to use and enhance. This paper has illustrated one of many possible applications to a classical geographical problem of current relevance to public debates, namely the study of residential segregation. It has also presented a small gallery of applications, in order to illustrate the very wide potential applicability of the CEL classification. Application of the CEL methodology to different research settings and contexts offers one way of improving our understandings of contemporary society and neighbourhoods, and these in turn will allow wider validation of the classification to take place.

There is evidence today that names are unfortunately still being used to discriminate against people's abilities to access the labour, housing, and credit markets (Carpusor and Loges, 2006; Williams, 2003), because of the prejudices that some retain about people's ancestry, language, religion, culture, or skin colour. Yet it is in using the same weapons as the 'enemy', in the 'The Causes and Consequences of Distinctively Black Names', that Fyer and Levitt (2004) develop a (albeit crude) picture of ethnic inequalities and discrimination in the US through an innovative analysis of forenames. A golden opportunity would be missed if social science researchers eschew a creative opportunity to find new ways of reducing persistent discrimination and inequalities between ethnic groups in today's ever increasingly multi-cultural cities. It is hoped that the methodology developed in this paper will assist them in this difficult task.

# 8 References

Altman DG and Bland JM. 1994a. Statistics Notes: Diagnostic tests 1: sensitivity and specificity. *BMJ* **308**(6943): 1552

Altman DG and Bland JM. 1994b. Statistics Notes: Diagnostic tests 2: predictive values. *BMJ* **309**(6947): 102

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

Aspinall PJ and Jacobson B. 2004. *Ethnic Disparities in Health and Health Care: A focused review of the evidence and selected examples of good practice.* London Health Observatory. Available at: http://www.lho.org.uk/viewResource.aspx?id=8831. Accessed: 20/07/2006.

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.

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

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

Carpusor AG and Loges WE. 2006. Rental Discrimination and Ethnicity in Names. *Journal of Applied Social Psychology* **36**(4): 934-952

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

Connolly H and Gardener D. 2005. *Who are the 'Other' ethnic groups?* Social and Welfare reports. Office for National Satistics. London. Available at: http://www.statistics.gov.uk/articles/nojournal/other_ethnicgroups.pdf. Accessed: 27/01/2006.

Cummins C, Winter H, Cheng K-K, Maric R, Silcocks P, et al. 1999. An assessment of the Nam Pehchan computer program for the identification of names of south Asian ethnic origin. *Journal of Public Health Medicine* **2**(4): 401-406

Department of Health. 2005. *A Practical Guide to Ethnic Monitoring in the NHS and Social Care.* Available at: http://www.dh.gov.uk/assetRoot/04/11/68/43/04116843.pdf. Accessed: 23/09/2005.

Fryer RG and Levitt SD. 2004. The Causes and Consequences of Distinctively Black Names. *The Quarterly Journal of Economics* **119**(3): 767-805

Gerrish K. 2000. Researching ethnic diversity in the British NHS: methodological and practical concerns. *Journal of Advanced Nursing* **31**: 918-925

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

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

Haut Conseil à l'Integration. 1991. *Pour un modèle francais d'integration.* Premier Rapport Annuel. La Documentation Francaise. Paris.

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

Lauderdale D and Kestenbaum B. 2000. Asian American ethnic identification by surname. *Population Research and Policy Review* **19**(3): 283-300

Leppard D. 2005. Race chief warns of ghetto crisis. *The Sunday Times* September 18, .

London Borough of Camden. 2007. *Camden Profile.* Available at: http://www.camden.gov.uk/ccm/cms-service/stream/asset/?asset_id=576779. Accessed: 08/06/2007.

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.

Majeed A, Bardsley M, Morgan D, O'Sullivan C, Bindman AB. 2000. Cross sectional study of primary care groups in London: association of measures of socioeconomic and health status with hospital admission rates. *British Medical Journal* **321**(7268): 1057-1060

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.

Martineau A and White M. 1998. What's not in a name. The accuracy of using names to ascribe religious and geographical origin in a British population. *Journal of Epidemiology and Community Health* **52**(5): 336-337

Mason D. 2003. *Explaining ethnic differences: changing patterns of disadvantage in Britain*. Bristol: Policy Press.

Mateos P. 2007. A Review of Name-based Ethnicity Classification Methods and their Potential in Population Studies. *Population Space and Place* **13**(4): 243-263

Mateos P, Webber R, Longley PA. 2007. *The Cultural, Ethnic and Linguistic Classification of Populations and Neighbourhoods using Personal Names.* CASA Working Paper 116. Rep. ISSN 1467-1298, Centre for Advanced Spatial Analysis. University College London. London. Available at: http://www.casa.ucl.ac.uk/working_papers/paper116.pdf. Accessed: 05/03/2007.

McAuley J, De Souza L, Sharma V, Robinson I, Main CJ, et al. 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

NHS Executive. 1994. *Collection of ethnic group data for admitted patients.* EL/94/77. NHSE. Leeds.

NHS Information Authority. 2001. *CDS, HES and Workforce: Ethnic data Finalised Coding Frame. DSC Notice: 02/2001.* Birmingham. Available at: http://www.connectingforhealth.nhs.uk/dscn/dscn2001. Accessed: 13/02/2006.

Office for National Statistics. 2003. *Ethnic group statistics: A guide for the collection and classification of data.* Available at: http://www.statistics.gov.uk/about/ethnic_group_statistics/downloads/ethnic_group_statistics.pdf. Accessed: 13/02/2006.

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

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.

Platt L, Simpson L, Akinwale B. 2005. Stability and change in ethnic groups in England and Wales. *Population and Trends* **121**: 35-46

Quan H, Wang F, Schopflocher D, Norris C, Galbraith PD, et al. 2006. Development and validation of a surname list to define Chinese ethnicity. *Medical Care* **44**(4): 328-333

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

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

The Economist. 2005. One man's ghetto. *The Economist*, 24th September: 16

The Economist. 2006. Hostility at home. 23rd November

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. 2007. Personal communication.

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.

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): 705-710

Williams A. 2003. Who will be hired: Stacey or Shakisha? *Journal of the National Medical Association* **95**(2): 109-110

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.