An ontology of ethnicity based upon personal names: with implications for neighbourhood profiling

Pablo Mateos

Department of Geography
University College London (UCL)

Thesis submitted in partial fulfilment of the requirements for the degree of:
Doctor of Philosophy (PhD)

July, 2007

Author's declaration

I, Pablo Mateos, confirm that the work presented in this thesis *'An ontology of ethnicity based upon personal names: with implications for neighbourhood profiling'* is exclusively my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis. This work was undertaken with the partial support of the Economic and Social Research Council (ESRC) and Camden Primary Care Trust who received a proportion of funding from the NHS Executive. The views expressed in this publication are those of the author and not necessarily of the NHS Executive nor of the ESRC or University College London.

Acknowledgements

I must thank, first of all, my supervisor Paul Longley, who has provided support, advice and guidance throughout this PhD. I will be always indebted to him for his constant encouragement and support from the moment I expressed an interest in studying a PhD at UCL. Furthermore, Paul's spirit, good mood and sense of humour in these three years have made the PhD experience most enjoyable.

Thanks must go to Camden Primary Care Trust (PCT) and the Economic and Social Research Council (ESRC) who funded this PhD through a Knowledge Transfer Partnership with UCL (KTP-037). The Public Health Intelligence team at Camden PCT was very supportive throughout the research. I am specially indebted to Richard Webber for his source of inspiration, input and supervision throughout this PhD. Through him I would like to thank *Experian* for providing free of charge part of the data that made this research possible. Many people gave me very valuable ideas, support or materials through my research, amongst them; Ken Tucker, Mario Cortina Borja, Kate Jones, Maurizio Gibin, Alex Singleton, and many others too numerous to mention here. I am grateful to you all.

I am also grateful to various anonymous referees who have provided useful feedback on aspects of the PhD that have been submitted for publication and for the helpful comments that have been made at conferences when material contained within this thesis has been presented and discussed.

I am also thankful to my colleagues at UCL Centre for Advanced Spatial Analysis (CASA), for having made me feel belonging to a big family in London, specially

through the irreplaceable experience of ‘CASA Wednesdays’, which were fundamental to the ideas and spirit of this PhD, as well as a vehicle for the many friendships made.

Finally, I would like to thank my friends, colleagues and family with whom I have shared different bits of the process of this PhD. Without the international support of my brother Ramón in getting me up to speed with Oracle databases I could not have coped with the massive name datasets involved in this project. My wife Brenda has been an amazing companion through these years, without whose support, patience and enthusiasm this PhD would have not been possible.

Abstract

Understanding of the nature and detailed composition of ethnic groups remains key to a vast swathe of social science and human natural science. Yet ethnic origin is not easy to define, much less measure, and ascribing ethnic origins is one of the most contested and unstable research concepts of the last decade - not only in the social sciences, but also in human biology and medicine. As a result, much research remains hamstrung by the quality and availability of ethnicity classifications, constraining the meaningful subdivision of populations.

This PhD thesis develops an alternative ontology of ethnicity, using personal names to ascribe population ethnicity, at very fine geographical levels, and using a very detailed typology of ethnic groups optimised for the UK population. The outcome is an improved methodology for classifying population registers, as well as small areas, into cultural, ethnic and linguistic groups (CEL). This in turn makes possible the creation of much more detailed, frequently updatable representations of the ethnic kaleidoscope of UK cities, and can be further applied to other countries.

The thesis includes a review of the literature on ethnicity measurement and name analysis, and their applications in ethnic inequalities and geographical research. It presents the development of the new name to ethnicity classification methodology using both a heuristic and an automated and integrated approach. It is based on the UK Electoral Register as well as several health registers in London. Furthermore, a validation of the proposed name-based classification using different datasets is offered, as well as examples of applications in profiling neighbourhoods by ethnicity,

in particular the measurement of residential segregation in London. The main study area is London, UK.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACL	Absolute Clustering Index
ACLS	American Council of Learned Societies
ACO	Absolute Concentration Index
AFPD	All Fields Postcode Directory
AIDS	Acquired Immune Deficiency Syndrome
AU	Australia
avg	Average
CA	Canada
CACI	Consolidated Analysis Centers Inc
CEL	Cultural, Ethnic and Linguistic group
CELG	Cultural, Ethnic and Linguistic Group (technique)
CLUTO	Clustering Toolkit
CSR	Compressed Row Format
DAFN	Dictionary of American Family Names
DNA	Deoxyribo-Nucleic Acid
EM	Ethnic Minority
ESDA	Exploratory Spatial Data Analysis
ESRC	Economic and Social Research Council
EU	European Union
F_List	Forename List
FCEL	Forename Cultural, Ethnic and Linguistic group
FHSA	Family Health Service Authority Register
FSC	Forename-Surname Clustering
GB	Great Britain
Gb	Gigabyte
GIS	Geographic Information System
GWR	Geographic Weighted Regression
H&F	Hammersmith and Fulham
HES	Hospital Episode Statistics
HIV	Human Immunodeficiency Virus
ID	Index of Dissimilarity
IDESCAT	Instituto De Estadistica de Cataluña (Catalan Statistical Institute)
IE	Republic of Ireland
ISO	International Standards Organisation
LA	Local Authority
LISA	Local Indicators of Spatial Autocorrelation
log	logarithm
LSOA	Lower level Super Output Area
MAUP	Modifiable Areal Unit Problem
MEUP	Modifiable Ehtnic Unit Problem

MS	Microsoft
NDTMS	National Drug Treatment Monitoring System
NHS	National Health Service
NI	Northern Ireland
NPV	Negative Predictive Value
NSPD	National Statistics Postcode Directory
NZ	New Zealand
OA	Output Area
OAC	Output Area Classification
ONS	Office for National Statistics
OS	Ordnance Survey
PCEL	Person Cultural, Ethnic and Linguistic group
PCT	Primary Care Trust
PLASC	Pupil Level Annual School Census
PPV	Positive Predictive Value
S_List	Surname List
SANGRA	South Asian Name and Group Recognition Algorithm
SAS	Statistical Analysis System
SCEL	Surname Cultural, Ethnic and Linguistic group
SMOBE	Survey of Minority-Owned Business Enterprises
SOA	Super Output Area
SOM	Self-Organising Maps
SOPHID	Survey of Prevalent HIV Infections Diagnosed
SPSS	Statistical Package for the Social Sciences
SURFING	SUbspaces Relevant For clusterING
TB	Tuberculosis
UK	United Kingdom of Great Britain and Northern Ireland
US	United States of America
z_tok	z-score based on Tokens
z_typ	z-score based on Types

Chapter 1. Introduction

'National identity requires a collective work of amnesia'

(Renan, 1990 [1882]: 11)

'It takes at least two somethings to create a difference (...) Clearly each alone is – for the mind and perception– a non-entity, a non-being (...) a sound from one hand clapping'

(Bateson, 1979: 78)

1.1. Ethnicity, Collective Identities and Multicultural Cities

The study of ethnicity in multicultural societies and cities is probably the most problematic phenomenon that social scientists face today. Ethnicity relates to a person's inner sense of collective identity, and – as the second quote above suggests – its definition requires contact between differently perceived groups to create a difference. Such contact has exponentially increased in the last decades as populations, cities and neighbourhoods are becoming increasingly multi-culturally diverse and globally connected (Castles and Miller, 2003). If – as stated in the first quote above from 19th century philosopher Renan – 'national identity requires a collective work of amnesia'(1990 [1882]: 11), it could be argued that in today's context of globalisation and erosion of 19th century nation-state identities, ethnic identity requires a collective work of 'remembrance and nostalgia'.

One of the multiple definitions of ethnicity states that '[a]n ethnic group is a collectivity within a larger population having real or putative common ancestry,

memories of a shared past, and a cultural focus upon one or more symbolic elements which define the groups' identity, such as kinship, religion, language, shared territory, nationality or physical appearance' (Bulmer, 1996: 35). However, the definition of ethnicity is controversial because ethnic identification is subjective, multi-faceted and changing in nature and because there is not a clear consensus on what constitutes an 'ethnic group' (Coleman and Salt, 1996; Office for National Statistics, 2003). Moreover, ethnicity classifications have become a key factor of political power in the growing arena of identity politics (Skerry, 2000). The power struggle between competing collective identities for institutional recognition through official ethnicity classifications is especially manifested at local level, where such recognition brings solutions for locally perceived problems, monetary resources, political representation, and benefits associated with positive discrimination initiatives (Kertzer and Arel, 2002).

However, the main purpose for which ethnicity started to be officially classified and measured in a number of developed countries in the last decades bore little correspondence with this identity politics struggle. It directly emanated from the need to monitor progress in equality legislation, introduced to prevent racial discrimination and reduce ethnic inequalities after the 1960s (Peach, 2000), in particular the American Civil Rights Movement (1954-1968) and the UK Race Relations Act (HM Government, 1976). Such legislation and the population classifications derived from them were only concerned with people seen of 'darker skin colour', following non-European post-war migration to America and Europe and the deeply rooted black discrimination in the US (Coleman and Salt, 1996). Today the much broader and cultural term of ethnicity is preferred to the biologically rooted

concept of ‘race’, to define and classify collective identities in increasingly multicultural populations.

The evidence of ethnic inequalities in most multicultural societies has grown strongly in the last decades (Mason, 2003; Nazroo, 2003a). One of the aspects in which such inequalities are manifest is in its spatial dimension, with debates about ethnic residential segregation and the ‘ghettoisation of society’ having acquired special prominence in the public debate of recent years (Dorling, 2005b; Phillips, 2005). Although a range of diverse and intertwined factors for such ethnic inequalities has been identified, research has fallen short of unveiling the true interaction between such factors, especially at the local micro-level (Karlsen et al, 2002). As Chapter 2 in this thesis will describe, the main problem has been a lack of availability of ethnicity data at sufficient quality and level of disaggregation, and an absence of adequate methods to interpret the problematic nature of measuring different ontologies of ethnicity. Therefore, new methods are required in the analysis of ethnic inequality in increasingly diverse populations and neighbourhoods, which are capable of being 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 and maintain accurate population statistics and plan for the future complex needs of our societies and cities.

This PhD aims to contribute to such methodological need. It contends that there is a strong relationship between the ethnic identities of human groups and their mother languages or those of their ancestors, and that an indication of these can be revealed by the analysis of personal name origins. This is the cornerstone of the

methodological innovation that this PhD aims to contribute: developing a new classification of populations and neighbourhoods along the multidimensional aspects of collective identity, through the cultural, ethnic and linguistic origins of personal names.

From the title of this thesis; 'An ontology of ethnicity based upon personal names: with implications for neighbourhood profiling', three main concepts arise: ontology of ethnicity, personal names, and neighbourhood profiling. These are the three core components around which the whole thesis is developed. Names are thus used to propose a new ontology and classification of ethnicity, developing an innovative methodological tool for investigating population ethnicity at neighbourhood level.

1.2. Aim and Objectives

The principal aim of this thesis is *to develop a new ontology and classification of ethnicity based on personal names origins*. This aim will be achieved by the following six objectives:

1) To investigate and review the methods and procedures involved in creating a name-based ethnicity classification at the level of the individual person.
2) To propose a detailed taxonomy of cultural, ethnic and linguistic groups (CEL) based on the common characteristics of the names present in Britain.
3) To develop a classification of all the surnames and forenames present in Britain with a frequency of three or more people into cultural, ethnic and linguistic groups (CEL classification).
4) To evaluate the CEL taxonomy and classification at the level of the individual person.

5) To demonstrate the value of the CEL classification through its application to the study of residential segregation in London at different geographical scales and ethnic group aggregations.
6) To provide a detailed description of the methodology in the creation of the CEL classification, in order to facilitate its future reproducibility.

The content of these six objectives will become clearer in each of the individual chapters in which they are tackled. However, it can be seen that the main purpose of this PhD research is of a methodological nature. Although the contextual and applied aspects of ethnic inequalities and residential segregation are present throughout the thesis, the distinctive contribution to knowledge of this thesis lies on the novel methodology developed to classify the whole population into ethnic groups using their names' origins.

1.3. Methods and Outputs

Because of the methodological essence of this PhD thesis, the research methodology employed is actually the end in itself, rather than the successful application of established methods. Therefore, and following objective (1) above, this thesis first investigates and reviews existing methods involved in creating a name-based ethnicity classification at the level of the individual person (Chapters 3 and 4). Most of the similar studies found in the literature come from the public health and population genetics literature, which use simple statistical methods applied to very large population datasets at the level of the individual person, following a 'data mining' approach.

Based on these findings, the thesis then describes the steps taken in the actual development of the names classification developed in this research, termed the

cultural, ethnic and linguistic (CEL) classification (Chapters 4, 5 and 6). These steps can be summarised as the following general methodological processes:

1- Develop a taxonomy of cultural, ethnic and linguistic groups
2- Collect and prepare several different names datasets with differing content and coverage
3- Select a suite of name classification techniques
4- Develop an exploratory names classification based on heuristic rules
5- Develop the CEL names classification following an automated and integrated approach based on forename-surname clustering
6- Validate the CEL names classification

The main output of this research is a new classification of names into cultural, ethnic and linguistic groups (CEL). Such classification is comprised of two tables, a surname-to-CEL table including 225,576 surnames, and a forename-to-CEL table including 98,624 forenames, each alongside their most likely CEL and a measure of the degree of association between the name and the CEL. The intention is to make these tables available on request to bona-fide academic researchers for further evaluation and enhancement. Other outputs from this research have been three publications in peer reviewed journals, one working paper, and twelve paper presentations at national and international conferences. For the details of these outputs see Appendices 1 and 7.

1.4. Thesis Structure

The thesis is organised in nine chapters. Each of the seven core chapters addresses one or two of the six objectives specified in Section 1.2. Table 1.1 describes the objectives that are addressed by each chapter.

Chapter	Objective
.1- Introduction	
.2- Concepts and measurements of ethnicity	1
.3- Names and ethnicity	1
.4- Taxonomy, materials and methods	1, 2
.5- Heuristic approaches to creating a name classification	1, 3
.6- An automated and integrated approach to name classification	3, 6
.7- Validating the CEL name classification	4, 6
.8- Applications: residential segregation and ethnic inequalities	4, 5
.9- Summary and conclusion	6

Table 1.1: Thesis structure and correspondence between chapters and objectives

Chapter 2 – 'Concepts and Measurements of Ethnicity' – sets out the general context of the different intersections between the ontology of ethnicity and its measurement, and debates of ethnic inequalities and residential segregation. The first part justifies the study of the classification of populations and neighbourhoods according to ethnicity, primarily in the context of British ethnic inequalities. The second part addresses the ontological issues behind the concepts of ethnicity and their measurement, and investigates how they affect the analysis of ethnicity. As a whole, the chapter justifies the need for new alternative methods to study and measure ethnicity, setting the ground for the developments of the rest of the thesis.

Chapter 3 – Names and Ethnicity – reviews in depth the multidisciplinary literature of name-based ethnicity classifications with a view to summarising the different approaches that have been independently developed, to identifying the research gaps, and to setting out a research agenda for potential contributions using the improved methods that will be described and developed in the rest of the thesis.

Chapter 4 – Taxonomy, Materials and Methods – introduces the three integrated components of the first phase in the development of a new name-based ethnicity classification: taxonomy, materials and methods. It formalises a new taxonomic classification of names based upon cultural, ethnic and linguistic (CEL) groups. It then discusses the potential universal name register data sources and presents the materials finally selected for use in this research. Finally, it describes the methodologies that are subsequently used to put the three components together; methods to classify the universal list of names into the CEL taxonomy.

Chapter 5 – Heuristic Approaches to Creating a Name Classification – describes the initial heuristic approach, which laboriously classified different groups of names into CELs following the techniques described in Chapter 4, in the order in which they were first investigated. These techniques were specified through different rules and applied to different stages in a dynamic and iterative process. Because of the cumulative yet fundamentally exploratory nature of this approach, it is described as ‘heuristic’.

Chapter 6 – An Automated and Integrated Approach to Name Classification – describes how an enhanced methodology was developed in what constitutes an

automated approach to classification of names into CELs. It builds upon the analysis of the limitations and achievements recognised in the heuristic approach, and sets out the requirement to develop a transparent and reproducible method. This automated approach has the objective of providing a simple and systematic method that can be easily explained and understood, and allows third parties to understand the explicit procedures that were used to develop the classification.

Chapter 7 – Validating the CEL Name Classification – explains the validation process performed 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. The validation is carried out against some populations for whom ethnicity is already known through an independent source (i.e. not based on names). More than one way to validate the effectiveness of the CEL classification is provided; a classification of individuals through a hospital admission register, and a classification of neighbourhoods, using 2001 UK Census ethnicity at several geographical scales.

Chapter 8 – Applications: Residential Segregation and Ethnic Inequalities – illustrates the potential applications of the CEL name classification to issues surrounding neighbourhood profiling and residential segregation debates. Given its high relevance to current debates in contemporary society, the CEL classification is applied to the study of ethnic residential segregation in London. Other potential applications are briefly described constituting a small gallery of applications, in order to illustrate the very wide potential applicability of the CEL classification.

Finally, the thesis closes with a concluding chapter (9) wrapping up the evidence gathered through the PhD and pointing out to promising avenues for future research in this area.

Chapter 2. Concepts and Measurements of Ethnicity

In the last decade and a half, there has been an explosion of interest in issues of ethnicity, nationalism, race and religion, around a renewed preoccupation with the question of defining and asserting collective identities. This trend has contradicted the prediction made in 1920s by Max Weber (1980 [1921]) who stated that 'primordial phenomena' such as ethnicity and nationalism would decline in importance and eventually vanish as a result of modernisation, industrialisation and individualism. On the contrary, the change of Millennium has brought opposition between an ever-expanding globalisation and an upsurge in identity, an antagonism that is key to understanding the way that our world and our lives are being shaped (Castells, 1997). Collective identities are formed and expressed as a resistance movement to cultural homogenisation (Castells, 1997) in a struggle for political power in multicultural societies (Kertzer and Arel, 2002). This is set in a context of a diminishing role of the nation-state, with political power being devolved to the regions and cities as well as taken away by international institutions and a new global order. Combined with these trends, long-established 19th century national identities are being eroded in an era of migration, characterised by the increased intensity and complexity of its flows (Castles and Miller, 2003).

In such circumstances, governments and social scientists have struggled to keep track of the reality of a rapidly changing population that is constantly re-defining its collective identities (Skerry, 2000). Although highly contested, the practice of classifying the population into discrete groups according to race, ethnicity or religion has made a strong re-appearance in many countries' recent national censuses

(Howard and Hopkins, 2005; Kertzer and Arel, 2002; Nobles, 2000). Such questions in the censuses not only quantify the size and geographical extent of collectively pre-perceived racial, ethnic and religious groups, but more interestingly help to reinforce the self-identity of those groups or accelerate the emergence of new identities (Christopher, 2002) by solidifying transient labels (Howard and Hopkins, 2005).

Because of the subjective nature of collective identities, the categorization process (the problematic definition of ethnic groups' boundaries and labels) has been a significant issue in social science (Peach, 1999b). Following an impassionate debate around the essentialism of ethnicity labels (Modood, 2005), there seems to be a consensus, at least in the demographic and public health literature, that the classification of the population into ethnic groups has proved useful to fight discrimination and entrenched health and social inequalities (Bhopal, 2004; Mitchell et al, 2000). There is a vast literature that demonstrates the persistence of stark inequalities between ethnic groups, especially in terms of health outcomes, access to housing and labour markets, educational outcomes and socioeconomic status (Frazier et al, 2003; Mason, 2003). As long as such inequalities between population subgroups persist, no matter how these are defined or perceived, the use of ethnic group definitions and labels will be useful to denounce them and fight against their causes. However, many of the current ethnicity classification practices have proved very inappropriate to uncover the true nature of specific factors of inequalities.

This chapter sets out the general context of these processes and the different intersections between the ontology of ethnicity and its measurements, and debates of ethnic inequalities and residential segregation. The first section, titled 'The

Geography of Ethnic Inequalities', describes the evidence in ethnic inequalities and population classification by ethnicity and geographical areas and summarises the current research shortcomings. The second section, 'Neighbourhood Profiling and the Segregation Debate', reviews in depth issues of ethnic minorities' uneven residential distributions as identified using different measurement metrics, and through synthesis of past debates identifies the main research gaps. Having justified the study of ethnic inequalities and the classification of populations and neighbourhoods according to ethnicity, the third section, 'Defining Ethnicity and Race', directly addresses the ontological issues behind the concepts of ethnicity and its multidimensional characteristics, taking a broad perspective drawn from the anthropological, sociological, geographical and health literatures. The fourth and final section, 'Measuring Ethnicity', complements the ground laid down in the previous section with an extensive review of the different ways in which ethnicity is measured in different contexts, identifies the key issues of measurement and investigates how they affect the analysis of ethnicity. Finally, a conclusion weaves together the main points and the research challenges established throughout the chapter, providing a firm justification for the further study of name origin analysis in the following chapter.

2.1. The Geography of Ethnic Inequalities

'[Recent ethnic disturbances] might be said to fit into a long-established pattern in the development of policies to address ethnic disadvantage in Britain, that is, the tendency, after a period of public hand wringing and spate of policy initiatives, for the issue of ethnic inequity to disappear from the agenda for a period, before

dramatically being forced back on – not infrequently by events on the streets'. (Mason, 2003, 1)

2.1.1. Ethnic inequalities

There is a vast literature that denounces the persistence and growth of social inequalities in the developed world (Dorling and Rees, 2003). Of particular relevance to this PhD, studies of the geographical aspects of such inequalities have demonstrated over and over again that there is a direct relation between where different social groups live and their overall chances in life, their level of income and outcomes in terms of health, education and material deprivation.

In the last twenty years or so, the way these 'social groups' have been defined has evolved to embrace a variety of social dimensions and not just the classic classification along socio-economic or social class criteria. Social inequalities have thus been analysed along lines of difference and identity, concerned with even starker inequalities according to gender, age, and ethnicity, a classic demographic triad (Mateos and Webber, 2006), in addition to other aspects such as disability and sexual orientation. Of relevance to this PhD is thus the literature that in this respect tackles the geography of ethnic inequalities in the developed world, especially in Britain. A detailed reflection on the concept of ethnicity is offered in Section 2.3.

Ethnic inequalities may be defined as the differential outcomes and processes of social disadvantage experienced between ethnic groups in a society where they co-exist. Differential outcomes by ethnic group have been repeatedly demonstrated along the following dimensions of quality of life: health (mortality and morbidity: Bhopal, 1997; Nazroo, 1997), education attainment and participation (Johnston et al,

2004; Parsons et al, 2004), housing (Boal, 2000), employment (Carter et al, 1999; Karn, 1997), material deprivation and social mobility (Loury et al, 2005). Furthermore, sustained differential access to and experience of public services, such as in healthcare, schooling, policing, social housing, transportation, social services, recreational activities, access to amenities, etc have been identified as being highly related to discrimination processes, whether passive or active, both at the individual and institutional levels. These differentials clearly show that members of some ethnic minority groups have a shorter life expectancy, are more likely to report bad health, have higher unemployment rates, lower wages, lower level of political participation, lower access to higher education, and live in more deprived areas than the national average. Special care needs to be put into not stigmatising ethnic minorities with all these conditions of disadvantage, as if they all experience these inequalities, or in the same way, or as if the mere fact of 'their difference' implied the cause of their condition of disadvantage. Past research has unfortunately assigned ethnic minorities with an accumulation of conditions that 'deviated from the norm' (i.e. the white majority), reinforcing an 'us and them' view of ethnicity based on assimilationist assumptions, that has consistently informed social policy in Britain (Mason, 2000).

Unfortunately, the literature has not been very successful in identifying and isolating each of the individual causes that lie behind the inequalities found. More often, racial discrimination, socio-economic, geographic, environmental, demographic, cultural, lifestyle, migration, historic, and genetic factors are closely intertwined in producing and reproducing the patterns of disadvantage by ethnicity that have been commonly observed in our cities and societies (Mason, 2003). Disentangling the separate causes and confounding factors of inequalities is only possible by measuring ethnicity in all

datasets where those same health, educational, employment or other socio-economic outcomes are recorded. This has only just started to happen in the last ten years, but as this PhD will discuss there is insufficient data and methods to analyse ethnic inequalities in the way that is required. As a starting point we will review how this has been attempted in explaining ethnic inequalities in health.

2.1.2. Ethnic inequalities in health

In no other area are ethnic inequalities more strikingly manifested than in health, where they have been extensively studied and denounced for at least over three decades in Britain. This area will be briefly summarised here as an example of the relevance of the disadvantages described in the previous subsection, and the problems in ascribing causality to observed outcomes.

Ethnic inequalities in health form part of the wider research field of health inequalities that has a long history and established methods, mainly in the epidemiological and public health literatures. Accounts of ethnic inequalities in health caused by socioeconomic disadvantage in Britain can be traced back in time to 1845 when Engels ascribed the poor health and high mortality rate of Irish migrants living in England to their poor social circumstances (Engels, 1987). The British public health literature has intensively analysed differential health outcomes by ethnic group, initially only for immigrants (Marmot et al, 1984; Wild and McKeigue, 1997), and in the last twenty years for different ethnic minority groups (Nazroo, 1997; NHS Health and Social Care Information Centre, 2005; Szczepura, 2005). In the US the literature has focused primarily on health inequalities between the Black population and the White majority (Gutgesell et al, 1981; Smith et al, 1998), with growing attention to other ethnic groups, such as Asian groups, Hispanics, and Arabs

(Chiapella and Feldman, 1995; Lauderdale and Kestenbaum, 2002; Lipson et al, 1987; Polednak, 1993).

Ethnic inequalities in health have become an increasing focus of research attention in Britain since the early 1970s, partly reflecting a growing public policy concern with the health of, and quality of healthcare provided for ethnic minority groups (Nazroo, 2003a). Research in this field was initially concerned with the health of immigrants and their children (Marmot et al, 1984; Oppé, 1967). However, it has experienced a continuous growth as ethnic groups have emerged, providing a wide range of evidence on the differential levels of morbidity and mortality between ethnic groups, specifically between the so-called ethnic minorities and the 'ethnic majority' (Aspinall and Jacobson, 2004; Bhopal, 2007). Figure 2.1 shows the odds ratio of people that have reported bad or fair health in the Health Survey for England by ethnic group, compared with the White British majority. It is striking to notice that the odds of people from groups such as the Caribbean, Indian, Bangladeshi or Pakistani ethnic groups that report bad or fair health is 2 to 3.2 times higher than that of the White British majority.

However, when considering unequal health outcomes it is surprising how common it still is to observe the stigmatisation of ethnic minorities. For example, in a report for a UK programme called '*Tackling Health Inequalities: A Programme for Action*' (Department of Health, 2003), ethnic minorities were lumped together into a 'vulnerable population group' that includes people '… living in remote and rural communities, as well as teenage parents, vulnerable older people, black and minority ethnic groups, looked after children and care leavers, homeless people, asylum

seekers and prisoners'(Department of Health, 2003, 10). This striking definition is very symptomatic of the fact that the very specific stereotypes that Bhopal (1997) denounced as the racist concept of a 'package of specific ethnic diseases', have not yet gone away even in government reports. Moreover, this definition supports the claim of Nazroo (2003b) that there is a common implicit and mistaken assumption that ethnic minorities are uniformly disadvantaged, leading researchers in the US to use 'race' as a surrogate for poverty in many analyses.

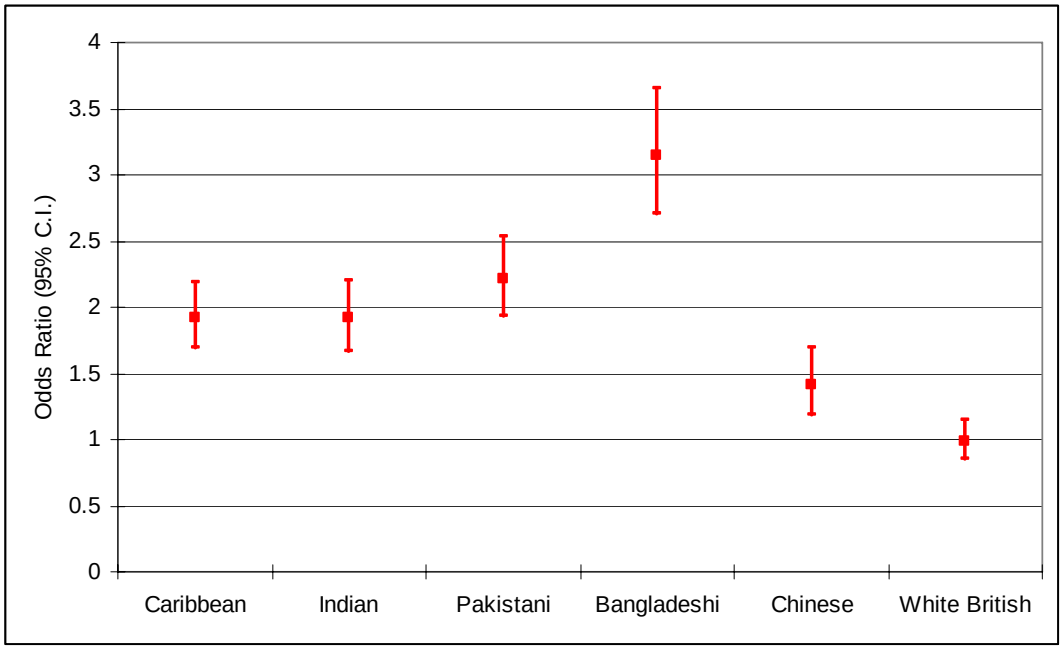

Figure 2.1: Health Survey for England reported 'Fair or Bad Health' by ethnic group
Age and gender adjusted odds ratio and 95% confidence interval compared with White British
Source: Health Survey for England 1999, reported in Nazroo (2003a)

As these examples from the study of ethnic inequalities in health show, there is a clear need to subdivide populations by ethnic group, and to access such disaggregated data from a wide range of social studies and government datasets in order to compare results, isolate the causes for such inequalities for each ethnic group, and disentangle the facts from the stereotypes. This demand for better data sources has been partially addressed since the UK 1991 Census of Population, yet,

sixteen years on, ethnicity is still not collected in most government datasets today, with evidence collection lagging far behind policy in this area. As will be explained in Section 2.4., research on ethnic inequalities has been hampered by a lack of data sources on ethnic groups and common methodologies in their collection and treatment (Aspinall, 2000; Whitehead, 1992).

Finally, although geographers have also been highly concerned with the spatial aspects of such health inequalities in the population (Boyle et al, 2004; Gatrell et al, 2000), they have typically not engaged with the study of ethnic inequalities in health, with the exception of studying migrant populations' health (Boyle, 2004). For example, in the landmark four volume series 'Ethnicity in the 1991 Census', written mostly by geographers after the ethnicity question was introduced in 1991 (Coleman and Salt, 1996; Karn, 1997; Peach, 1996b; Ratcliffe, 1996), there is not a single chapter devoted to ethnic inequalities in health, while there are several dedicated to employment and the labour market, education, housing, segregation and demographic aspects of ethnic groups. This is unfortunate, because geographers and health researchers have much to gain from mutual collaboration in the study of ethnic inequalities, since both collectives would benefit from methods developed independently in each field. This PhD intends to make a small contribution to this gap, adopting an interdisciplinary approach to expand a methodology to subdivide populations by ethnic group initially developed in the public health domain, applying it to geographical problems, and promoting new cross-collaborations in this field.

2.1.3. Subdividing populations by ethnicity and geography

Ethnic inequalities in the population vary starkly between and within ethnic groups, not only between ethnic minority groups and the ethnic majority, but also between

and within the ethnic minorities (it could also be argued that they vary within the ethnic majority as well, but this falls outside the scope of this PhD). As pointed out in the previous subsections, such variation is very difficult to explain, because of the complex interaction of many confounding factors at the level of the individual, neighbourhood, and ethnic group. One approach to disentangle the sometimes confounding influences of these levels is to isolate them across geographic space.

The standard research approach to studying geographical inequalities begins with the subdivision of the population into societal groups according to specified criteria, and then into geographical areas to analyse regional or local variation within each group. Following this approach, in the study of ethnic inequalities the population needs to be subdivided into ethnic groups and geographical areas, in order to measure certain outcomes and to compare them with the overall population or with the ethnic majority. In this way it is possible to highlight any differentials that merit further investigation and explanation.

The role of geography in the explanation of social inequalities cannot be underestimated (Curtis and Jones, 1998; Dorling, 2005a), and this also applies to ethnic inequalities (Peach, 2000). Geographical analysis of ethnic inequalities can provide very important clues to understanding ethnic group differentials and observed patterns of within group heterogeneity. Furthermore, geography can be then used to explain the patterns found, through enhanced knowledge about those same areas and their related populations (Harris et al, 2005). This is especially relevant at the neighbourhood level, where most processes of ethnic inequalities and difference take place.

The subdivision of populations by ethnic group and geography does not only provide public policy makers with the evidence required to reduce ethnic inequalities at local level. It also gives them the means to better understand their populations and neighbourhoods, their characteristics and processes of change, in order to plan and deliver public services anticipating future needs at local level.

One area where this is key in the future planning of public services for the whole of the population, is population forecasting by ethnic group (Haskey, 2002), since there are significant variations in the demographic characteristics (Coleman and Salt, 1996) and geographical distributions of ethnic groups (Ratcliffe, 1996). The classic stepping stones of population studies; fertility, mortality and migration, have very different behaviours when broken down by ethnic group, and more importantly, they experience very rapid changes (Owen, 1996). This arises because of distinct patterns of population age and gender structure, migration stocks and flows, family formation and household composition, and so forth (Coleman, 2006).

Amongst these demographic variables, international migration is the one most difficult to measure, and even more so to predict (Rees and Butt, 2004). Although migration statistics in Europe have significantly improved in the last ten years (Salt, 2006), the situation is far from ideal, and as Rees and Boden (2006) put it; ‘There is an increasing desire for more comprehensive and more consistent information on new migrants at an international, national, regional and local level. International migration is now the dominant driver of population change in the UK and is set to remain so for at least the next 25 years.’ (Rees and Boden, 2006: 1). Political

representation and local funding all depend on accurate population statistics. The scale of the problem in this area of official statistics is such that a group of Local Authorities have brought legal action against the Office for National Statistics (ONS) since it recently introduced a change in the way international migration is imputed to local areas, as a consequence of which their income is set to be substantially reduced (The Economist, 2007a).

The populations of contemporary societies and cities are becoming increasingly multi-culturally diverse and globally connected, and the assumptions that worked well in the past in the demographic and geographic literature no longer hold true. The evidence of ethnic inequalities in Britain is now strong and encompasses a range of diverse and intertwined factors. New methods are thus required in the analysis of inequality in increasingly diverse populations and neighbourhoods, at the junction between 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. The next section will focus on one of these key aspects of policy, analysing residential segregation at the neighbourhood level.

2.2. Neighbourhood Profiling and the Segregation Debate

'[T]he professionals (..) too (..) [would have to] move, swapping their comfortable suburban homes for inner-city flats, moving into areas they would never normally consider living in, sharing streets with people whose skin is a darker shade than white.' (Dorling and Rees, 2003: 1287)

2.2.1. The community cohesion debate

Two major events in 2005 reopened a long-standing debate about the causes and consequences of residential segregation of ethnic minorities in European cities: the London bombings of July 7th, 2005, perpetrated by home-born terrorists; and the urban riots in France's *banlieues* in November 2005. Recent precedents of these events in the UK were the racialised disturbances in three northern English cities in the summer of 2001. These events triggered a heated public debate in which diverse issues were all linked to an apparent failure of European society to assimilate immigrant communities (Leppard, 2005; The Economist, 2005). A perceived manifestation of such failure is the residential segregation of ethnic minorities, whether through constraint or choice, to the frustration of ethnic minority integration policies, now relabelled into a broader buzzword; 'community cohesion'.

An independent inquiry chaired by Ted Cantle prepared a report for the Home Office after the 2001 riots, that highlighted ethnic segregation and the so-called problem of 'parallel lives' between ethnic groups as the major root cause of the community tensions and divisions (Cantle, 2001). After the London bombings in July 2005 the head of the Commission for Racial Equality, Trevor Phillips put the problem bluntly in these terms:

> *'But the aftermath of 7/7 forces us to assess where we are. And here is where I think we are: we are sleepwalking our way to segregation. We are becoming strangers to each other, and we are leaving communities to be marooned outside the mainstream.' (Phillips, 2005)*

Although several geographers have contested Phillips' and Cantle's view of ghettoisation of society as lacking or contradicting empirical evidence on the different dimensions of segregation (Dorling, 2005b; Simpson, 2006), these perceptions have frequently hit the headlines in Britain since then.

The current ghettoisation debate in Britain emanates from a broader shift in policy from a phase of celebrating an increasingly multicultural society in the late 1990's and early 2000's to a post-September 11th 2001/ post-July 7th 2005 that is fearful of 'parallel lives' and favours integration into the 'mainstream of society', a term difficult to define. These worries are linked to an increasing anxiety about individual atomisation and the lack of community cohesion, which is then projected on to those who are seen different from 'the norm' (Bunting, 2006). Fortuijn et al (1998), in a special issue of *Urban Studies* on ethnic segregation, have summarised public perception of 'ghettoisation' as an unwanted sequence of events in which 'increasing spatial segregation will lead to increasing separation of different social and ethnic classes and population categories; in its turn, that will produce ghetto-like developments and will finally result in the disintegration of urban society' (Fortuijn et al, 1998: 367).

As a result of this shift in public perception and government policy in Britain, initially directly stemming from Cantle's (2001) report, a series of policies have been implemented to promote 'community cohesion', a concept at the core of a re-launched 'new localism' effort (through area-based policies, such as Neighbourhood Renewal Areas, Health Action Zones, etc). A 'cohesive community' has been defined as:

> *'(…one) where there is a common vision and a sense of belonging for all communities; where the diversity of people's different backgrounds and circumstances are appreciated and positively valued; where those from different backgrounds have similar life opportunities; and where strong and positive relationships are being developed between people from different backgrounds in the workplace, in schools and within neighbourhoods' (Commission for Racial Equality, 2002: 1).*

A Commission for Integration and Cohesion was created in June 2006, within the Department for Communities and Local Government, with the aim of establishing how 'local areas can make the most of the benefits delivered by increasing diversity...[and] respond to the tensions it can sometimes cause.' (Department of Communities and Local Government, 2006). Unfortunately, the measurements to establish a baseline and progress on 'community cohesion' are not as easy to build.

2.2.2. Measuring residential segregation

The study of ethnic residential segregation has attracted geographers for a long time (Clarke et al, 1984; Jackson and Smith, 1981; Peach et al, 1981), and more recently they have made interesting contributions to the ethnic segregation debate (Dorling and Rees, 2003; Johnston, Burgess et al, 2006; Peach, 1996d; Simpson, 2005a), that begin with the difficult task of ascertaining the degree to which neighbourhoods and cities are considered segregated (Peach, 1981; 1996a). Underlying 'geographers and politicians [fascination] with measuring residential segregation' (Simpson, 2007: 406), seems to be an implicit association between high values of measured segregation and a lack of social integration (Simpson, 2007) that is threatening the social fabric of society (Fortuijn et al, 1998). What is more, Peach denounces the fact

that 'the literature has taken for granted that high levels of segregation are characteristic of marginalized and racially discriminated groups only, and that all high levels of segregation are negative, imposed, involuntary and transitory' (Peach, 2000: 622). But in reality, when looking at policies that have reversed residential segregation, such in post-apartheid South Africa, it has been found that desegregation initiatives alone do not solve the problem of lack of social integration (Lemanski, 2006).

However problematic, measuring segregation is useful to compare differences in population compositions between neighbourhoods and their changes across time. There is an established consensus in the quantitative sociological and geographical literature that there are different dimensions of residential segregation, and that separate indices should be used to independently measure these dimensions. Massey and Denton's (1988) seminal work proposed five major dimensions termed as; evenness, exposure, concentration, centralisation, and clustering, which have become very established as the standard in the last twenty years of segregation studies. *Evenness* relates to the degree of geographical spread of two groups among small areas; *exposure* measures the degree of potential contact between members of two different groups within the same small area; *concentration* speaks to the relative amount of physical space occupied by a group; *centralisation* addresses the degree to which a group is spatially located near the centre of an urban area; and *clustering* refers to the spatial clustering of groups (or their adjacency). Simpson (2007) has recently suggested two additional dimensions that measure migration over time and the overall composition of ethnic groups as: *movement* which relates to the extent of movement towards localities that already have relatively high proportions of the

same group; and *diversity* that concerns how close a set of groups are to equal numbers within an area.

Geographers are contributing to the methodological debate of segregation measures by moving the argument in two directions; measuring change over time and better reflecting the spatial aspects of segregation. The first contribution relates to the task of measuring whether segregation is increasing or decreasing through time and at different scales (Rees and Butt, 2004; Stillwell and Phillips, 2006), and the apportionment of any such change to demographic factors such as immigration, out-migration, and natural growth (Simpson, 2004; 2006; Stillwell and Duke-Williams, 2005). The processes behind each of these different factors have totally different meanings in the social integration debate, and relate to aspects of the so called 'good and bad segregation' (Peach, 1996c). For details of this debate on segregation change through time see Johnston et al (2005) and Simpson (2005b). The second contribution is concerned with measures that better reflect the spatial aspects of social interaction and segregation, which many of the established composite indices ignore. This builds upon well-established concepts in Geography such as the Modifiable Areal Unit Problem (MAUP) (Openshaw, 1984) and spatial autocorrelation (Goodchild, 1986), proposing innovative solutions to remove the effects of change of scale (Voas and Williamson, 2000), of shape and size of statistical areas, or the topological arrangement of ethnicity data (O'Sullivan and Wong, 2007; Wong, 2003; 2004).

A third aspect of the segregation debate that seems to be overlooked in the geographical literature is the effect of changes in the definitions and meanings of

ethnic groups. This is probably because the majority of the research in this field in most countries depends solely upon the information provided by censuses of population (Logan and Zhang, 2004), and the inert nature of ethnic categories is thus taken for granted. However, and as will be fully described in Section 2.4, censuses typically utilise only a few self-assigned ethnic categories, the number, definition and reporting of which may be highly inconsistent through time (Kertzer and Arel, 2002). Therefore, the spatial distribution of alternative ontologies of ethnicity, arranged in multiple possible combinations will have a strong impact on the level of segregation of particular neighbourhoods. This raises the issue of what we mean by a 'segregated' or an 'integrated' neighbourhood or city? (Peach, 1996d)

2.2.3. Ethnic segregation and neighbourhood profiling

After the UK 2001 Census results were released, the debate on the measurement of ethnic residential segregation and subsequent classification of neighbourhoods in Britain has experienced a new momentum of research and publications, linked to the debate on cohesive communities and the new localism of area-based policies. There are two key aspects of this new research stream. Firstly, it exploits the new features of the 2001 Census, typically employing more than one variable to measure ethnicity, and analysing it in combination with the new question on religion (Brimicombe, 2007; Peach, 2006): it also uses several levels of geographical disaggregation down to very fine Output Areas (Martin, 2002), and compares changes over the inter-censal period 1991-2001 (Simpson, 2005a). Secondly, it starts to use alternative ethnicity data sources to the census, primarily using the Pupil Level Annual School Census (PLASC), which reports ethnicity of children in state education every year and hence allows analysis of residential and school segregation as well as change over time (Burgess et al, 2005; Johnston, Burgess et al, 2006).

This *renaissance* of geographical research has also reignited an old debate in geography, concerning the motivation of inferring process from pattern. Two general strands can be identified, defined by those who prefer to use relative measures of segregation and focus on investigating processes (Peach, 2006; and Simpson, 2007) and those who prefer to use absolute measures and focus on patterns (Johnston, Poulsen and Forrest, 2002a; 2002b; 2003; 2005; 2006; Poulsen, 2005). The latter group seems to have caught the attention of the media and policy makers, because they develop a classification of neighbourhoods according to certain interpretations of segregation, that once labelled as 'ghettos' or 'enclaves' easily catches the eye of the public. The declarations by Trevor Phillips quoted above were based on research presented at the Royal Geographical Society Annual Conference by Poulsen (2005), that depicted highly segregated areas as increasing. Johnston et al (2002a; 2002b; 2003; 2005; 2006) and Poulsen et al (2005; 2002; 2001) have built classifications of neighbourhoods in different countries that include the terms 'ghettos' and 'enclaves'. This typology is further subdivided into; 'polarised', 'encapsulated', 'concentrated', or 'assimilated' enclaves according to a set of absolute thresholds for different measures of segregation deemed to mark the transitions between different neighbourhood types in a pre-conceived model of segregation-multiculturalism-assimilation. A similar approach is followed by Brimicombe (2007). Other authors do not agree with this interpretation of segregation because it does not take into account temporal change and demographic dynamics (Dorling, 2005b; Simpson, 2004; 2005a; b; 2006; 2007), or because they prefer to use relative measures, such as the Index of Dissimilarity rather than absolute ones (Dorling and Rees, 2003; Peach, 2006).

This neighbourhood profiling exercise follows a tradition in the urban sociological literature that started with the Chicago school of urban ecology and Burgess and Park's famous concentric rings model of the growth of Chicago (Park et al, 1925). They marked areas of the city as 'Black belt', 'Chinatown' 'Deutschland', 'Ghetto', 'Little Sicily', or 'Second Immigrant settlement', depicted by adjectives such as 'Vice', 'Under-world' or 'Slum' (see Figure 2.2). Since then, many authors have classified neighbourhoods and cities according to the ethnic composition of their populations, establishing links between these communities, their position in the socio-economic hierarchy of the city and the type of urban environment that they occupy.

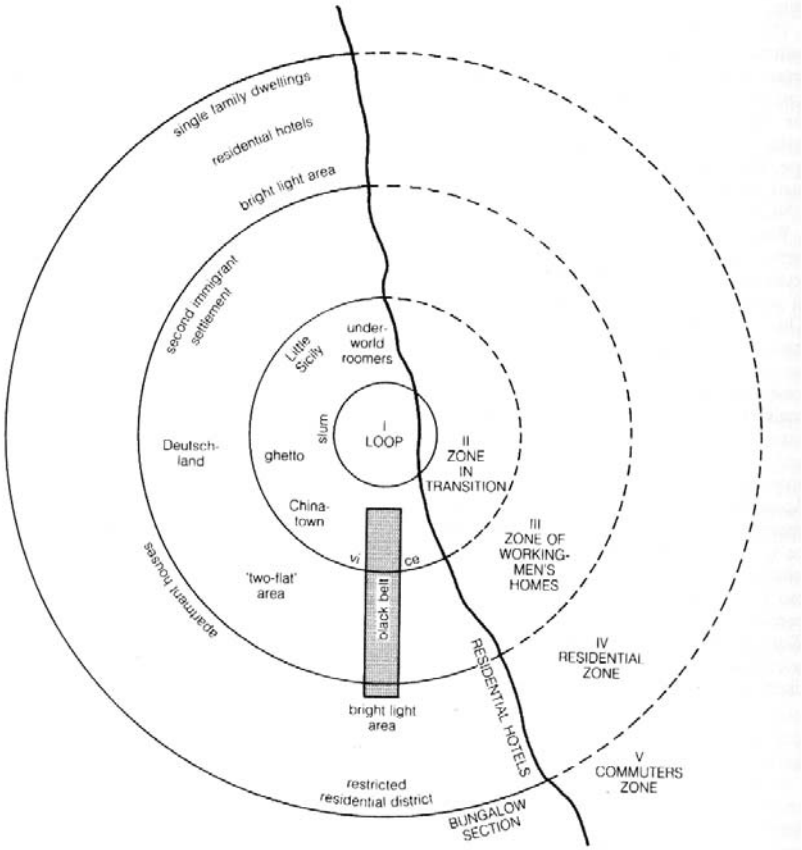

Figure 2.2: Burgess' Concentric Rings Model of the growth of Chicago
Source: Park et al (1925) reproduced in Johnston et al (2000:904)

Philpott (1978) studied these classifications and concluded that there was a profound distinction between European immigrant ethnic enclaves, where only a minority of the total populations of these groups lived, and the true ghetto of African Americans, where most of the population of this group lived. These distinctions have been used by Peach (1996a; 1999a) and Poulsen (2001) to build their classifications of areas and levels of segregation, and continue to influence the literature of urban segregation.

2.2.4. Meanings of segregation and the geography of ethnic inequalities

It is interesting to note the persistence of the skin colour criterion when creating these segregation classifications of neighbourhood, where the White/Non-White divide seems to be a perennial axis of segregation, and is construed as 'a problem', both in the early twentieth century as well as today. It seems that the definition of the segregation/integration problem lies upon a racial dichotomy, which seems to depict Non-White concentrations as negative (Simpson, 2004), when the reality looks more like a complex spectrum of 'skin tones' and cultures.

Even when several ethnic groups are analysed, most research on segregation has been reduced to just a few of the largest minorities in a country, which in the UK have typically been South Asian (Indian, Pakistani and Bangladeshi) (Peach, 1998) and Black minorities (Phillips, 1998), or increasingly a Muslim minority (Peach, 2006; Peach and Owen, 2004). Little consideration is usually paid in segregation studies to what it means to be 'White' (Peach, 2000), who the 'Other' ethnic groups are (Connolly and Gardener, 2005), or whether it is meaningful to use overarching groups such as 'Asians' (Aspinall, 2003) or 'Hispanics' (Choi and Sakamoto, 2005). On the contrary, the question that seems to be missing from segregation studies is;

what are the key cultural differences or discrimination factors that justify the study of segregation as a vehicle for understanding the wider geography of ethnic inequalities?

Therefore, although the meaning of segregation is hotly debated in geography (Johnston et al, 2005; Peach, 1996a; d; Simpson, 2006), the meaning and ontology of ethnicity nevertheless seems to go unnoticed in the discipline, whereas it constitutes one of the central issues in contemporary public health research (Aspinall, 2007; Bhopal, 2007). It is in this precise gap where this PhD intends to make a contribution, helping advance the debate about the ontology of ethnicity and how it may affect the results of geographical analysis at the neighbourhood level.

2.3. Defining Ethnicity and Race

Ethnicity and race are very controversial variables in scientific inquiry, and during over 150 years of speculation biologists, anthropologists and geneticists have demonstrated over and over again that these terms are both socially constructed and lack any biological reality (Cavalli-Sforza, 1997).

2.3.1. Race

The history of how, during the age of European colonialism, scientists identified races and ranked them according to their biological and social value, with the 'White-European race' always ranking on top, is unfortunately well known (Gould, 1984). They justified such rankings based on claims of intelligence hierarchies using measurements of the size and shape of the head, and even the contents of the brain (Gould, 1984), with the underlying value that biology determined social position; in short, biological determinism (Bhopal, 1997). This type of research whereby human

populations were divided into sub-species, mainly on the basis of visible physical characteristics, was used to justify slavery, imperialism, anti-immigration policy, and the social status quo (Bhopal, 1997). It was dominant for most of the 19th century and beyond until its abandonment with the defeat of the Nazis at the end of the Second World War (Bhopal, 2004). Attached to the ideas of the Nazis were the Eugenic theories, which sought the improvement of the 'human race', in particular the 'Aryan race'. A book titled 'Outline of Human Genetics and Racial Hygiene' was published in 1921 by the geneticist Fritz Lenz, a leading advocate of the Aryan ideology, and is claimed to have been very influential in Hitler's (1925) own book 'Mein Kampf', where he set out his political beliefs about German racial superiority (Olson, 2002).

Today race is defined in the Cambridge English Dictionary as 'a group, especially of people, with particular similar physical characteristics, who are considered as belonging to the same type, or the fact of belonging to such a group' (University of Cambridge, 2004). Therefore, it is a subjective 'consideration of belonging' that makes it a social construct. There is a general agreement, forged through the last four decades of population genetics research, that the concept of race is socially constructed, and cannot be explained by genetic differences between human groups (Cavalli-Sforza, 1997). But even though none of the numerous 'scientific' racial classifications has stood the test of time (Bhopal, 2004), current 'race' classifications remain influenced by 'biologically rooted' racial stereotypes (Graves, 2002), and the concept of 'race' is still strongly used in many countries, such as the U.S., when subdividing populations according to their ancestral origins. The persistence today in the U.S. of the concept of race, and hence the use of racial classifications in administrative records and academic studies, may be traced to the legacy of the

American Civil Rights Movement (1954-1968) and the legislation subsequently introduced to prevent racial discrimination. Although the contemporary concept of 'race' has partially lost its roots in distinguishing differences in physical appearance alone (phenotypes), it is still loaded with ideological assumptions about innate, hereditary, ranked differences between groups of people (Chapman and Berggren, 2005).

However, the debate surrounding biological differences of human groups has not been closed, but actually moved on to a new stage in the era of individual human genetics. In a recent issue of *Science* magazine commemorating its 125th anniversary, two of the '125 big questions that face scientific inquiry over the next quarter-century' are very closely related to this debate:

> *'What are human races, and how did they develop? Anthropologists have long argued that race lacks biological reality. But our genetic makeup does vary with geographic origin and as such raises political and ethical as well as scientific questions'*
>
> (Science, 2005 : 100. Emphasis added)
>
> *'To What Extent Are Genetic Variation and Personal Health Linked?'*
>
> (Couzin, 2005:85)

There is a growing belief, in the health and anthropological literature, that the biological concept of race made a strong come back at the turn of the Millennium, hand in hand with the genetics revolution in science (Kahn, 2005). In this era of race genetics and genetic medicine (Nature Genetics, 2001), 'Gene hunting [has become] the new research colonialism' (Pearce et al, 2004: 1071), in which scientists try to identify key differences in gene frequencies between different 'populations'. The key

to mapping DNA groups therefore lies in the definition of such 'populations', which are again socially constructed based on geographical, anthropological and historical assumptions (M'charek, 2005).

In order to overcome the biological determinism implicit in the term 'race', and to include other non-biological factors that make us perceive human groups as different from each other, the concept of 'race' has been rapidly abandoned in favour of that of 'ethnicity'. This trend has been observed in the last three decades of the 20th century, primarily outside the US (Oppenheimer, 2001), and is especially well documented in the health literature (Afshari and Bhopal, 2002). However, this trend is not without problems, since it assumes that both terms can be used interchangeably, as if it both described the same quality – despite this assumption being disproved by many authors (Bhopal, 2004).

2.3.2. Ethnicity

The word ethnicity derives from the Greek word ethnos, meaning a nation, and the term ethnic group is considered to have been introduced by Max Weber in 1922. He defined ethnic groups as 'Those human groups that entertain a subjective belief in their common descent because of similarities of physical type or of customs or both, or because of memories of colonization and migration (...) it does not matter whether or not an objective blood relationship exists' (Weber, 1922, cited in Guibernau et al, 1997). Therefore, at the core of the concept of ethnicity is a subjective belief of common origins without the necessary existence of genetic linkages or physical similarity. This concept is thus closely linked to the question of an individual's identity, which is defined by the characteristics of the ethnic group to which he or she recognises belonging. Amongst the main reasons for such perception of self-

identity are certain shared characteristics, including physical appearance, but most importantly geographical and ancestral origins, cultural traditions, religion and language (Bhopal, 2004). Therefore, the most widely accepted notion of ethnicity is a multi-dimensional concept that encompasses different aspects of a group's identity, in relation with kinship, religion, language, shared territory, nationality, and physical appearance (Bulmer, 1996). Understood as such, ethnicity is considered to differ from race, nationality, religion, and migrant status, sometimes in very subtle ways, although it is considered to include traits of these other concepts as well (Bhopal, 2004).

Therefore, at the core of the concept of ethnicity is the question of an individual's identity, which is defined by the characteristics of the ethnic group that he or she considers herself to belong to, always understood in a contextual rather than in an essentialist way (Peach, 1996b: who himself might be considered Welsh in England, British in Germany, European in Thailand, and White in Africa). The social context in which the ethnic group is defined is therefore key to understanding its identity. This idea stems from one of the more interesting facts observed during the processes of ethnic group formation; not only a firm belief in group affinity is required for group identities to emerge, but this is usually defined in opposition to other groups perceived as being culturally different and with whom contact is required (Eriksen, 2002). In other words, if there is no contact with other groups that are perceived as 'culturally different', the identity of an ethnic group to which one belongs does not emerge. For example, the concept of a Hispanic ethnic group only emerged in the US during the 1960s and 1970s, when large numbers of Spanish speaking immigrants from many countries and their descendents found a common identity through a

shared language and migration history in an English-speaking country. Not only did the 'host culture' consider them as one group, but Spanish-speakers from Latin America considered that this new identity would make them stand out in the US in a much stronger way than with their individual national identities (Skerry, 2000). The paradox is that no Spanish speaker outside the US would consider himself or herself as 'Hispanic', and the group's homogeneity is difficult to sustain (Choi and Sakamoto, 2005). This important appreciation of contact between differently perceived groups explains why the debate on ethnic identity has grown since the end of the Cold War in developed countries (Castells, 1997). This recent trend is explained by the disappearance of the communist-capitalist bipolar world and its political antagonism that prevented mass population movements and the redrawing of national borders, the diminishing role of the nation-state, the growth of nationalisms, and a growing number of different human groups living amongst each other in large numbers (Castles and Miller, 2003).

2.3.3. Criticisms

Nonetheless, the characteristics that together define ethnicity are not fixed or easily measured, so ethnicity is considered in science as a subjective, contextual, transient and fluid concept (Senior and Bhopal, 1994), and probably the most controversial subject of study in social science (Nobles, 2000). The fluidity of the concept of ethnicity is at the root of the anti-essentialists' critiques, who challenge the whole idea of trying to classify people into discrete and immutable categories, such as social classes but especially ethnic groups (Brubaker, 2004). These authors favour the concept of 'identities' which are subjective, fluid and always evolving, where people can assign themselves to several categories which, taken together, may better reflect the complexity of their lives (Pfeffer, 1998). Even the American Sociological

Association describes race (in the US research context) as 'a social invention that changes as political, economic, and historical contexts change' (American Sociological Association, 2002: 7). Although, as has been mentioned above, there is a consensus that the modern concept of race is not equivalent to ethnicity, the differences between the two are still widely ignored by researchers (Comstock et al, 2004). This confusion makes the understanding of the separate processes of inequalities, arising from racial or ethnicity factors, even more difficult and controversial.

Other authors such as David Harvey relate current issues of ethnicity and race difference with more traditional structural differences in class identity;

> *'Popular as well as elite class movements make themselves, though never under conditions of their own choosing. And those conditions are full of the complexities that arise out of race, gender, and ethnic distinctions that are closely interwoven with class identities.' (Harvey, 2005: 202)*

This contention seems to suggest a situation of 'old wine in new bottles', in which new identities formed around minority groups (according to race, ethnicity, gender, sexuality, age, or disability) have replaced old divisions along social class lines in the explanation of socio-economic inequalities.

Going back to the concept of ethnicity, because it is considered a core element of personal identity, the current preferred method for ascribing one's ethnicity in research and government statistics is self-assessment. However, since the categorizations of ethnic groups are usually pre-classified and individual choice is constrained to choosing amongst them, the concepts of ethnic groups themselves are

also considered an externally imposed identity (Senior and Bhopal, 1994). Therefore, the definition and measurement aspects of identity are closely related and cannot be studied in isolation. The problem of ethnicity measurement is dealt with in the next section. Ethnicity, rather than the more biologically rooted concept of race commonly used in the US, will be used from now on since it is the concept most widely used to define the identity of population groups by ancestral and cultural origin.

2.4. Measurements of Ethnicity

Following from the complex definition of ethnicity presented in the previous section, and with the aim of studying ethnic inequalities at neighbourhood level, this section will review the issues around the difficult task of measuring ethnicity in order to classify people into ethnic groups.

2.4.1. Measurement issues in official ethnicity classifications

It should be obvious by now why the measurement of ethnicity is problematic; because ethnic identification is subjective, multi-faceted and changing in nature and because there is not a clear consensus on what constitutes an 'ethnic group' (Coleman and Salt, 1996; Office for National Statistics, 2003). However, as has been justified in the introduction of this chapter and further explained in Section 2.1, the measurement of ethnicity is today useful for a wide range of purposes in many countries, especially to reduce ethnic inequalities. This puts pressure upon government statisticians who try to cope with surges of interest in collective identity formation and with the struggle of States to monitor and sometimes try to shape these processes (Kertzer and Arel, 2002). Even when a consensus in social statistics is reached, with time the action of statisticians cannot be detached from their

consequences on the reality being measured, and as Barrier (1981) puts it; 'The census imposes order of a statistical nature. In time the creation of a new ordering of society by the census will act to reshape that which the census sought to merely describe' (Barrier, 1981: 75).

The national Census of Population comprises the major classificatory effort of a society, and has been described as a sort of communal 'family photograph' that is only taken every ten years (Skerry, 2000). Therefore, the social processes and groups that appear in such photograph are of high importance, since census enumeration brings with it political and economic power (through representation and funding). As such, the classification of the population into the groups of common ancestry used in the census brings with it an official statistical recognition that transcends the census enumeration exercise and determines all sorts of possibilities in the arena of identity politics during decennial inter-censal periods and beyond (Skerry, 2000). As such, in most countries the de facto 'gold standard' for ethnicity measurement usually emanates from the categories created by the national population censuses (Kertzer and Arel, 2002).

The UK Office for National Statistics (ONS) recognises that measurement of ethnicity should be done in a way that is sound, sensitive, relevant, useful, and consistent over some period of time (Office for National Statistics, 2003). However laudable these statisticians' principles, Skerry (2000) depicts very well the tension in the US Census Bureau 'between the extremely technical character of the census and the emotional, highly symbolic nature of race politics' (Skerry, 2000, 4). These types of frictions were behind the reasons why, despite having been considered since 1971,

an ethnicity question was not introduced in the UK until the 1991 Census (Coleman and Salt, 1996), why it is still not asked in many countries (for example in France or Spain), and why it has created so much controversy before and after each US census during the last decades (Nobles, 2000). An early quote from the introduction in the US of an official racial and ethnicity classification summarises well this point:

> *'These classifications [set in the Racial and Ethnic Standards for Federal Statistics and Administrative Reporting] should not be interpreted as being scientific or anthropological in nature' (Office for Management and Budget, 1978: 19269)*

Even when national consensus is reached, a further problem arises when trying to perform international comparisons between national censuses, since the terms used to describe ethnic groups are developed within each country in response to their own particular historical processes of ethnogenesis (Aspinall, 2005). In the round of population censuses conducted at the turn of the Millennium, 141 countries collected information about the ancestries or identities of their populations, using questions on one or more of the following dimensions of identity; ethnicity, race, indigenous/tribal origin, and nationality (Morning, 2008). However, international comparisons are highly limited because of the different ontologies of ancestral origin and identity that underlie each of the classifications. Henceforth this PhD thesis will mainly focus on the UK context, unless otherwise specified.

2.4.2. The UK Census ethnicity classification

After two consecutive UK censuses collecting ethnicity information, the census ethnicity classification has become a 'gold standard', currently being used by most public bodies and many private institutions as a template for data collection. The set

of ethnic categories used in both the 1991 and 2001 Censuses, as well as that proposed for the next 2011 Census, are listed in Table 2.1. This shows an expansion in the number of categories and the specificity of the descriptions.

The 2001 classification addressed some of the omissions in the 1991 question (mixed ethnicities, and breakdown of the white category) (Bulmer, 1996; Rankin and Bhopal, 1999), expanding the 8 original categories to 16. An additional and voluntary question on religion complemented the 2001 Census ethnicity classification.

However, the 2001 classification has been criticised for continuing to stress the importance of skin colour across the whole classification (Aspinall, 2000), the problematic mix of 'racial', 'ethnic', and 'geographical' domains, the vague definitions of the smaller minority groups (buried under the 'other' categories) (Connolly and Gardener, 2005), the mismatch with self-descriptions of identity given by a high volume of write-in answers (that the ONS assigns back to one of the 16 categories) and the move to pan-ethnic racial groups in some census outputs (White, Black, Asian, Other). Furthermore, the classification has been criticized for failing to reflect the internal heterogeneity of some groups such as 'Black African' (Agyemang et al, 2005), 'Asian' (Aspinall, 2003), 'White' (Peach, 2000), or 'Other' (Connolly and Gardener, 2005). As a result, Aspinall (2000) states that the 2001 classification falls short of what is needed in the ethnicity and health context, because it does not capture adequately the multi-dimensional character of ethnic identity.

1991 Census	2001 Census	2011 Census (proposed for England)
0 White	White	A - White
	A British	English
1 Black - Caribbean	B Irish	Other British
	C Any other White	Irish
2 Black - African		Any other white
3 Black - Other	Mixed	B - Mixed
	D White and Black Caribbean	White and Black Caribbean
4 Indian	E White and Black African	White and Black African
	F White and Asian	White and Asian
5 Pakistani	G Any other mixed	Any other mixed
6 Bangladeshi	Asian or Asian British	C – Asian
	H Indian	Indian
7 Chinese	J Pakistani	Pakistani
	K Bangladeshi	Bangladeshi
8 Any other ethnic group	L Any other Asian background	Chinese
		Any other Asian background
9 Not Given		
	Black or Black British	D - Black or Black British
	M Caribbean	Caribbean
	N African	African
	P Any other Black background	Any other Black background
	Other Ethnic Groups	E - Other ethnic group
	R Chinese	Arab
	S Any other ethnic group	Gypsy / Romany / Irish Traveller
		Any other ethnic group
	Z Not stated	Not stated

Table 2.1: UK Census ethnicity classifications in 1991, 2001 and 2011 (proposed) in England
Source: (Coleman and Salt, 1996; Office for National Statistics, 2000; 2006a)

Despite the problems with the UK 2001 Census ethnicity classification it has been widely adopted as the standard to be used by public institutions when collecting ethnicity data in order to facilitate comparisons (Office for National Statistics, 2003), especially in the health arena (Department of Health, 2005a). When the 16 categories are not sufficient to reflect the diversity of the population being measured for a

particular purpose (for example to analyse equity of admissions to a Hospital in Inner London), the Department of Health condones breaking down any census category into sub-categories, but requires that these always 'nest-together' back into the 16 census categories (i.e. no overarching group is permitted, such as 'Arab'), thereby preserving levels at which data aggregations should always be available (NHS Information Authority, 2001). A similar arrangement is adopted by the Department of Education and Skills for reporting ethnicity in the Pupil Level Annual School Census (PLASC), resulting in a large number of 95 ethnic groups. A full list of the ethnic group categories contained in PLASC data is offered in Appendix 2.

Plans for the UK 2011 Census are already well under way, and after a period of public consultation the Office for National Statistics is testing a full questionnaire for England and Wales in May 2007. It includes a total of 19 ethnicity categories, the 16 from 2001 plus a new breakdown of White into English and Other British, and two new categories of Arab and 'Gypsy / Romany / Irish Traveller', as listed in Table 2.1. Furthermore, the test questionnaire also includes two new multiple response questions on languages spoken and national identity (Office for National Statistics, 2006a), and the question on religion and country of birth. This set of questions for 2011 partially addresses some of the often-rehearsed issues with official ethnicity classifications, which are summarised in the next subsection.

2.4.3. Issues with official ethnicity classifications

Despite their widespread influence, there are three major problems with the way ethnicity is currently officially measured in most developed countries. First, ethnicity is usually measured as a single variable, that of an 'ethnic group' into which the individual self-assigns his or herself from a classification of a reduced number of

classes, with no leeway to represent any characteristics of the multi-faceted nature of self-identity described above. This problem has been partially addressed in the U.S. 2000 Census in which respondents were able to choose from more than one 'race/ethnic group', although it has created a new issue of comparability across time and between different combinations.

A second problem is that pre-set ethnic classifications are used as opposed to just an open question, and the responses are then arranged according to the most meaningful common identities. This is of course justified with the need to facilitate the creation and comparison of the resulting statistics over time and between different information sources (Office for National Statistics, 2003). However, as mentioned before these categories have proved not to reflect the complex heterogeneity found within each group (Agyemang et al, 2005; Connolly and Gardener, 2005; Rankin and Bhopal, 1999).

A third problem arises from the method of determining ethnicity by self-assessment, which comprises the current consensus across datasets and the literature (Bhopal, 2004), as opposed to it being assigned by a third person or a computer according to some established measurable criteria. As a result of self-classification, the ethnicity of the same person can vary through time, since perceptions of individual and social identity changes over time (Aspinall, 2000) and are influenced by the type of ethnicity question asked (Arday et al, 2000), the definitions of categories offered (Olson, 2002), and the country and method of data collection. Although this is not the aspect of ethnicity classification that is the most highly debated, self-defined ethnicity has been deemed as 'unhelpful' (McAuley et al, 1996).

In addition to these three major issues of official ethnicity classifications, an additional recognised problem is the lack of routine collection of ethnicity data in most government or public service datasets.

2.4.4. The limits to ethnicity data in the UK

Despite the UK Census ethnicity classification having become the standard for ethnicity information collection, ethnic group is still not recorded in most public sector datasets, including population registers such as birth, death, electoral and health general practice registrations (London Health Observatory, 2003; Nanchahal et al, 2001). In the health area, even though such collection has been mandatory in hospital admissions statistics since 1995 (NHS Executive, 1994), data coverage is still a poor 74% of all episodes (London Health Observatory, 2005) and of low quality when compared with other research sources (Bhopal et al, 2004).

Table 2.2 shows the results of a recent study by the Association of Public Health Observatories (2005) that analyses the percentage of records with incomplete ethnicity coding in eight separate datasets. The study concludes that a substantial proportion of events are not assigned an ethnic group because of organisational issues, rather than because of the relative size of ethnic minority groups at the local level (i.e. when ethnic minorities represent a small share of the population it might be argued that there is little incentive to record ethnicity for all of the population).

Dataset	England	London
	(percentage)	
Pupil Level Annual School census (PLASC), 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 2.2: Percentage of records with incomplete ethnicity coding in different datasets
PLASC= Pupil Level Annual School census; TB= Tuberculosis; AIDS= Acquired Immune Deficiency Syndrome; HIV= Human Immunodeficiency Virus; SOPHID= Survey of Prevalent HIV Infections Diagnosed; NDTMS= National Drug Treatment Monitoring System
Source: (Association of Public Health Observatories, 2005, 12)

However, that these datasets include any ethnicity information at all makes each the exception rather than the norm (Bhopal et al, 2004; Harding et al, 1999). This lack of ethnicity information has been described as critical in frustrating attempts of social science and health researchers to measure ethnic inequalities, produce accurate population forecasts, assess public service use, and demonstrate compliance with policy and legislation. There has been a specific call for an effort to record ethnicity at least for all births and deaths (London Health Observatory, 2003), although this has been a classic unattended demand (Cook et al, 1972). The problem is such that in the latest *Status Report on the Programme for Action in Health Inequalities* the UK Department of Health states in several points that ethnicity, although a powerful factor of health inequalities, is not systematically covered in the report because of lack of data availability (Department of Health, 2005b). They add; ‘this is crucial

since targeting sectors of the population is a key characteristic of most effective [health] interventions' (Department of Health, 2005b, 61). This gives a sense of the scale of the problem of lack of availability of ethnicity data.

2.4.5. The limits to comparability between research studies

Even when ethnicity information is collected, its consistency and comparability is usually very poor. As a consequence, research on ethnicity has been hampered by a lack of common methodologies in the collection and treatment of ethnicity information (Whitehead, 1992). Different studies define ethnicity in different ways, and create independent classifications and non-comparable methods of data collection (Choi and Sakamoto, 2005). Decisions taken in this respect are only based on the tactical considerations deemed most appropriate for each context, while making explicit neither the methodology nor the classification. The inevitable consequence is that results cannot be correctly interpreted and compared between studies.

The problem of lack of comparability is especially critical in research about differential outcomes by ethnic group. Comstock et al (2004) summarise very well the extent of this problem in public health research. They conducted a comprehensive review of 1,198 articles published in the *American Journal of Epidemiology* and the *American Journal of Public Health* from 1996 to 1999, and found 219 different terms to describe just 8 core 'ethnic groups'. Moreover, the authors denounce the frequent failure of researchers to explicitly define the ethnic categorizations and their context of use, to differentiate between race and ethnicity, to state the study methods used, and to significantly discuss the results. Bearing in mind that the large collection of articles were drawn from just two journals of the same scientific discipline in the

same country, where research on ethnic disparities has a longer tradition, this issue poses a crucial problem that requires 'continued professional commitment […] to ensure the scientific integrity of race and ethnicity as variables' (Comstock et al, 2004:611). This problem has been also identified by other authors, and defined as an ontological problem that constitutes 'a problem with basics' (Bhopal, 2004: 441).

It is important to mention here the efforts made, especially in health research, to overcome the comparability issues in ethnicity studies. In the UK, this debate began following the 1991 Census inclusion of the ethnicity question and its mandatory recording in hospital admissions since 1994. Most of the main issues with the official ethnicity classifications described in this section have already been pointed out by Senior and Bhopal (1994), and have been highly debated during the last decade, with important contributions by Peter Aspinall (2002; 2005; 2007), and Raj Bhopal (2004; 2007; 1998). These and other authors agree that researchers in health and ethnicity should use comparable ethnic classifications and make explicit the meanings of the ethnic group categories selected, the criteria use for such selection, their method of ascribing ethnicity to individuals, and give precise explanations of differential health outcomes by each of the ethnic groups studied. Unfortunately, this objective is still far from becoming a reality, and even more so outside ethnicity and health research.

Taken together, the issues of lack of reflection of the multi-dimensional nature of ethnicity, the use of just a few pre-defined coarse categories, the variability of self-assignment of ethnicity, the lack of routine collection of ethnicity information, and its low quality and comparability, present major impediments for researchers and

public policy decision makers. Their consequences are that researchers are prevented from measuring socioeconomic inequalities, equity of access to and uptake of public services by ethnic group, and demonstration of compliance with anti-discrimination and equal opportunities legislation, in an increasingly multicultural population.

2.4.6. Alternative measurements

As a consequence of the lack of ethnicity data availability, other proxies, such us country of birth, have been used to ascribe a person's ethnicity when it is not known (Marmot et al, 1984; Wild and McKeigue, 1997). Despite its utility to classify migrant origins, with growing numbers of second generation migrants, the proportion of the 'ethnic majority' people born abroad, and migrants born in 'intermediate' countries (i.e. East African Indians that migrated to the UK), this method has become increasingly inappropriate (Harding et al, 1999). In the UK 2001 Census, only half of the ethnic minority population was recorded as born outside the UK. Furthermore, many health and demographic studies use country of birth from death certificates, which rely on an informant and may be less accurate than the census, when the person is still alive to provide the information (Gill et al, 2005).

In some countries where the concept of 'foreigners' (as opposed to nationals or citizens) is still used as a proxy for ethnic minority, such as Germany, Spain or France, the main variable used to classify populations by origin is nationality, which is not recorded by the UK Census. This proxy is also problematic since it can change over time, some people retain more than one nationality, and usually second or third generation migrants acquire the host country's nationality.

A third alternative method employed as a proxy for ethnicity is the analysis of name origins. Personal names are in principle good indicators of ethnicity, at least in relation to the immediately previous generations, that gave the forename to their descendants and probably exercised some preference in the surname. After migration to another country or region, names can probably be viewed as a kind of 'self-assignment' of ethnicity that is likely to have strong links to the language, culture and geography of a person's ancestry. Names have been used in particular to identify the main ethnic minority populations in some 'destination countries', with a relatively good degree of accuracy. This alternative method forms the core methodology of this PhD thesis, and as such will be further reviewed in detail in Chapter 3, and built upon through an innovative methodology in subsequent chapters. Therefore, repetition is avoided here.

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 classify a person into an ethnic group if these six dimensions were to be measured separately. This conclusion has been reached by several researchers in ethnic inequalities in health, that call investigators to use a range of variables instead of just one summary measure that is deemed to encapsulate language, religion, country of birth, family origins, and length of residence (Bhopal, 2004; Gerrish, 2000; McAuley et al, 1996). Physical appearance seems to be a much more sensitive aspect to ask about, and even more to classify.

Even the trend in national censuses is now towards measuring these different dimensions separately, with a religion question having been introduced in the 2001

UK Census, in addition to ethnicity and country of birth questions: there are also plans for a question on language spoken and national identity in the 2011 Census, and these are currently being evaluated (Cope, 2005; Office for National Statistics, 2006a). The collection of data on languages spoken in the UK Census will not only provide a richer insight into the culture of ethnic minorities in Britain, for example allowing public services to be better targeted to different languages, but it will also allow for a key aspect of the individual identity to be revealed. This is of course, assuming that the final census question makes it possible to differentiate between the primary mother tongues and other languages learnt outside the household.

Therefore, and constituting one of the hypotheses that this PhD will try to test, it is believed that there is a strong relationship between the ethnic identities of human groups and their mother languages, and that an indication of these can be revealed by the analysis of name origins. If this hypothesis can be proved correct and a methodology developed for the purpose of studying ethnic inequalities at neighbourhood level, this research may be invaluable in overcoming the problems arising from ethnicity being measured as a single variable, the difficulties in classifying and generalising about ethnicity, the lack of data between censuses, and the coarse categorisations that census-type surveys adopt.

2.5. Conclusion

The evidence on ethnic inequalities in the developed world that has accumulated over the last two decades has allowed researchers to identify consistent and stark differences between ethnic groups along the most important dimensions of quality of life: health, education, employment, housing, general well-being and social support.

A range of diverse and intertwined factors lie behind these inequalities, but they have at best only been described in terms of associations and not as cause and effect relationships. This is largely because of the lack of data that are fit for purpose.

The subdivision of populations according to ethnicity and geography has allowed better understandings of contemporary society and neighbourhoods, as populations and cities are becoming increasingly multi-culturally diverse and globally connected. Today and in the immediate future, improved methods of understanding the processes of population composition and change by ethnic group and small area are desperately required to improve our knowledge of the whole population's dynamics and the impact of welfare provision.

The debate about residential segregation in Britain has gathered increased public interest in recent years along headlines of 'ghettoisation', 'parallel lives' and lack of community cohesion. New policies have been implemented within a re-launched 'new localism' agenda implemented through area-based policies, but unfortunately such policy analysis has not been able to establish the baseline measurements necessary to define concepts such as 'community cohesion'. There are different meanings and dimensions of segregation, and geographers have contributed to their measurement, adding spatial and temporal dimensions to the extensive sociological literature. However, research remains divided between those who favour the use of relative measures and emphasise the processes of change, and those who prefer absolute measures and focus on patterns of segregation. Much more complex combinations of ethno-religious groups and spatio-temporal scales are now

facilitated by more detailed classifications (e.g. 2001 UK Census) and alternative datasets to the census (e.g. Pupil Level Annual School Census - PLASC).

As pioneering methods to study ethnicity begin to mature, and society experiences very rapid changes, the problems of what ethnicity means and how should it be measured, assume ever greater importance. Ethnicity is a multi-dimensional concept that encompasses various aspects of individual identity expressed in reference to a group's origin. As such, it is socially constructed, contextual, and fluid, and thus very problematic to define and even more to measure.

Measurements of ethnicity have become increasingly standardised through the wide adoption of official ethnicity classifications throughout public datasets. However, such classifications have a series of problems: they fail to reflect the multi-dimensional nature of ethnicity; they are restricted to just a few pre-defined coarse categories; and they are subject to the variability of self-assignment of ethnicity. Moreover, the lack of routine collection of ethnicity information, and its low quality and comparability across datasets and periods of time, present major shortcomings for researchers and public policy decision makers. As a result, they have turned to alternative methods to classify population by ethnicity such as analysing country of birth, nationality and name origins as proxies for ethnicity, whenever existing ethnicity information is clearly not fit for purpose.

Through all of the above conclusions, it is clear that new methods are required in the classification of populations according to ethnicity in increasingly diverse societies and neighbourhoods. Such improved methods are fundamental in informing public

policy in the reduction of ethnic inequalities, such as in planning health services at local level, providing more robust evidence in the residential segregation debate, and producing accurate population statistics by ethnic group and small area.

It is in this methodological research area where this PhD intends to make a contribution. As such, it is located at the intersection between the debate on the ontologies of ethnicity and the development of alternative measurements in the classification of populations and neighbourhoods along new and complex multidimensional aspects of identity. One specific alternative measurement will be fully developed and will form the core of the PhD: name origin analysis will be used to ascribe populations to ethnic cultural and linguistic groups. The next chapter reviews the literature in this area with a view of identifying the research gaps and potential contributions of improved methods to ascribe ethnicity for the purposes that have been described here.

Chapter 3. Names and Ethnicity

A core argument of this thesis is that the analysis of personal names can be used to ascribe populations to a robust and defensible taxonomy of ethnic, cultural and linguistic groups, and that name taxonomies are valuable when ethnicity classifications are not available at the desired quality, geographical scale or nominal groupings. The previous chapter has justified the need to classify populations by ethnicity, and considered the problems with its measurement primarily from a population geography perspective. This chapter will review in depth the multidisciplinary literature of name-based ethnicity classifications with a view to summarising the different approaches that have been independently developed, to identifying the research gaps, and to setting out a research agenda for potential contributions using the improved methods, that will be described and developed in the rest of the thesis.

Name origin analysis techniques have been independently developed in several disparate research fields, such as human genetics, anthropology, public health/epidemiology, geography, history, demography, linguistics, computer science, economics and marketing. The common motivation in applying name origin analysis in these various disciplines is typically to reveal some hidden process in the population, with names analysis serving as a method for applied research. Therefore, although the objectives of such research might be very different, all of these disciplines have developed a set of relatively similar ‘tools of the trade’ as it were. Since the object of this thesis is to create an innovative method in population geography, the methods upon which it is based will be garnered from all the other

disciplines that have used names for the purpose of classifying populations by ethnicity. This chapter brings together the different approaches to developing these methods and the evidence to justify them for this application.

Hereinafter two types of personal names will be distinguished 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. Where the term 'names' is used on its own it will usually refers to both forenames and surnames indistinctively.

The chapter is organised in five sections. The first section reviews the human genetics literature on name analysis and weaves together the evidence of how languages, names, genes and human origins all tell a similar story about our ancestral origin and historic migration flows. The second section summarises the history of name-based ethnicity analysis in the twentieth century, reviewing its use to study historic domestic and international migration patterns as well as contemporary ethnic inequalities. The third and fourth sections carry out a systematic review of the literature with a focus on identifying and analysing the most representative research studies in which a new name-based ethnicity classification has been developed and evaluated. The third section describes how 13 studies were selected and analysed, comparing how these built their own name-based ethnicity classifications and how they applied them to target populations. The fourth section evaluates these 13 studies, and summarises the limitations and advantages of the overall methodology. Finally, the fifth section explores alternative approaches to building universal name-

based ethnicity classifications from computational, marketing and onomastic perspectives, and suggests how they offer promising methodological developments that complement the rest of the literature.

3.1. Languages, Names, Genes and Human Origins

> *'It may be worthwhile to illustrate this view of classification, by taking the case of languages. If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; and if all extinct languages, and all intermediate and slowly changing dialects, were to be included, such an arrangement would be the only possible one'* (Darwin, 1859, 422).

3.1.1. Human and language evolution

The quote above is from Charles Darwin's 'On the Origin of the Species' (1859) which included a parallelism between the evolution of languages and of humans, suggesting that the genealogical arrangement of the 'races of man' necessarily had to follow a taxonomy of languages. With subsequent advances in modern genetic techniques, population geneticists have recently demonstrated the existence of such a relationship in human evolution, mapping human origins, gene evolution, and geographical spread and intermixing across the planet and comparing it with the evolution of language and the archaeological record (Cavalli-Sforza and Cavalli-Sforza, 1995; Piazza et al, 1987).

One problem that population geneticists face when analysing the genetic linkages between human groups, is how precisely to define those groups in the first place in order to analyse genetic similarities within the group and differences between groups. Luigi Luca Cavalli-Sforza, at Stanford University is considered the 'father of population genetics', and has carried out very successful research in this area for over 40 years, summarized in his masterpiece 'The History and Geography of Human Genes' (Cavalli-Sforza et al, 1994). He has primarily used a mother language criterion to define such human groups, avoiding cases where there is known to have been a historic language replacement (e.g. Spanish imposed to Native Americans, or Finno-Urgic language to Hungarians: Cavalli-Sforza, 1997). His justification for so doing is that the classification of languages, as opposed to the classification of places or regions, or of anthropological human groups, is well standardised and commonly accepted (Cavalli-Sforza and Cavalli-Sforza, 1995). Moreover, the 6,000 or so languages currently in existence in the world may be arranged into a hierarchical taxonomy that relates each to one of six major families – that is, according to the most widely accepted language taxonomy of Greenberg and Ruhlen (Ruhlen, 1987). Population geneticists are able to compare such language trees with the genetic linkages between populations (i.e. an evolutionary tree), to corroborate the geographical spread or explain any differences with historical data (Cavalli-Sforza et al, 1988), following the path suggested by Darwin in 1859.

In the age of DNA (deoxyribonucleic acid) research, population geneticists have demonstrated that there is a continuum of human genes across continents that have migrated and intermixed over the millennia. However, the relative frequencies of certain genetic markers do vary between regions and human groups, sometimes in

very marked ways, but the way in which differences between these groups are defined in the first place, lies to a certain extent 'in the eye of the beholder' (Cavalli-Sforza, 1997). The Human Genome Diversity Project, which intends to trace human genes across the globe back to ancient migrations, has also defined human groups, called 'populations', by the common mother language of the subjects to be studied (M'charek, 2005).

There is a fascinating genetic history of human evolution that links the places and regions our ancestors inhabited, their migration flows initially out of Africa and then back and forward between continents and regions, with the way cultural heritage, including language, has developed, evolved and been transmitted from one place to another and from one generation to the next. There is a vast literature on population genetics and molecular anthropology that studies these relationships, having made great advances in disentangling ancestral human movements, cultural exchange and distant historical settlement and migrations. Various researchers have used languages, as well as surnames, to study populations genetic structure, endogamy, and the cultural evolution of populations.

This PhD will use some of the evidence collected in the analysis of surnames in the human genetics field, although it will only focus on name origin analysis to classify 'contemporary populations' according to recent migrations (their own or that of their ancestors to three or four generations back), and not ancient ones. For a full review of the literature on names and genetics see Lasker (1985) and Colantonio et al (2003).

3.1.2. Isonymy studies in genetics

The 'discovery' of a statistical relationship between family names distribution and population structure, is attributed to George Darwin (1875), son of Charles Darwin. George's parents were first cousins (relatively common amongst the more affluent classes of 19th century Victorian Britain) and he was very interested in demonstrating that there was not a statistical association between first cousin marriages and a higher frequency of mental disorder or congenital disease. To do so he computed the actual frequencies of same surname marriages compared with the true rate of first cousin marriages, and then extrapolated it to the general population in an area comparing variations with that of the rates of disorders or diseases. His most interesting discovery is in the association he established between same surname marriages with the probability of endogamy.

Several researchers throughout the 20th century have studied the relationship between surnames and population structure (Crow and Mange, 1965; Yasuda et al, 1974), building models that were then compared with the genetic evidence. The base premise in these studies is the fact that '[s]urnames are not distributed homogeneously in different places and among different social groups. The general purpose of surname studies in human biology is to measure the different probabilities of finding the same surnames in different times, places, groups and, especially, in marital partners' (Lasker, 1985: 5). This probability is defined as the degree of 'isonymy', and can be calculated between marital partners or between places or population groups. Marital partners' isonymy is not relevant to this PhD, so the meaning given here to isonymy will be that between places or regions and between human groups.

In human genetics, differences in isonymy across space or groups are compared with differences in gene frequencies to isolate common boundaries, or barriers to human exchange that create 'geographic patterns of genetic, morphologic, and linguistic variation' (Manni et al, 2004: 173). The underlying hypothesis of isonymy studies is that areas or groups that have freely exchanged individuals between them, since the time when surnames were introduced, will tend to have similar proportions of common surnames, while areas or populations that remain isolated from each other will have very distinct surname frequency distributions. This is well illustrated by Figure 3.1, which shows two maps of France comparing surname clusters derived from isonymy between departments (left) with dialect clusters derived from dialectometric distances between departments (Scapoli et al, 2005). The similarities between surname and linguistic regions are striking, given the two independent data sources.

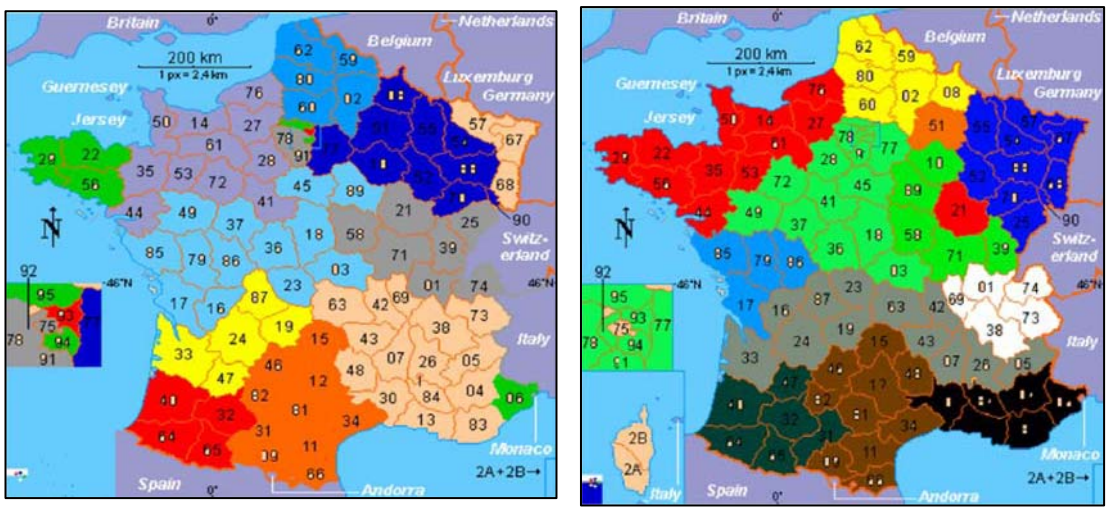

Figure 3.1: Maps of France's surname (left) and dialect (right) clusters
The map on the left shows surname clusters derived from isonymy (Lasker distance) between departments, and the map of the right represents dialect clusters derived from dialectometric distances between departments. Source: Reproduced from Scapoli et al (2005: 83-84)

The methods used to study isonymy in entire populations were initially very limited and human biology studies of surnames were traditionally reduced to the name structure of isolated villages, valleys, or ethno-religious groups (e.g. the Mormons: Jorde and Morgan, 1987), for which researchers were able manually to process individual Parish or marriage records (Lasker, 1985). However, in the early 1990s large population registers, such as Electoral Registers or telephone directories, started to become available in digital form, and this triggered an explosion of interest in analysing the name structure of whole populations. A team based at the University of Ferrara in Italy has been a pioneer in these studies, and has published their findings for the whole population of the following countries: Austria (Barrai et al, 2000); Switzerland (Barrai et al, 1996; Rodriguez-Larralde, Scapoli et al, 1998); Germany (Rodriguez-Larralde, Barrai et al, 1998) ; Italy (Manni and Barrai, 2001); Belgium (Barrai et al, 2003); the Netherlands (Manni et al, 2005); Spain (Rodriguez-Larralde et al, 2003); Venezuela (Rodriguez-Larralde et al, 2000) ; Argentina (Dipierri et al, 2005); the US (Barrai et al, 2001); and France (Scapoli et al, 2005). The same research group has produced an extensive comparative review of the surname distribution of the total population of eight countries in Western Europe (Scapoli et al, 2007) in which they conclude that the present surname structure of Western Europe is intimately linked to local languages.

This PhD author obtained the inspiration to carry out the research presented in this thesis from the link between surname structure, geography and languages established by these authors. Furthermore, at the core of this research are parallels between how these authors had tackled the problem of finding regional and structural patterns in the name distributions of entire populations using the entries in telephone directories

and Electoral Registers, and the task of classifying ethnic groups in multicultural societies using the same materials and principles.

3.1.3. DNA, surnames and population structure

Another stream of research in names and genetics has analysed the relationship between individual surnames and Y-chromosomes, since both are patrilinearly inherited (Jobling, 2001; McEvoy and Bradley, 2006). This has allowed researchers to group surnames according to regional frequencies and to study the genetic linkages between the contemporary bearers of such surnames. An example of this is shown in Figure 3.2, which plots the network of genes (haplotypes) shared between individuals with four related surnames. If the necessary caveats of name change, migration, and large differences in name frequency are taken into account, surnames are deemed to be good markers for establishing individual ancestry and for intra-group relationship histories (King et al, 2006).

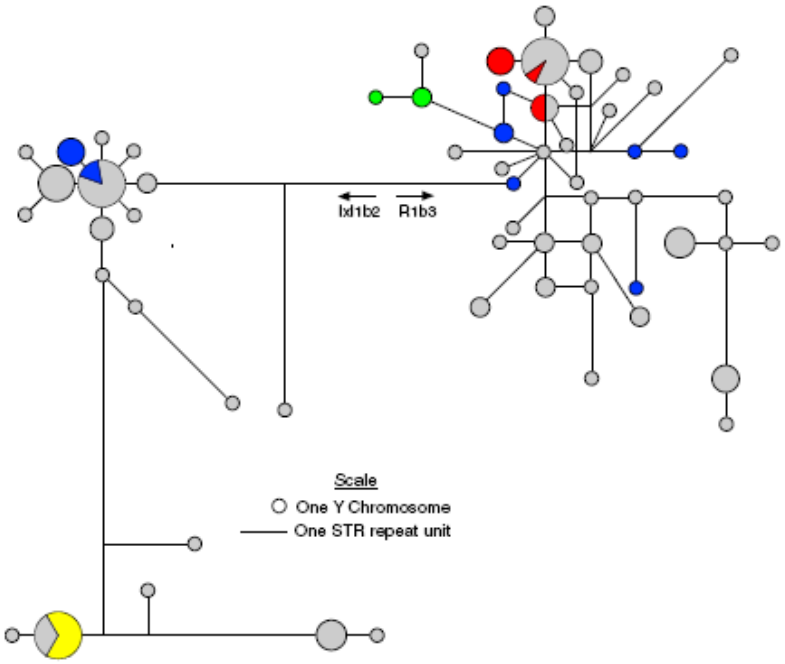

Figure 3.2: Median-joining network of Y-chromosomes in the surname McGuinness and four putatively related surnames

(Reproduced from McEvoy and Bradley, 2006: 218) The network illustrates Y-chromosome phylogeny, where each circle represents a distinct haplotype, with circle area proportional to frequency, and line length between haplotypes indicates their mutational divergence. Different colours represent different surnames with proportional 'pie-slices' for haplotypes shared across surnames. McGuinness Y-chromosomes are shown in grey (n=99), McCartan in blue (n=13), McCreesh in red (n=7), Neeson in yellow (n=8) and Guinness in green (n=3).

As a consequence of this type of relationship, a new stream of research has been rapidly developing in what has been termed the 'era of genetic genealogy' (Shriver and Kittles, 2004). People who do not know who their immediate ancestors were can search DNA databanks for their most similar genetic markers, and obtain the common surnames with those markers (Sorenson Molecular Genealogy Foundation, 2007). Recent research in this area has also suggested that the police might be able to identify a suspect's surname by his/her DNA traces (Jha, 2006).

More relevant to this PhD are applications of DNA and surnames at the group level, rather than at that of the individual. It is relevant to mention here the case of genetic medicine, a thriving and polemic field of research (BBC Radio 4, 2005) which investigates the associations between genetics and disease by isolating the complex multigene reactions to different environments. What these studies require, in order to isolate individual genes, are individuals that can be easily ascribed to similar groups according to a genetic tree. As with many other classic studies in biology and geography, researchers have turned to study islands, and not surprisingly one of the favourite places in the world for studies of genetic medicine is Iceland, a classic field study area for human and physical geographers (Haggett, 2001). The company *deCODE Genetics* has built a genealogical database of all Icelanders tracing each individual's ancestors forename and surname and place of birth back to 1650 through very detailed parish records and censuses, with the aim of isolating genetically based diseases in present day DNA through geo-historical stratification of Icelandic

population structure and their gene pool (Wade, 2005). Their results indicate that sampling strategies to isolate human groups with common origins in medical research need to take account of substructure even in a relatively homogenous genetic isolate. This fact is usually ignored in studies that sample their cases and controls from groups that share the same nationality or self-reported ethnic background, with the implicit assumption that no substructure exists within such groups (Helgason et al, 2004). This information, ultimately based on genealogical relationships, has allowed *deCODE Genetics* to test certain drugs to specific genes and produced findings that are just starting to unwind the braids binding DNA, drugs and disease (Couzin, 2005). However, this type of research is polemic because of its obvious parallels with the methods employed in 19th and early 20th century racial medicine studies to isolate discrete 'races' with a view of justifying biological determinism.

3.1.4. Wrapping up the evidence

Taken together, it is apparent that: first, surnames correlate well with Y-chromosomes at the regional and national levels; second, several genetic markers also significantly correlate with languages at a continental and global scale; third, there are distinct geographical patterns in surname structure identified through isonymy; and fourth, there is an obvious link between names and the languages from which they originate. All of this indicates that personal name analysis can offer a reliable method to ascribe individuals to common human groups, where such groups are defined as having a common linguistic, geographical and ethnic origin.

3.2. The History of Name-based Ethnicity Analysis

3.2.1. Names and domestic migration

Most of the studies of surnames distributions across space and time pertain to finding patterns, structures and changes across the regions of a single country, as opposed to international comparisons. When studying population migration, therefore most surname analysis studies have studied the nature of internal or domestic migration within a specific country, and usually over the last 200 years for which reliable written records and statistics exist. The number of studies of this kind is not very large, but they have been attempted from very different research fields, such as demography, geography, genetics, history, and linguistics.

In order to illustrate their value to this thesis a few examples of this type of study will be mentioned here, taken from France, Belgium, Britain and Spain. From a human genetic perspective, Degioanni and Darlu (2001) propose the application of a Bayesian method to estimate the probability of geographical origin of migrants in an area using surname frequencies measured in two time periods. By using birth registers in France in 1891-1915 and 1916-1940, they demonstrate the validity of using surnames to estimate the probable region of origin of migrants between these two time periods. Poulain et al (2000), used a similar approach in a historic demography study that proposed the use of patronyms (surnames derived from the father's forename, such as Johnson) to measure historic migration flows of Flemish people to Wallonia (Belgium) and Northern France in the early 20th century. This method removed the requirement for two datasets from subsequent periods of time, since frequency of Flemish patronyms, such as those starting with 'Van' or 'Ver',

will clearly indicate the major destination areas of Flemish migrants. From a geographical and geodemographic perspective, Longley et al (2007) present a set of applications of the classification of surnames in groups of regional origin (e.g. Cornish surnames) to study historic migrations and social mobility at very fine geographical scales. In a similar guise Aranda Aznar (1998) studies migration patterns in the Basque Country in Spain and of people taking Basque surnames to the rest of Spain, and the degree of intermarriage between Basque and non-Basque, using very detailed population registers.

What differentiates these three last studies (Aranda Aznar, 1998; Longley et al, 2007; Poulain et al, 2000) from the rest of studies mentioned so far in this chapter, is that rather than studying individual surnames (which is what, for example, isonymy does), they group them according to their language or culture of origin (Flemish, Cornish, Basque). This is an important step for reasons that will become clear later in this chapter, and the use of classifications of surnames into groups according to origin to analyse migration patterns will be further developed in this thesis.

In these and similar studies, surnames are proven to offer very revealing insights not only into population or regional structures over the last centuries, but also into the migration flows of specific human groups. However, whereas a substantial proportion of the surname analysis literature has focussed on domestic migration (an exception being Chiarelli, 1992 who used surnames to study the regional origins of Italians who emigrated to Toronto, Canada), this PhD is concerned with the use of names as a proxy for ethnicity, hence its primary focus on international migrants and their descendants.

3.2.2. Names and international migration

The use of people's name origins to subdivide contemporary populations into ethnic groups has been applied in population studies, especially in the US where attempts have been made to segment the 'melting pot' into its alleged original constituents, at least since the beginning of the twentieth century. Rossiter's (1909) 'A Century of Population Growth, from the First Census of the United States to the Twelfth, 1790-1900' was the first in a series of studies concerned with calculating immigration quotas, which were set according to the estimated ethnic composition of the 'original national stock' of the population of the US in the 1790 Census.

In the 1920's the US Federal Government attempted to control the size and character of the streams of immigration, and passed the Immigration Act in 1924 which established that from 1927 'the flow of immigrants should be related to the 'national origins' of the existing population' (Akenson, 1984: 103). In this context 'national origins' was the term used then for contemporary's ethnicity (McDonald and McDonald, 1980). Therefore the government needed to establish a baseline determining what was the original stock of white population in the 18th century when the country was created (US Senate, 1928), and the only method available was to examine the origins of surnames in the individual responses of the first US Census of Population, that of 1790 (Purvis, 1984).

A thorough study was commissioned to the American Council of Learned Societies, and led by Howard Barker, who was a leading linguistic scholar specialising in names and very keen on using name frequency statistics (Barker, 1926; 1928). The results were published in 1932 as the 'Report of Committee on Linguistic and

National Stocks in the Population of the United States' (American Council of Learned Societies, 1932). Therefore, 'fundamental to a determination of who would be let into the United States from the late 1920s (…) until 1965 (…) was an analysis of the ethnic character of the American population in 1790' (Akenson, 1984: 103), more precisely through the surnames origin of the 1790 population.

	Estimate by	
National Origin (1790 Census)	**Rossiter (1909) %**	**ACLS (1932) %**
English and Welsh	82.1	60.1
Scottish	7.0	8.1
Irish	1.9	9.5
German	5.6	8.6
Dutch	2.5	3.1
French	0.6	2.3
Swedish	n.a.	0.7
Spanish	n.a.	0.8
All others/ unassigned	0.3	6.8
Total	100	100

Table 3.1: Estimates of national origin breakdown of the US white population in 1790, using surnames
Source: Rossiter's (1909) reported in Akenson (1984: 103) and the American Council of Learned Societies (ACLS) (1932: 124)

The results of both the Rossiter (1909) and the American Council of Learned Societies (1932) studies are summarised in Table 3.1. However, both studies have been heavily criticized for being inherently flawed and full of errors, with biases that served well the purpose of limiting 'undesired' migration (Akenson, 1984; McDonald and McDonald, 1980; Petersen, 2001; Purvis, 1984). Amongst the major critiques, are the lack of expertise in name transformations (Anglicisation, transliteration, transcription, etc) that the clerks who undertook the work had, the use of non-random sampling, and the attempts to make international comparisons of

name frequencies using different time periods and sources of various qualities (Akenson, 1984; McDonald and McDonald, 1980).

Despite the problems in their methods, both studies set an important precedent justifying their approach on '[t]he fundamental assumption (…) that in the absence of direct data on ethnicity, surnames provided the most accurate of all possible informational surrogates' (Akenson, 1984: 103). This is actually the same assumption upon which this PhD thesis is based, and the literature since these studies in the early twentieth century will be analysed in the following sections.

3.2.3. Names and ethnicity

Following the Second World War, there were large population movements in Western Europe and the US, political boundaries were re-drawn, new nationality and citizenship rules were applied to migrants, and ethnic minorities and refugees started to be recognised. In this context the concept of ethnicity started to take root displacing that of national origins or country of birth, as an expedient to describe first and second generation migrants.

At that time names origin analysis began to be used to ascribe ethnicity, and to be validated in the fields of demography and public health, especially with respect to US Hispanic populations (US Bureau of the Census, 1953; Winnie, 1960). The key factor for the early success of surname and ethnicity analysis in the US was that the Census Bureau was involved in the development and validation of these techniques, lending robust official support to the use of these methods and their derived statistical results.

A key figure was Robert Buechley, an epidemiologist who conducted studies on the health of Mexican migrants using Spanish surnames in the 1950's and 60's (Buechley et al, 1957). He used counts of Spanish surnames by area as provided by the US Census Bureau in 1950 and 1960 for the five southern border states, and used these figures as denominators to calculate the incidence and prevalence rates of certain conditions of Mexicans in California relative to the general population (Buechley, 1961). He quickly realised that 'the difficulties arise in explicit listing and definition of 'Spanish Surname'' (Buechley, 1961: 88), and devoted several studies to overcoming them (Buechley, 1961; 1967; 1976).

This work was paralleled by statisticians in the US Census Bureau, who for over forty years kept improving the official Spanish Surname list (US Bureau of the Census, 1953), using census country of birth information, geographical distribution analysis and text string mining (Fernandez, 1975; Word et al, 1978). This work resulted in the widely used Word-Passel Spanish surname list in the 1980's (Passel and Word, 1980; US Bureau of the Census, 1980) and the Word-Perkins surname list in the 1990's (Perkins, 1993; Word and Perkins, 1996) which have both been distributed as official statistics by the US Census Bureau (US Census Bureau, 2006), and used by many researchers in population and public health studies. Similar attempts were made by the US Census Bureau to produce a list of Asian surnames (Passel et al, 1982) which was used as a sampling frame for the Survey of Minority-Owned Business Enterprises (SMOBE) using both forenames and surnames (Abrahamse et al, 1994). However, no Asian surnames list has been published or documented in the same way as the Spanish surnames list (Lauderdale and Kestenbaum, 2000).

Many other researchers have developed and applied these techniques, and interest has grown very rapidly through the past 30 years, following increasing relevance of research in international migration, improvements in computer processing power, and (most importantly) with the wider availability of digital name datasets covering entire populations at the individual person level. Given the level of interest in name-based techniques, and the known limitations to their accuracy (Choi et al, 1993), a few studies have concentrated upon measuring the accuracy of different name-based ethnicity classification methods, a stream of research that was opened by Nicoll et al (1986) and which has been sustained over time (Nanchahal et al, 2001). There is a vast range of studies that developed, evaluated or applied these techniques, and therefore the purpose of the next two sections is to carry out a thorough review of the literature of those studies that developed their own surname classification methods, comparing them in a systematic way.

3.3. Name-based Ethnicity Analysis: Building the Classifications

A literature search and systematic review has been carried out to identify the most representative research papers that specifically deal with the problem of classifying lists of names of individuals into ethnic groups through the development of new methods, and that provide a full evaluation of their accuracy. The objective of this and the following section is to bring together isolated efforts in the literature and provide a coherent comparison, a common methodology and terminology in order to identify new research gaps to be tackled in the rest of the thesis. Through this review,

the methodological commonalities, achievements and shortcomings of the selected studies have been extracted.

This section presents a summary of this review, the main characteristics of the studies evaluated, and how they built the name to ethnicity classification. The following section will then separately analyse their evaluation and the results of the comparison.

3.3.1. Literature review

The literature search was carried out using three databases of scholarly publications; PubMed Medline, the ISI Web of Knowledge (CrossSearch), and Google Scholar. The keywords and search string used to search these databases were:

(1) [ethnic* OR race OR racial OR minorit* OR migrant* OR immigrant*]; in the title, keywords or abstract of the publication (abstract not used for Google Scholar)

AND

(2) [name* OR surname* OR forename*]; only in the title or keywords of the publication (due to the common use of the word 'name' in abstracts).

This search retrieved 186 unique publications at the time (January 2006).

The *inclusion criteria* were to select any study; (a) that developed or used a name-based ethnicity classification method to subdivide contemporary populations at the individual level, and (b) that evaluated its accuracy in a systematic way. On the other hand, the *exclusion criteria* were; (a) studies that neither offered a new method of name-based ethnicity classification, nor evaluated a previously developed method that had not been tested before; (b) studies that did not validate the classification

using an alternative ethnicity information source (i.e. non-name-based); (c) studies that provided insufficient detail of their research process and results as to support this systematic review, for which at least the method's sensitivity and specificity needed to be explicit, and (d) studies that were not published in English.

The 186 publications retrieved by the search were filtered through a three-tier process. First, potentially relevant publications were evaluated against the inclusion criteria, using solely the information offered in their title, with non-relevant publications being rejected, most of them using surnames in the genetic domain to study ancient migrations or isonomy. In cases of doubt, the publication was left included in this phase. This reduced the number of publications to 129. Second, these were then evaluated against the exclusion criteria using the information provided in their abstract, which reduced the number of selected publications to 37. Finally, the full text of these 37 publications was analysed against the exclusion criteria, ending up with 11 publications that met all the selection criteria. These 11 publications were analysed in-depth, and all of their references were retrieved and also checked against the inclusion and exclusion criteria. This last step contributed two additional publications that were not found by the original search, one of them because the word 'name' or its equivalents did not appear either in the title or in the keywords (Sheth et al, 1999), and the second because it is a government report only published on-line (Word and Perkins, 1996).

Paper Reference	Geographical area of study *Country and (Region)*	Ethnic Minorities (E.M.) classified	Name to Ethnicity Assignment	
			Method --- ***A**utomatic* ***M**anual*	***Name components*** --- ***S**urname* ***F**orename* ***M**iddle name*
Choi, *et al* (1993)	Canada (Ontario)	Chinese	A	S
Coldman, Braun & Gallagher (1988)	Canada (British Columbia)	Chinese	A	F, S, M
Lauderdale & Kestenbaum (2000)	US (National)	Chinese, Japanese, Filipino, Korean, Indian, & Vietnamese	A	S
Razum, Zeeb, & Akgun (2001)	Germany (Rhineland-Palatinate & Saarland)	Turkish	A	F, S
Word & Perkins (1996) / Stewart et al (1999)	US (National)	Hispanic	A	S
Harding, Dews, & Simpson (1999)	UK (Bradford & Coventry)	South Asian + Hindu, Muslim & Sikh	A	F, S
Cummins, *et al* (1999)	UK (Thames, Trent, W.Midlands & Yorkshire)	South Asian	A	F, S
Nanchahal, *et al* (2001)	UK (London, W.Midlands, Glasgow)	South Asian	A	F, S, M
Sheth, *et al* (1997)	Canada (National)	South Asian and Chinese	A/M	S
Martineau & White (1998)	UK (Newcastle; 4 General Practices)	Bangladeshi, Pakistani, Indian Muslims, Non-South Asian Muslims, Sikh, Hindu, White, Other	M	F, S and Gender
Bouwhuis &. Moll (2003)	Netherlands (Rotterdam; 1 Hospital)	Turkish, Moroccan, Surinamese	M	F, S
Nicoll, Bassett, & Ulijaszek (1986)	UK (Selected areas)	South Asian	M	F, S
Harland, White & Bhopal (1997)	UK (Newcastle)	Chinese	M	F, S

Table 3.2: Summary of the general characteristics of the 13 studies reviewed
Method of name to ethnicity assignment: 'A' = Automatic, 'M' = Manual. Name components used in the classification; 'S'= Surname, 'F'= Forename, 'M'= Middle Name.

The final selection of publications consisted of 13 papers representing five countries (Canada, Germany, Netherlands, UK, and the US), and most of them from the field

of public health. Table 3.2 shows the key characteristics of these studies, whose findings will be analysed in the following sections. The subsets of ethnic minorities studied represent the biggest and most recently arrived groups in each country: (a) South Asians (Indian, Pakistanis, Bangladeshis, Sri Lankans); (b) Chinese; (c) other East and South-east Asians (Vietnamese, Japanese, Korean, and Filipino); (d) Hispanics; (e) Turks; and (f) Moroccans (see third column in Table 3.2 for the correspondence between these groups and each study).

Amongst the publications excluded in the last phase of the selection strategy (n=26) there were some other interesting research papers in which an independent name-based approach was developed, although not explicitly explained or independently evaluated. However, some of these studies are worth mentioning, since they typically used telephone directories to select names from a particular ethnic group as a sampling strategy for their surveys, showing the usefulness of the name-based approach to classify Vietnamese (Hinton et al, 1998; Rahman et al, 2005), Korean (Hofstetter et al, 2004), Cambodian (Tu et al, 2002), Chinese (Hage et al, 1990; Lai, 2004), South Asian (Chaudhry et al, 2003), Japanese (Kitano et al, 1988), Irish (Abbotts et al, 1999), Jewish (Himmelfarb et al, 1983) Iranian (Yavari et al, 2005) and Lebanese (Rissel et al, 1999) names, in the US, Canada, UK and Australia.

3.3.2. Structure of the selected studies

The 13 selected papers aimed to demonstrate a satisfactory accuracy rate in separating individuals of either one, or just a few, ethnic minority groups from the rest of the resident population in some developed countries. None of them tried to classify the whole population into all of the potential ethnic groups in a country, something that remains a research gap. The studies differ substantially in the sizes of

the target populations to be classified (from 137 to 1.9 million people), the numbers of unique forenames or surnames in the reference list used in the search (from fewer than 100 to 27,000 names), and hence the method to allocate them (manual vs. automatic classification). However, each of the studies includes a number of common methodological processes and research components: firstly a name *reference list* is independently built or sourced from another study or from 'an expert'; secondly a separate *target population* is manually or automatically classified into ethnic groups; and thirdly the *accuracy* of the method is *evaluated* against a previously known 'gold standard' for ethnicity in the target population. These common structures and processes are summarised as a flow chart in Figure 3.3.

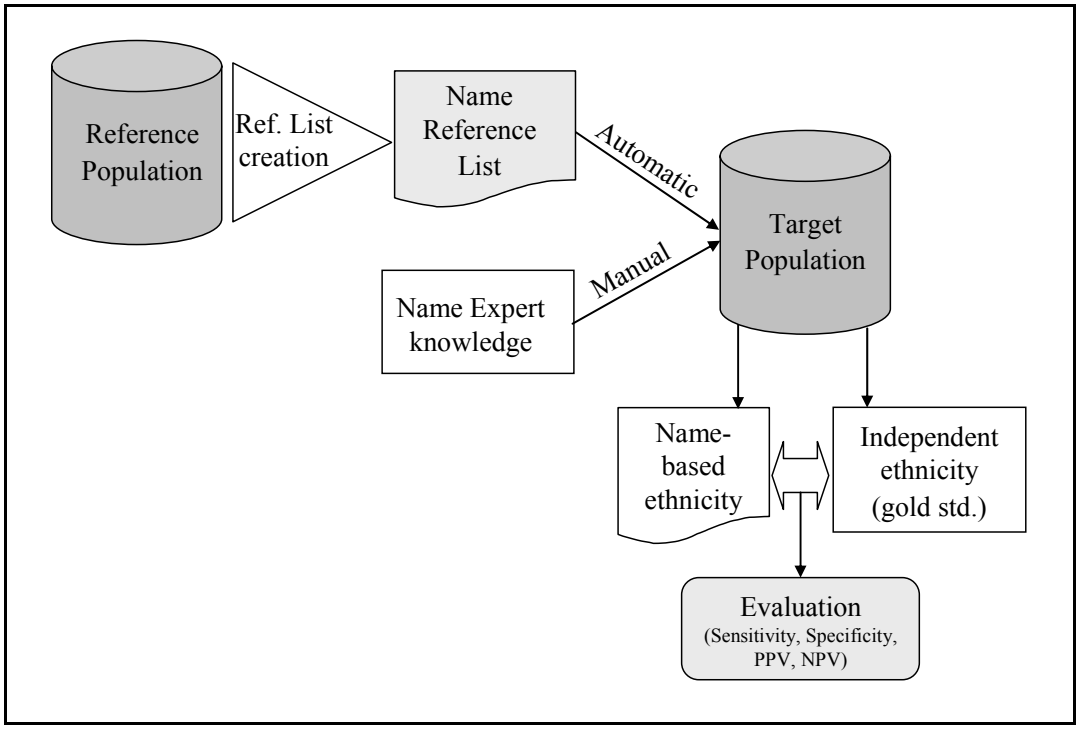

Figure 3.3: Common structures and processes of name classifications

3.3.3. Source data, reference and target populations

The primary source material for each of the studies are datasets of individuals' personal data that are usually sourced from population administrative files, health

registers or surveys. *Target population* is the term given to the list of individuals to be classified into ethnic groups using their names, either manually or automatically. Automatic classification methods require an independent *reference list* of surnames or forenames with their pre-determined ethnic origin, which is used to perform the computerised search and allocation of ethnicity for each individual in the target population (in the manual methods the equivalent to the *reference list* is the expert's knowledge). This distinction between *reference* and *target* lists of names is key to the understanding of the methodologies analysed here.

3.3.4. Building reference lists

The first step thus involves building *reference lists* or borrowing them from previous studies. These would typically include several hundreds or thousands of surnames, each one of them with a pre-assigned ethnic group (e.g. Nguyen - Vietnamese; Chang - Chinese). The characteristics of how the reference lists in the eight studies that used automatic classification were developed are further detailed in Table 3.3. Two of these studies used a software application previously developed to identify South Asian names in the UK, *Nam Pehchan* (Cummins et al, 1999; Harding et al, 1999), which contains 2,995 unique South Asian surnames, and was derived from the Linguistic Minorities Project (1985). The study of Nanchahal et al (2001) developed similar software called *SANGRA,* but did not offer sufficient information about how they built their reference list of 9,422 South Asian names. In the remaining five studies, purpose-built reference lists were constructed, containing between 427 and 25,276 unique surnames. These *reference lists* were typically built from a source independent of the *target population*, a second population generally described as the *reference population* (see the left half of Table 3.3). Exceptions to this are Choi et al

(1993) and Coldman et al (1988), with important consequences for their results, as will be discussed below.

Paper Reference	Reference Population					Reference List		
	Total Population	E.M. pop. identified	% E.M.	Source	Dates	Production Method	Nr. Unique E.M. Surnames	E.M. people / Surname
Choi, *et al* (1993)	270,139	1,899	0.7%	Mortality database	1982-1989	Country of Birth + Manual cleansing	427	4.4
Coldman, Braun & Gallagher (1988)	203,354	5,430	2.7%	Death registrations	1950-1964	Ethnicity (family)	544	16 (Chinese) 1.7 (other)
Lauderdale & Kestenbaum (2000)	1,765,422	1,609,679	91.2%	Social Security Card Applications (MBR)	Born <1941	Country of Birth	27,000	59.6 (avg.)
Razum, Zeeb, & Akgun (2001)	4,000,000	108,500	2.7%	Rhineland-Palatinate Population Register	c.2000	Nationality + Manual cleansing	12,188	12.8 (in Germany) / 3.1 (in Turkey)
Sheth, *et al* (1997)	2,782,000 (estimated)	N/K	N/K	Canadian Mortality Data Base (CMBD)	1979-1993	Country of Birth (deceased & parents)	4,271	N/K
Word & Perkins (1996) / Stewart et al (1999)	5,609,592 people; 1,868,781 households.	597,533	10.7%	1990 US Census Post-enumeration Sample	US Census Day 1990	Ethnicity (self-assigned)	25,276	23.6 (avg.)
Harding, Dews, & Simpson (1999)	List of 2,995 surnames in *Nam Pehchan* program			*Nam Pehchan* program	1981-1998	Experts' knowledge	2,995	N/A
Cummins, *et al* (1999)	List of 2,995 surnames in *Nam Pehchan* program			*Nam Pehchan* program	1981-1998	Experts' knowledge	2,995	N/A
Nanchahal, *et al* (2001)	List of 9,422 surnames in *SANGRA* program			Surveys and Hospital Records	1995-1999	From list of voluntary organisations and ONS	9,422	N/A

Table 3.3: Characteristics of reference populations and reference lists in the automatic methods
E.M. = Ethnic Minority, N/K= Not Known, N/A= Not Available.
Reference Population: 'Total population' is the input dataset used, of which 'E.M. population identified' is the ethnic minority population identified within the 'total population'. Reference List: 'Production Method' is the technique or piece of ethnicity information in the reference population used to produce the reference list; 'Nr. Unique E.M. Surnames' is the final number of ethnic minority surnames present in the reference list. 'E.M. People / Surname' is the average number of people of the ethnic minority sharing the same surname (column 3 / column 8).

Despite big differences in the sizes of the reference populations, the methods employed to derive the name reference lists were broadly similar. Generally, they all used some type of 'ethnic origin information' in the reference population, such as

self-reported ethnicity, country of birth, or nationality, to classify individuals into ethnic groups, and they then aggregated them by surname and produced a frequency count for each surname and ethnic group combination (and the same for forenames when available). Each surname or forename was then assigned to the ethnic group with the highest frequency, using a series of rules or thresholds in some cases (Lauderdale and Kestenbaum, 2000; Word and Perkins, 1996), producing the final *reference list*.

In general, there are four factors affecting the accuracy and coverage of the *reference list*, as will be explained in the accuracy evaluation section: the independence between reference and target populations, the size of the reference population, its spatio-temporal coverage (the countries and regions where it was sourced and the time period covered), and the method used to ascribe ethnicity (using proxies vs. self-reported ethnicity). Therefore, the desired qualities of the reference list are to be large enough to maximise coverage in the target population, and accurate enough as to minimise misclassifications (Coldman et al, 1988; Nanchahal et al, 2001). These two qualities are usually mutually exclusive, and hence there is a trade-off to be made between extra coverage of a larger number of names and marginal extra accuracy of the classification, as each extra name tends to be rarer than the last. The final decision concerning the size of the reference list will depend on each specific type of application. A similar issue arises regarding the nominal resolution of the ethnic group categorisations used: the finer the groups that are defined (e.g. Hindu, Bengali, Tamil, Urdu, Gujarati, Punjabi, vs. 'Indian' or 'South Asian'), the less accurate the name classification becomes, and vice versa.

3.3.5. Minimum size of the reference list

For calculating the ideal size of the reference population from which a robust reference list will be produced, the best attempt has been proposed by Cook et al (1972: 40) using the following formula:

$$n \geq \frac{\log(1-x)}{\log y}$$

where n is the required minimum size of the reference population, x is the desired level of confidence for the allocation of an individual to his or her appropriate ethnic group, and y is the required level of confidence that a particular surname will perform as desired. For example, for x= 80% and y=95% the minimum size of the reference population required will be $n \geq 13.4$, meaning that for every surname to be classified a list of at least 13.4 individuals with that surname and their known ethnicity is required within the reference population.

The minimum value of n (in the above example equal to 13.4) refers to the unlikely situation that all individuals with the same surname in the reference population had the same ethnicity, and hence the size would have to be extended in proportion to the 'noise' found in each specific reference population. Cook et al (1972) proposed multiplying n by a -'rule of thumb'- factor of 4 to obtain a realistic reference population size. The actual reference population sizes used in the five studies evaluated here, that built their own reference lists, have been compared against these two 'Cook et al criteria': *first criterion*; n=13.4 people per surname, and *expanded criterion*; n=13.4 x 4= 53.6 people per surname and the results are presented in Table 3.4. It is surprising to find that only two of the five studies' reference populations satisfy the first 'Cook first criterion' (Lauderdale and Kestenbaum, 2000; Word and

Perkins, 1996), with the remaining three below 75% of the required size. Moreover, only one satisfies the 'Cook expanded criterion' (Lauderdale and Kestenbaum, 2000), with the rest below 45% of the required minimum reference population size.

Paper Reference	Reference List	Ethnic Minority Reference Population Size (Nr. People)				
			Minimum Ref. Pop. Size Required			
	Nr. Unique Ethnic Minority Surnames	Actual Ref. Pop. Size Used	Cook first criterion (13.4)	Actual size as % of Cook first criterion	Cook expanded criterion (13.4x4)	Actual size as % of Cook expanded criterion
	a	*b*	*c* = *a* * 13.4	*b/c*	*d* = *a* * 13.4 * 4	*b/c*
Choi, *et al* (1993)	427	1,899	5,722	33%	22,887	8%
Coldman, Braun & Gallagher (1988)	544	5,430	7,290	74%	29,158	19%
Lauderdale & Kestenbaum (2000)	27,000	1,609,679	**361,800**	**445%**	**1,447,200**	**111%**
Razum, Zeeb, & Akgun (2001)	12,188	108,500	163,319	66%	653,277	17%
Word & Perkins (1996) / Stewart et al (1999)	25,276	597,533	**338,698**	**176%**	1,354,794	44%

Table 3.4: Comparison of actual reference population sizes used in five studies with the minimum reference population size criterion established by Cook et al (1972)
The five studies included are the only ones that developed their own name reference lists from reference populations. Actual reference population size used in each study is compared against two Cook et al criteria: first criterion; n=13.4 people per surname, and expanded criterion; n=13.4 x 4= 53.6 people per surname. Only two studies satisfy the first 'Cook first criterion' and only one satisfies the 'Cook expanded criterion' (highlighted in bold).

3.3.6. Classification of target populations

The second step in the 13 studies analysed consisted of classifying the target population into ethnic groups, using either a manual (i.e. human expert) or an

automatic method (through computer algorithms). The characteristics of the target populations selected in each of the 13 studies are summarised in Table 3.5 ('Target Population' section).

Manual methods have the advantage of not requiring a name reference list and also of being amenable to a rich number of 'fuzzy rules' that the experts performing the classification can apply in order to decide the group into which an individual should be assigned. However, the manual method has a series of major limitations, the main one being that it is cumbersome and time-consuming (Bouwhuis and Moll, 2003) and this seriously constrains the size of the target population to be coded. In order to increment the number of individuals to be coded, additional experts need to be recruited, which also causes inconsistency in the subjective decisions taken by different human subjects. Additionally, most of the manual classification studies focus on a two-group classification problem, which only requires a simple binary decision on whether the individual belongs to a specific ethnic minority group or not, but when more groups are introduced, several experts from different cultural backgrounds are required, and hence the number of misclassifications quickly rises, especially when names overlap across similar ethnic groups (Martineau and White, 1998). For these reasons, no further specific attention will be given here to those studies using manual methods (last four papers in Table 3.2).

On the other hand, automatic methods to classify target populations rely on the availability of an appropriate name reference list. The studies analysed here applied an automated algorithm to search for the name of each individual in the target population against the reference list, and then assign the pre-coded ethnic group for

that name to the individual. One of the main differences between the studies is whether they used only one name component of the individual (surname) or more (forename and surname, or even middle name) (see last column of Table 3.2 for details). *Nam Pehchan* includes a set of rules that use name stems if the name has no match in the reference list (Cummins et al, 1999), but this is avoided by *SANGRA* since it is deemed to produce an unacceptable number of false positives (Nanchahal et al, 2001).

A second difference between studies is whether one or several ethnic groups are to be classified. It must be emphasised that almost all of the studies that used automatic classification were designed to classify individuals using a binary taxonomy in mind, that seeks to identify members of a particular minority group or macro group (i.e. South Asians) from a general population. The exception is Lauderdale and Kestenbaum (2000) who classify six substantially different Asian ethnic groups (Chinese, Vietnamese, Japanese, Korean, Asian Indian and Filipinos). A third difference, is the use of certain name scores or thresholds related to the strength of the association between each name and the ethnic group of origin (e.g. heavily Spanish, moderate Spanish, etc.), to the final user's advantage when fine-tuning the classification to their specific target population and purpose. Only two studies use such thresholds (Lauderdale and Kestenbaum, 2000; Word and Perkins, 1996).

Paper Reference	Division of Reference & Target Population	Target Population					Method Evaluation (*single value or a range*)			
		Total Population	Nr. E.M. classified	% E.M.	Source	Dates	Ethnicity Gold Standard	Sensitivity	Specificity	PPV	NPV
Choi, *et al* (1993)	Random split	270,138	1,910	0.7%	Same as Reference	1982-1989	Country of Birth	0.73	N/K	0.81 - 0.84	N/K
Coldman, Braun & Gallagher (1988)	Chronological split sample	155,629	3,205	2.1%	Same as Reference	1965-1973	Ethnicity	0.89-0.97	1.00	N/K	N/K
Lauderdale & Kestenbaum (2000)	Different sources	1,900,000	N/K	N/K	1990 US Census Sample	1990	Ethnicity	0.55 - 0.70	N/K	0.76 - 0.83	N/K
Razum, Zeeb, & Akgun (2001)	Different sources	NK	192	N/K	Saarland Population Register	c.2000	Nationality	0.40 - 0.84	0.99	0.14 - 0.98	1.00
Word & Perkins (1996) / Stewart et al (1999)	Different research papers	7,232	780	10.8%	Greater Bay Area Cancer Register	1990	Ethnicity (self-reported)	0.61	0.98	0.70	0.96
Sheth, *et al* (1997)	Different sources	200	100	50%	Telephone survey	1990s	Ethnicity (self-reported)	0.96	0.95	N/K	N/K
Harding, Dews, & Simpson (1999)	Different sources	275,353	6,585	2.4%	a) Resident Survey, b) School Survey, c) Death Register, d) Census Longitudinal Study	1981-1998	Ethnicity [self-rep. (a)&(d) parents(b)], c)Visual inspection	0.94	0.99	0.96	N/K
Cummins, *et al* (1999)	Different sources	356,555	3,845	1.1%	Thames, Trent, W. Midlands & Yorkshire Cancer registers	1990-1992	Visual inspection + computerised dictionary	0.90	N/K	0.63	N/K
Nanchahal, *et al* (2001)	Different sources	130,993	15,390	11.7%	London and Midlands Hospital Admissions	1995-1999	Ethnicity (self-reported)	0.89 - 0.96	0.94 - 0.98	0.80 - 0.89	0.98 - 0.99
Martineau & White (1998)	N/A	137	107	78.1%	Family Health Service Authority Register (FHSA)	Born Oct 93 - Sep 94	Ethnicity (3rd party reported)	0.87- 0.98 (outlier 0.5)	0.60 - 0.97	N/K	N/K
Bouwhuis &. Moll (2003)	N/A	335	99	29.6%	Hospital Internal Survey to parents of children	Sep - Dec 99	Parents' country of birth (COB)	0.40 - 0.95	0.80 - 0.99	0.61 - 0.86	N/K
Nicoll, Bassett, & Ulijaszek (1986)	N/A	846	348	41.1%	(a)Child Register, (b)School Survey (c)Stillbirth Certificate	N/K	Ethnicity [(3rd pty. (a),parents (b)]; Mother COB (c)	0.67-1.00	0.92 - 1.00	0.72-1.00	0.96-1.00
Harland, White & Bhopal (1997)	N/A	129,914	1,702	1.3%	Family Health Service Authority Register (FHSA)	1991	Individual contact	N/K	1.00	0.95	N/K

Table 3.5: Summary of target population characteristics and results of the evaluation of classification accuracy in the 13 papers reviewed
'E.M.' = Ethnic Minorities; COB = Country of Birth; 'N/K' = Not Known; 'PPV'= Positive Predictive Value; 'NPV'= Negative Predictive Value.
A range of values is included here when a study reports several values of results for different subpopulations (e.g. by gender or ethnic group), or under different evaluation criteria

3.4. Name-based Ethnicity Analysis: Evaluating the Classifications

All of the 13 studies measure the accuracy of the name-based classification, by comparing it to a 'gold standard' for the ethnicity of the individuals in the target population, which had to be previously known through an independent source (the exception is Word and Perkins, 1996, but another study that evaluates their method is used here: Stewart et al 1999). This 'gold standard' is either the person's ethnicity (self-reported, by a next-of-kin, or by a third party), or a proxy for it such as country of birth or nationality (of the person or of his/her parents), all of which are assumed to represent the individual's 'true ethnicity'. However, such assumption should be interpreted with caution, as an objective entity such as the 'true ethnicity' does not exist, and hence *'there can be no such thing as a completely correct method of classifying individuals into ethnic groups'* (Cook et al, 1972 : 39), but to a certain extent a more appropriate one.

3.4.1. Accuracy evaluation

The studies reviewed here self-evaluated their accuracy using the epidemiological measures of *sensitivity*, *specificity*, *positive predictive value* (PPV), and *negative predicted value* (NPV). *Sensitivity*, is the proportion of members of 'Ethnic Group X' (gold standard) who were correctly classified as such; *specificity*, the proportion of members of 'Other 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 'Other Ethnic Groups' (predicted) who were actually from 'Other Ethnic Groups'. These concepts are better explained in Table 3.6 in a more visual fashion using a 'confusion matrix'

(Longley et al, 2005). Any classification's objective is to maximize the number of correct classifications across the main diagonal ('a' and 'd') and to minimise the number of misclassifications ('b' and 'c').

	Gold Standard ('true' ethnicity)	
Classification (predicted ethnicity)	Ethnic Group X	Other Ethnic Groups
Ethnic Group X	**a**	b
Other Ethnic groups	c	**d**

Measures of classification accuracy:

Sensitivity = a / (a + c)

Specificity = d / (b + d)

Positive Predictive Value (PPV) = a / (a + b)

Negative Predictive Value (NPV) = d/ (c + d)

Table 3.6: Explanation of measures of classification accuracy: Sensitivity, Specificity, PPV and NPV

The results for these four variables in the 13 studies are given in Table 3.5 ('Method Evaluation' section) and a range of values is offered where the study evaluated different populations, or made separate evaluations for subpopulations (e.g. by gender). If certain isolated outliers are excluded, the *sensitivity* varies between 0.67 and 0.95, the *specificity* between 0.8 and 1, the *PPV* between 0.7 and 0.96, and the *NPV* between 0.96 and 1 (only reported in four studies).

It is striking to notice that there are no substantial differences between the accuracy of the manual (bottom four in Table 3.5) and automatic classification methods, removing the theoretical advantage, in accuracy terms, of the former over the latter. In general the studies tend to reach a high specificity and NPV (near to 1), to the detriment of a slightly lower sensitivity and PPV (e.g. see Razum et al, 2001), a fact linked to the aforementioned trade-off between the extra coverage of a classification

and its marginal extra accuracy. The differences between the statistics of the 13 studies do not seem to imply substantial differences in the quality of the methods adopted. Rather, they reflect variations between the degree of distinctiveness of each subpopulation's names in the particular context of the general population studied, as well as constraints imposed by the characteristics of the datasets used.

All authors read into these results a validation of the name-based classification method to ascribe ethnicity, when other data sources are not available, giving further details of their advantages and the limitations found which will be discussed in the next two sections. However, one could argue the factor of publication bias, by which studies that did not achieve satisfactory results may have not been published.

3.4.2. Limitations found in the methodology

The 13 studies list a series of issues and limitations, many of them common between them, which are summarised below complementing them with other studies (Jobling, 2001; Senior and Bhopal, 1994) under the following eight major themes:

(a) *Temporal differences in name distribution between the reference and target populations:* different migration waves and changing geographical distributions through time, introduces misclassification and reduced coverage in classifications. For example, Lauderdale and Kestenbaum (2000) used a population reference list of people born in Asia before 1941, which might not represent the current distribution of common Asian names in the US across all age groups, and a similar problem is present in Coldman et al (1988) with Chinese names in Canada.

(b) Regional differences in the frequency distribution of names, whether these are *between* the origin and the host country, *within* either of them, or between *different host countries*. Such differences arise from geo-historical processes and migration flows. If this heterogeneity in name distribution is ignored when sampling the reference population, the subsequent name reference lists will be biased and names from a single region might not represent well the names present in other regions. Some examples found are: different Pakistani names present in the north of England, compared with the South East (Cummins et al, 1999); different Turkish names between a region in Germany and Istanbul, Turkey (2001); or Chinese migrant names that seem to be common in Australia but not so in Canada (Choi et al, 1993).

(c) Differences in the average frequency of surnames (i.e. the average ratio of people per surname). The following differences in the average frequency of surnames have been observed; between the ethnic minority (typically with higher average surname frequencies) and the host population (with a lower average), and between the ethnic minority in the host country (with a higher average) and in the origin country (with a lower average). These differences are depicted in the last column of Table 3.3: 'E.M. People / Surname'. This asymmetry is caused by a combination of the phenomenon of 'family autocorrelation' in the data (Lasker, 1997), and the uneven initial distribution of migrant names arising because of selective migration (a few initial names that can be rare in the origin country but grow rapidly because of intra-group marriages in the host country, or transcription and transliteration issues). This invites the false assumption that a common name in the host country might also

be common in the origin country, which together with item (b) above makes a strong case for sourcing name reference lists from the entire population of both the origin and host countries.

(d) Name normalisation issues; data entry misspellings, forename and surname inversions and name corruptions, all need to be normalised both in the reference and target populations in order to cleanse the datasets. However, such normalisation entails making the difficult decision whether to keep the ones that might be accepted as official names, even for several generations (Lasker, 1985). This could arise through different *transcriptions* of a name into a different language's alphabet and/or pronunciation (called *transliteration*); and creates name duplications and long lists of name variants that present a barrier to the accuracy of the reference lists. This problem is linked to other processes of name change, the 'acculturation of a name' in a host country, and the degree of inter-marriages between groups, which are all well documented for 'older' immigrant groups in the US such as Norwegians (Kimmerle, 1942), Finnish (Kolehmainen, 1939), Italian (Fucilla, 1943) or Polish (Lyra, 1966). In a non-research context this lack of name normalisation has serious consequences for tracing individuals worldwide in an era of 'global terrorism' (The Economist, 2007b).

(e) Names usually *only reflect patrilineal heritage*; and thus the methodology assumes a high degree of group endogamy, and is incapable of identifying mixed ethnicity or women's ethnicity in mixed marriages (when maiden names are unavailable) (Harland et al, 1997). If exogamy increases, as is anticipated in

the near future, the method's discriminatory ability may decline. This has already happened in highly mixed populations such as the US or Argentina, where more than three generations have passed since immigration of the traditional European migrant groups. In such instances, populations are assimilated into the general population, and the male surnames that are passed on do not normally reflect a perceived ethnic identity (Petersen, 2001), although distinct (fore)naming practices nevertheless do survive after generations (Tucker, 2003).

(f) There are *different histories of name* adoption, naming conventions and surname change that vary from country to country (e.g. Caribbeans have British surnames, Spanish women do not change surname at marriage), leading to the overlapping of certain names between ethnic groups (Martineau and White, 1998) which is difficult to accommodate in a single classification.

All of the above issues result in *differences in the strength of association* of a particular name with an ethnic group, measured by the proportion of people with a name ascribed to a certain ethnic group that actually consider themselves to be from that ethnic group. The effects of issues (a) (b) and (c) can be mitigated by sourcing broad reference populations from both the origin and host country and from a wide enough time period, using the Cook et al (1972) formula mentioned above to calculate its minimum size. This would ensure that the name reference list would reflect all of the potential names and true frequencies from the regions of the origin and host countries in more equal probability than has been the case with the methods analysed here. Moreover, when aggregating the reference population by household

surname, the issue of family autocorrelation can be avoided (Word and Perkins, 1996). The effects of issues (d) to (h) can be ameliorated by the use of 'name scores' to measure the strength of the association between a name and its ethnic group (Lauderdale and Kestenbaum, 2000), and using such scores in different ways alongside other contextual information (e.g. such as address of residence, which can be linked to census information on the distribution of ethnic groups in an area).

3.4.3. Advantages of the methodology

According to the authors of the studies analysed here, name-based ethnicity classification methods present a valid alternative technique for ascribing individuals to ethnic groups through their name origins, where self-identification is not available. The criterion for such validity is that the methodology makes it possible to subdivide populations to a sufficient degree of accuracy at the ethnic group aggregate level, and not necessarily at the individual level (i.e. it produces reasonably accurate total figures and orders of magnitude). In general, there is a consensus in the literature that although this methodology cannot entirely replace self-assigned ethnicity information, it provides a sufficient level of classification confidence to be used in the measurement of inequalities and in the design and delivery of services that meet the needs of ethnic minorities. In predicting these types of outcomes, name-based classifications have proved a very cost effective method compared with conventional collection of self-assigned ethnicity information (e.g. projects aiming to collect all patients' self-reported ethnicity in the UK have had an average response rate of 56%: Adebayo and Mitchell, 2005).

Some of the methods evaluated here also provide a measure of the degree of strength in the assignment of an ethnic group to each name (Lauderdale and Kestenbaum,

2000; Word and Perkins, 1996), and others offer the probable religion and language associated with each group of names (specifically those using *Nam Pehchan* or *SANGRA*). These efforts have produced three computerised name classification systems, *Nam Pehchan* (Cummins et al, 1999) and *SANGRA* (Nanchahal et al, 2001), designed to classify South Asian names in the UK, and GUESS (Generally Useful Ethnicity Search System) (Buechley, 1976) which identifies Hispanic names in the US. These computer systems have been used in a wide variety of studies in public health, having proven very useful in identifying areas of inequality and health needs within populations (Coronado et al, 2002; Honer, 2004).

Furthermore, name-based methods have been successfully applied to sample members of particular ethnic groups using Electoral Registers or telephone directories (see discarded studies listed in Section 3.3.1), presenting significant cost advantages over other alternatives (Cook et al, 1972). Moreover, this methodology has also proven useful in combination with conventional ethnicity classification information (Coronado et al, 2002). When some degree of ethnicity information is already available for a population, name-based classification can provide complementary information to detect errors, complete missing data, or correct bias introduced by proxies of ethnicity used, such as country of birth (e.g. second generation migrants).

Despite having found some inconsistencies between *Nam Pehchan* and *SANGRA*, when trying to classify the entire UK population (using the Electoral Register), Peach and Owen (2004) concluded that name-based methods are of potential value to health organisations, local authorities, commerce and academics, but further research to

improve the classifications is needed. A similar conclusion was reached by Bhopal et al (2004), who also used *Nam Pehchan* and *SANGRA* in an extensive study linking census and health data in Scotland, highlighting that name-based methods are valuable in the absence of alternative information sources, and more crucially, suggesting that they produce important information at low cost (Bhopal et al, 2004).

3.5. Alternative Approaches to Building Universal Name Classifications

The 13 research studies reviewed in the previous two sections have demonstrated the advantages of name-based methods as well as their principal current limitations. With respect to the latter, three general priorities for improvement arise, as justified in the previous section: (a) a need for a reference population with high spatio-temporal coverage including name frequency data sourced both in the host and origins countries, (b) the need to use name scores to measure the probability of a name being associated with a particular ethnic group, and (c) the need for a system that classifies the whole population into all of the potential ethnic groups, and not just one or a few. This section will review some alternative approaches in the literature that have attempted to build such 'universal' name classifications, making partial contributions to these three general needs.

These tasks are made much easier today by the use of population registers that cover most of the population, such as Electoral Registers or telephone directories, providing very valuable name frequency information, name spelling variants, linkages between surnames and forenames, precise addresses, etc. A few of the studies analysed in the previous review make use of some of these resources,

although they only cover parts of a country, or use manual methods such as counting names in a paper telephone directory. Electronic versions of such registers can today be accessed through special requests or purchased from data providers, making this type of analyses much simpler.

However, such directories or registers do not obviously contain any ethnicity information associated with people's names. Therefore, by using these registers population coverage is maximised, but knowledge about the origin of the names is minimal. Researchers in marketing, computer science, and linguistics have made independent attempts to impute the language or culture of origin to a name using different data mining techniques. The field within linguistics that studies proper names is called 'Onomastics', and includes personal names, place names, and unique new naming in general (objects, companies, brands, etc). Other fields that have tackled the problem of identifying the origin of personal names are computational linguistics, an interdisciplinary field dealing with the statistical and rule-based modelling of natural language from a computational perspective, and in marketing and geodemographics, where imputation of ethnic group membership may be used in order to target potential customers and neighbourhoods. These approaches will be reviewed here to try to illuminate alternative ways of assigning the linguistic or cultural origin of each name in large lists derived from population registers (i.e. >25,000 names), when no ethnicity or related surrogate data are present, and without having to code them individually.

3.5.1. Computational and marketing approaches

The task of building a name classification system covering a large number of ethnic groups, when comprehensive name reference populations with ethnicity information

is not available, has been tackled specially in the US since the 1980s (Abrahamse et al, 1994). All of these attempts have been based on particular applications for which they were developed, usually in the commercial sector or under commercial relationships with the public sector. Therefore, most of these approaches have not been properly documented or published, their methods are opaque and external validations if done are not made explicit. One exception is Abrahamse et al (1994), from Rand Corporation, who evaluate two name-to-ethnicity databases in order to identify Hispanics and Asians in the US, the latter built by Donnelley Marketing. They conclude that the best approach to developing a comprehensive Asian surname dictionary entails combining three stages: take a seed of 1,000 Asian names provided by the US Census Bureau; expand it, identifying the most common surnames in areas of high concentration of Asians by crossing names from the Electoral Roll with the Census information at small area level; and then subdividing them by country of origin using country of birth information from tax records.

There are at least four companies in the US that have commercially exploited such databases, but unfortunately their methods have not been published. Language Analysis Systems (LAS; Herndon, Virginia) developed an extensive knowledge base to manage names in large databases, involving de-duplication of names about the same individual, name translation and transcription, and name-matching techniques, and also assigning names to its language of origin using a proprietary ‘name classifier algorithm’. LAS did publish some of their name classification techniques (Williams and Patman, 2005), but after the company was sold to IBM in 2006, a lot of their public papers disappeared from their website (Dance, 2007). Ken Williams, the former owner of LAS and ex-president of the American Names Society, now

working for IBM, has filed a US patent protecting his *name classifier algorithm* (Williams, 2007). The importance of this business is such that IBM has created a 'Global Name Recognition' business unit (http://www-306.ibm.com/software/data/globalname/), which is very successful in security applications dealing with international lists of names in a post-September 11th world (The Economist, 2007b).

Other companies focus upon an applications area often termed 'multicultural marketing', and offer similar products to perform the ethnicity profiling of names, such as:

- Donnelle Marketing, now a branch of InfoUSA
  (http://www.donnelleymarketing.com/)
- List Service Direct Inc (LSDI)
  (http://www.listservicedirect.com/ethnic_religious.html)
- Ethnic Technologies
  (http://www.ethnictechnologies.com/index.html)

The applications of these name-based ethnicity profiling techniques not only cover the segmentation of customers or public service users, but also tasks such as survey sampling (Hage et al, 1990; Himmelfarb et al, 1983), drawing members of jury services and electoral redistricting (Abrahamse et al, 1994), and improving automatic document archival and speech recognition and synthesis systems (Bonaventura et al, 2003). However, the majority of these computational and marketing approaches generally ignore the detailed issues of ascertaining the geographic, cultural and linguistic origin of the name forms that are found in their databases, once names have

arrived in the countries of destination of migrants. The study of such linguistic processes lies in the field of Onomastics.

3.5.2. Onomastic studies: the cultural ethnic language group (CELG) technique

In Onomastics, the classical way to study the origin of a surname is to investigate the genealogies of people with that surname, using the earliest historical documents available that mention that surname and linking it to a place and period of time (Reaney, 1958). Through this method, a linguistic expert may be able to assign a language of origin and an etymological definition of a name (original meaning of the name or explanation of its origin). The main problems that onomastic researchers face in this task is identifying reliable genealogical sources and accommodating the regularities of language change so as to recognise true mutations in the way that a surname has been written and pronounced in one or several languages through history.

This clearly involves a very cumbersome and slow process, and it is estimated that a experienced name researcher would have a productivity of only four surnames a day (Hanks and Tucker, 2000). Adopting a rule of thumb that a surname dictionary should represent the names at least of 70% of the people in a population, this would require the explanation of several tens of thousands of names, a task that would be too time-consuming for any single researcher if it were attempted manually (Tucker, 2003). Furthermore, only a small percentage of most common names in the UK or the US have been studied genealogically, and most of the successful genealogies have dealt with rare and unusual surnames (Hanks and Tucker, 2000). These are the

reasons why there have been so few surname dictionaries published: forenames dictionaries, by contrast, are more numerous since forenames are relatively easy to investigate and fewer in number.

This has led to a proposal for a semi-automatic onomastic means of developing a 'surname to language' reference list by Tucker (2005). Hanks and Tucker (2000) pre-classified the 70,275 most common surnames in the US into 44 'Cultural, Ethnic and Linguistic' groups (CELG), to be further studied by each of the etymologists that wrote the descriptions of the entries in the Oxford Dictionary of American Family Names (DAFN) (Hanks, 2003). Tucker (2005) developed a technique termed *Cultural-Ethnic-Language Group (CELG)* in which a database of individuals with both forenames and surnames is required. To do this he used the US telephone directory with 88 million subscribers, from which he computed forename and surname frequencies and established relationships between the two.

This entailed a number of stages. First, a set of 'diagnostic forenames' (good predictors of ethnicity) was manually classified into cultural-ethnic-linguistic groups (CELG) by onomastic experts (Hanks and Tucker, 2000). This manual coding was achieved in a much more efficient way than with surnames, since there are less forenames than surnames, as will be explained later, and forenames are usually derived from a common set of religious or popular characters that are distinctively written in each language or culture and hence easier to be intuitively classified than surnames. In this way, Hanks and Tucker (2000) created a reference list of 85,000 unique forenames with a frequency greater than 9 in the US, which for simplicity

will be termed the 'F_list'. For each forename in F_list, an entry was created that included the following fields (Hanks and Tucker, 2000):

- Diagnostic Forename: (Yes/No) Indicating whether the forename is a good predictor of ethnicity or not
- Gender: (Female, Male, Both, Unknown)
- Cultural Ethnic Linguistic Group (CELG): One of 44 CELGs, assigned manually

The 85,000 forenames were not manually coded all at once, but in a series of steps starting with the most common forenames and using their surnames' CELG, as will be explained in the next paragraphs.

Second, the F_list was linked to the forenames of the individuals in the telephone directory. Third, a new surname reference list was produced, comprising all unique surnames in the telephone directory (1.75 million) and which will be termed 'S_list', which was also linked to the telephone directory through the surnames of individuals. The structure of the database at this point is represented as follows (the arrows indicate the relationships between the three tables in the database):

F_List		**Telephone Directory**	**S_List**
(n= 85,000)		*(n= 88 million)*	*(n= 1.75 million)*
Forename	→	Forename	
		Surname	← Surname

Fourth, using the three linked tables, the objective was to calculate for each surname in the S_list the percentage of people in the telephone directory with forenames assigned in the F_list to each particular CELG. However, in performing this calculation, Tucker (2005) introduced different weightings to two types of forenames:

(a) Forenames considered as 'diagnostic' in F_list are given double weight when computing the counts, i.e. counts are multiplied by a value of 2

(b) Female forenames are weighted down to 80% of their count values, i.e. counts are multiplied by a value of 0.8 (Tucker, 2007a). This is a 'rule of thumb' value to counteract the fact that women's forenames are less indicative of their surname's ethnicity because of intermarriage between ethnic groups and subsequent adoption of their husband surname.

These weights improve the efficiency of the classification, since diagnostic forenames are more representative of a CELG than non-diagnostic which tend to overlap between groups (e.g. Maria). Moreover, married women usually carry their husband surname (especially as listed in the Telephone Directory, e.g. *Maurizio & Tünde Moretti*) and could introduce a misinterpretation of the true CELG of the female's forename. One additional problem of anomalies in the CELG connectivity between forenames and surnames is child naming fashions, since a forename from a different CELG can be chosen by a family following a fashion (e.g. French girl names being popular amongst Anglo-Saxons). Unfortunately this problem cannot be avoided in the absence of other data, but it is deemed to be of a small relative importance (Tucker, 2005).

The weighting mechanism is here illustrated through a worked example using hypothetical figures. Say the surname *Moretti* had a total count in the US telephone directory of 645 people. The distribution of the forenames' CELG of these 645 people according to the F_List was as follows:

English; 311, Italian; 162, Spanish; 142, Others; 30

These people counts for each CELG were weighted according to the two criteria mentioned above; diagnostic forenames and female forenames. For example, the 162 people with forenames associated to the Italian CELG, were weighted as follows:

10 male diagnostic forenames; 10 x 2 = 20

12 female diagnostic forenames; 12 x 2 x 0.8 = 19.2

80 male non-diagnostic forenames; 80 x 1 = 80

60 female non-diagnostic forenames; 60 x 0.8 = 48

Total weighted count = 167.2

The total re-weighted count of 167.2 for the Italian CELG contrasts with the original 162 people, meaning this surname is slightly more prone to be associated with Italian forenames. The same exercise was repeated for all CELGs deriving the following re-weighted counts and relative sizes in brackets;

English; 192.7 (42.8%), Italian; 167.2 (37.2%), Spanish; 58.7 (8.6%), Other; 41.4 (11.4%); Total weighted count = 460 (100%)

The percentages above indicate relative weighted frequencies per CELG.

Finally, each surname was assigned to the CELG of highest relative weighted frequency other than 'English', this group being excluded since it is the 'default CELG' in the US, due to a 'host-country' assimilation effect. Moreover, only CELGs with a relative weighted frequency of at least 4% were considered (e.g. if the largest CELG other than English had a relative frequency of 3.7% the surname was left unclassified). This minimum threshold was introduced to make sure that there was a sufficient minimum number of weighted counts associated with the CELG finally selected. As a result, in the previous example the surname Moretti was finally

classified as Italian. This technique can be repeated iteratively to increase the number of diagnostic forenames classified and then the number of surnames and so forth.

At the end of the process, the 70,275 surnames included in DAFN were classified into 44 CELGs, 40,098 of them into the British/English/Welsh/Irish categories, and the remaining 30,177 of them into the rest of non-Anglo Saxon CELGs. The performance of the CELG technique is deemed to have an accuracy of between 88% and 94% (Tucker, 2005), based on a range of rates of misclassification identified by the DAFN language experts to which the surnames were sent for further study.

The uniqueness of Tucker's method is that it exploits the patterns of cross-occurrences between forenames and surnames that are more common amongst groups of the population that may be defined by their common ancestry. Hanks and Tucker (2000) contend that forename naming practices survive generations after immigration, even when the original language may have been lost by descendants. This method is very efficient because it leverages the differential skew of the name frequency distribution between forenames (extremely positively skewed) and surnames (largely positively skewed). To illustrate this with an example, 10% of the surnames in the US are sufficient to cover 91% of the population, while 1% of forenames is sufficient to cover 95% of the population. There are 1.25 million unique forenames in the US, so concentrating upon just 1% of them (12,500 forenames) allows one to code forename ethnicity of 95% of the US population, and hence the corresponding surname ethnicity (Tucker, 2001). Furthermore, by applying the CELG technique this population coverage can be increased to nearly 100%, while improving the overall accuracy of the names classified. This is further eased by the

use of etymology dictionaries of forename origins to code 'diagnostic forenames', with larger coverage and availability than surname dictionaries.

3.6. Conclusion

'The classificatory role of names proves very useful. By studying names we can find out how the human race divides up and then sorts into groups the many people living in a single society' (Smith-Bannister, 1997: 15)

Languages have evolved, branching out from a few original ones, following a parallel evolution with human genes across space and time. Surnames are derived from contemporary languages or others spoken in the last three to ten centuries. They have also been transmitted or extinguished through the generations following their populations' dynamics. Surnames have been proven to correlate well with several human genetic markers that are patrilinearly inherited, as well as with other cultural, religious and linguistic indicators that are vertically transmitted (such as forenames, dialects, accents and customs). The geographic distribution of surnames follows very distinct patterns that parallel zones of cultural interaction and marked linguistic boundaries. The very few exceptions to this trend can be explained by recent domestic and international migrations. Although the geography of forenames and their naming practices varies much more than that of surnames, it also reveals aspects of a group's ancestry that can complement the analysis of surnames. These factors indicate that personal name analysis can offer a reliable method to ascribe individuals to common human groups, where such groups are defined as having common linguistic, geographical and ethnic origins.

Within the wide research field of names origin analysis, this thesis focuses on its applications to classify contemporary populations according to recent migrations, meaning their own or those of their most recent three or four generations of ancestors. A large number of researchers have developed and applied name-based ethnicity classification techniques, in disciplines as diverse as human genetics, anthropology, public health/epidemiology, geography, history, demography, linguistics, computer science, economics and marketing. Research in this area has grown very rapidly over the past 30 years, following increasing interest in international migration and ethnicity, improvements in computer processing power, and the wider availability of digital name datasets that cover entire populations at the individual person level.

Only a few studies have focused upon measuring the accuracy of these different name-based ethnicity classification methods, by trying to identify their limitations and advantages. This chapter has presented a thorough review of 13 research studies representative of these efforts, which have demonstrated the advantages of these name-based methods as well as their current principal limitations. These studies share a number of common methodological processes and research components: first, a name reference list is independently built or sourced from another study or from 'an expert'; second, a separate target population is manually or automatically classified into ethnic groups; and third, the accuracy of the method is evaluated against a previously known 'gold standard' for ethnicity in the target population. The claimed prediction success of the different classifications are measured using the epidemiological concepts of sensitivity, specificity, positive predicted value and

negative predicted value that summarise the measures from a confusion matrix. The different classifications' sensitivity varies between 0.67 and 0.95, their specificity between 0.80 and 1, their positive predicted value between 0.70 and 0.96, and their negative predicted value between 0.96 and 1. These evaluation results prove the value of name-based ethnicity classifications for most applications.

This methodology makes it possible to subdivide populations to a sufficient degree of accuracy when ethnicity information is not available, especially at the aggregate ethnic group level, producing relatively accurate total figures and order of magnitude estimates. Moreover, name-based classifications have proved a very cost effective method compared with conventional collection of self-assigned ethnicity information, suggesting ways to complement or replace self-assignment depending on the type of application. Amongst its limitations, its classification accuracy and coverage needs to be improved for some groups and contexts. Three general needs for improvement arise from the review presented here: (a) a need for a reference population with greater spatio-temporal coverage, including name frequency data sourced both in the host and origin countries; (b) a need to use name scores to measure the probability of a name being associated with a particular ethnic group; and (c) a need for a system that classifies the whole population into all of the potential ethnic groups, and not just one or a few. These are the research gaps towards which this thesis contributes.

Alternative but rather obscure computational and marketing approaches have been attempted to build 'universal' name classifications, especially when little or no ethnicity and name information is available, making partial contributions to these

three general needs. Outstanding from the rest, the Cultural, Ethnic and Linguistic Group (CELG) technique, developed in the computer linguistics and onomastics field, has great potential for efficiently classifying hundreds of thousands of names into all of the potential ethnic groups present in a given population, with little ethnicity information. Furthermore, this makes it possible to create the desired 'surname scores', measuring the degree of association between a surname and an ethnic group by setting thresholds to the ethnicity distribution of its bearers forenames. The CELG technique is the main methodology used in this thesis, which has been further explored, enhanced and evaluated, as it will be described through the next chapters.

Finally, in order to create an improved ethnicity classification covering all of the potential ethnic groups present in a population, the name reference list has to be created using reference populations originating in a large number of countries, as is possible today through the use of electronic telephone directories, population registers and a growing realm of genealogical internet resources. Starting with these materials, two other ingredients are required: a taxonomy of cultural, ethnic and linguistic (CEL) groups into which to classify the names; and a set of techniques to perform the classification that puts each name in a slot of the taxonomy. These are no trivial tasks and form the first step in the development of the methodology presented in this thesis, which is discussed in detail in the next chapter.

Chapter 4. Taxonomy, Materials and Methods

By this stage it should be clear why the ontology of ethnicity supported by this PhD thesis is based upon personal names, as its title suggests. However, in order to create a universal ethnicity classification of names (i.e. one covering all potential ethnic groups as opposed to the partial ones which have dominated the literature to date), a formal taxonomy of ethnicity for this purpose must be conceived as a preliminary step. Name origins follow a very particular set of rules and patterns which usually follow linguistic and cultural criteria. Therefore, a classification of human groups using name origins will necessarily have to be based upon those linguistic criteria. The issues about creating a taxonomy of human groups for the purpose of classifying names will be further discussed in this chapter, as a precursor to proposal of a cultural, ethnic and linguistic taxonomy of personal names that will form the basis for a posterior classification of people into groups of common ancestry.

In the construction of the desired universal ethnicity classification of names, not only are large lists of surnames and forenames required as the input materials, but also data about their frequency and geographical distribution, whenever possible sourced from different geographical areas and periods of time – for the reasons set out in Chapter 3. Such materials can only be obtained from population registers with universal coverage, such as Electoral Registers or telephone directories.

Finally, once an extensive list of names has been sourced, and a universal taxonomy of cultural, ethnic and linguistic groups has been created, innovative methods are required that make it possible to assign the names to one of the categories in the

taxonomy, that is, a classificatory methodology. Several approaches have been proposed by different researchers from multiple disciplines in the literature, as reviewed in Chapter 3. This chapter will summarise these and other key classificatory and clustering methods and explain how they have been adapted to the classification exercise tackled in this thesis.

Therefore, this chapter describes the three integrated components of the first phase in the development of a name-based ethnicity classification; taxonomy, materials and methods. The first section deals with the justification and explanation of the concepts used to formalise a new taxonomic classification of names based upon cultural, ethnic and linguistic (CEL) groups. The second section discusses the potential universal name register data sources and presents the materials finally selected for use in this research. The third section describes the methodologies that are subsequently used to put these two components together; the universal list of names classified into the CEL taxonomy.

4.1. A Taxonomy of Cultural, Ethnic and Linguistic Groups (CEL)

4.1.1. Approaches to building taxonomies of human groups

As set out in the introduction to this chapter, before creating an ethnic classification of names it is first necessary to build a taxonomy of ethnicity that is tailored specifically to the characteristics of human groups as reflected in the contemporary names of multicultural populations. This is no trivial task, and is closely tied to the ontology of ethnicity supported by this thesis. This subsection will review a range of approaches to create such a taxonomy for the UK society at the beginning of the

second millennium. Rather than reviewing approaches to the creation of taxonomies of ethnicity in general, only those considered to be closely related to names will be mentioned, and as such most of them adopt a linguistic perspective.

The initial premise of this research is that current official ethnicity classifications are of only limited use for the purpose of classifying names into all of the potential ethnic groups present in a society. As explained in Chapter 2, they combine aspects of race, skin colour, geography and nationality into a single classification. Furthermore, they usually only cover the largest 10 to 15 ethnic groups present in a society, which sometimes are reduced to just 8 workable categories (if we discard the 'other' and 'mixed' groups, that are not very useful for most research purposes).

In order to gain a better idea of the kind of ethnicity classification that is actually useful to researchers in Britain, as well as to members of the population that is classified, two contemporary lists of meaningful ethnic groups will be consulted. One is the ethnicity classification that is currently part of the England and Wales Pupil Level Annual School Census (PLASC), which contains 95 ethnic groups of pupils and is listed in Appendix 2. Another is the UK 2001 Census 'write-in' answers to the ethnicity question, used by people who selected an 'other' category, allowing the Office of National Statistics to compile a detailed list of people's perceived ethnic identities that extends beyond the 16 standard groups. This list has been published for London and is reproduced in Appendix 2: it comprises 77 loosely defined ethnic groups many of them with several thousands people in London. These groups emanated directly from people's responses, and although have been standardised by ONS they present a high degree of overlap and fluidity (e.g. 'Moroccan', 'Arab' and

'North African'). These two lists, PLASC and 'Census write-in', which primarily follow a list a of languages spoken in British schools and in London neighbourhoods, provide a useful picture of the type of ethnicity classification that is actually required in order to carry out research on ethnicity at a much finer level than current official classifications permit. But how can it be ensured that such a list is exhaustive, non overlapping, hierarchically organised, and connected to name origin groupings?

Anthropologists have built classifications of human groups and their relationships based on cultural customs, languages and also fossil and archaeological records (Eriksen, 2002: xviii). Almost separately, linguistics have established a genealogical tree of the language families of the world, based on similar observable characteristics of languages, such as the phonetic, morphologic, semantic, or syntactic common origins and their evolution through history (Ruhlen, 1994). More recently, human geneticists have also attempted such classification of ancient human groups, usually borrowing anthropological and linguistic taxonomies to corroborate them with the genetic record (Cavalli-Sforza, 1997). This taxonomy work has spanned almost two centuries, and sometimes has been surrounded by a great deal of debate and speculation between different schools.

Surnames derive from the languages in which they were created, generally between two to ten centuries ago, and that have been passed down to us or modified in written form following specific morphologies and alphabetical rules through those same languages or those of the places to which those holding particular surnames have migrated. Forenames follow similar linguistic rules but are voluntarily picked up and almost freely modified by parents following a set of cultural, religious, linguistic,

social interaction and identity conventions, being propagated through a specific temporal and social medium. Therefore, language represents the primary factor in the processes of creation, modification, transmission and migration of surnames and forenames. Other secondary factors are religious, cultural and geographic aspects that together with linguistic considerations constrain the choice of forenames or the choice of marital partners and thus influence the ways in which both surnames and forenames are transmitted through generations within specific human groups. Therefore, a classification of names into human groups according to four criteria (linguistic, religious, geographic, and cultural factors) would necessarily primarily follow a classification of languages, being locally modified by the other factors. Surprisingly enough, this is the same proposal to divide human groups made by Charles Darwin (1859) in the *Origin of Species* quoted at the beginning of Chapter 3.

There are several classifications of the world's languages, which aim to list the languages currently spoken and organise them into linguistic families. The language family classification most widely accepted (Cavalli-Sforza, 2001) is that of Greenberg and Ruhlen, that attempts to relate all existing languages to a set of approximately 20 families, each grouping a larger number of languages related by descent from a common proto-language (Ruhlen, 1987). The list and coding system most commonly used is the *Ethnologue* system in combination with the international standard for language codes ISO 639-3. The ISO standard provides an extensive enumeration of languages, including living and extinct, ancient and constructed, major and minor, written and unwritten languages (International Organisation for Standardisation, 2007). ISO 639-3 is the third version of the international coding of languages and was released in February 2007, containing 7,618 languages.

Ethnologue 15th edition, released in 2005 (Gordon, 2005) contains 7,299 languages, most of them considered alive, providing a taxonomy of languages giving the ISO 639-3 code, the number of speakers, locations, dialects, and linguistic affiliation which relates all of them to a multilevel hierarchy of subfamilies that connect to 108 language families at the top (see www.ethnologue.com for the complete list and hierarchy). However, most of these 108 language families are considered language isolates, and most of the languages are assigned to the core of Ruhlen's 20 families.

The tandem *Ethnologue* - ISO 639-3 language classification forms the basis for the taxonomy of ethnicity based upon personal names developed in this thesis. As such, this taxonomy initially distinguishes ethno-linguistic groups through the names currently present in the UK, and is later modified by cultural, religious and geographic criteria where required to reflect the uniqueness of the group's names in the UK.

4.1.2. The CEL taxonomy

Following Hanks and Tucker's (2000) onomastic method developed for the Dictionary of American Family Names (DAFN) (Hanks, 2003), the taxonomy of names developed in this thesis is called the 'Cultural, Ethnic and Linguistic' classification, abbreviated by the acronym 'CEL'. It is based upon the *Ethnologue* - ISO 639-3 language classification for the languages found today in the UK and modified by cultural, religious and geographic classifications that were considered appropriate.

In this thesis 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 a

frequency of three or more occurrences per surname or forename. 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). The assumption underlying this thesis is that, the four dimensions that define a CEL, religion, geography, ethnicity and language, have left a 'trail' which can be today discerned from the characteristics of the forenames or surnames that belong to each CEL. These characteristics can be a name's morphology (elements, letter patterning, 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 also the 'raw materials' used by researchers in the field of onomastics.

The criterion used to create the CEL taxonomy, both in DAFN and in this thesis, 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 4.1 for the complete list). The resulting CEL taxonomy is thus comprised of a series of homogenous categories of various resolutions (in terms of size and scope) that primarily follow an onomastic criterion to classify names according to their common origins. The individual CELs form the building blocks of a multidimensional system, in which they can be aggregated into higher level groups not only following onomastic criteria, as applied here, but also using alternative combinations according to religious, geographic, ethnic or linguistic criteria. These different aggregations of CELs can then be applied to classify a population according to the criterion that best fits the purpose of each application (see Appendix 3 for the correspondence between CELs and the different aggregations proposed).

The process by which the CEL taxonomy was created is therefore a heuristic one, and has been developed in parallel with the overall classification of names, since the original very coarse groupings of languages, religions or continents (e.g. Hispanic, Muslim, or African categories) have been subdivided into finer categories during the

process by which the classification rules explained in Section 4.3 and Chapter 5 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 Groups’. A list of these CEL Types, ordered by CEL Group, is presented in Table 4.1, while the full details by CEL Type are described in Appendix 3.

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 4.1: The CEL Type taxonomy and its assignments into CEL Groups
See Appendix 3 for a full lookup table between CEL Types and their various groupings

4.2. Data sources

'I imagine Alexander Graham Bell rejoicing if he could see my private library of those typically very bulky books, usually printed with fonts verging on the microscopic, that have been defined as works "with a lot of characters and little action". I am referring to telephone directories. (…) To the surname scholar these books are of fundamental importance.'

(Tibón, 2001: xviii, translated from original in Spanish)

4.2.1. Some discussion of potential data sources

The literature review presented in Chapter 3 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*. By contrast, 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 be 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'. For reasons of clarity and ease of reading flow, the data sources discussed here will only be those for reference populations, which will be then used to create a name-to-ethnicity reference list. Data sources used for the evaluation of the classification, that is, the target population, will be discussed in Chapter 7 when describing the evaluation of the classifications.

In the 13 studies described in Chapter 3, 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) and are listed in Table 3.3 ('Reference List Production Method'). In building their name reference lists, these researchers had access to the names and ethnicity of people, always understood as a self-assigned identity, or other proxies for ethnicity, such as country of birth (of the person and / or parents), the parents' ethnicity, nationality, or judgement by a third person or name expert. However, as it was concluded in Chapter 3, all but two studies used very small reference populations to build their reference lists and to make universal inferences. The two studies which did have large enough reference populations as to satisfy the 'Cook et al (1972) first criterion' (of at least 13.4 people per surname) were Lauderdale and Kestenbaum (2000), which included 1,609,679 individuals from ethnic minorities, and Word and Perkins (1996), with 597,533, derived from patient registers and Census responses. The remaining studies used reference populations that were below 75% of the required minimum size of 13.4 people per surname. Even so these two studies only classified their populations into, respectively, six Asian and one Hispanic group.

The aim of this thesis is to develop an ethnicity classification of the names of the whole population of the UK into a taxonomy that reflects the systematic variability in inherited names and naming conventions. Consequently, if it were to follow similar methods to those developed in the literature reviewed in Chapter 3, it would require a dataset covering the whole population and collecting ethnicity at the individual level. There is only one dataset available with these characteristics; the decennial Census of

Population. However, for reasons of privacy protection, data on individuals are not available until 100 years after the Census is carried out and the first time this question was asked in the UK was in 1991, so the first studies that will be able to do this will come out in 2091. There is a precedent to the use of names information from census sources, in that the US Census Bureau has used individual answers to the ethnicity question to create their own Hispanic Names file (Word and Perkins, 1996). However, initial attempts by this author to contact relevant people at the UK Office for National Statistics who had access to names in the Census were not encouraging. A very rich, but highly restricted dataset was indeed compiled by Bhopal et al (2004) in Scotland, linking individual people in the 2001 Census individual returns, hospital admissions, and patient, birth and mortality registers, for a longitudinal study on health. This linkage was done using the individual's full name and address, and the South Asian name analysis techniques *Nam Pechan* and SANGRA were also applied to classify the names into these groups. Were a list of names and ethnicity or proxies to ever be produced from such dataset, it would definitely afford the best resource available for the purposes described in this research. However, early enough in this research it became obvious that Census data on names would not become available, at least within the life of the project.

One alternative could be to go back over 100 years in the Census and try to use country of birth information to compile a list of names and most common countries of birth. However, this route would not have been very useful since the foreign names in British 19th century population bear little resemblance to today's multicultural population. An analysis comparing the names of the 1881 Census and 1998 Electoral register used in this research has been published by Tucker (2004a;

2004b), indicating that the main difference between the two registers are foreign names that have migrated through the twentieth century.

Several attempts were also made to collect lists of names with ethnicity information or proxies such as nationality or country of birth from large population registers, such as from National Insurance registrations (Department for Work and Pensions), Workers Registration Scheme (Home Office) and the General Practice Patient Register (Department of Health), but these were not fully successful. A full list of potential datasets with data about migrants has been recently compiled by Rees and Boden (2006), some of which must hold names data and hence could produce valuable name dictionaries.

However, through the Knowledge Transfer Partnership between UCL and Camden Primary Care Trust (PCT) that co-funded this PhD research, access to some of these lists – Patient Register, Birth and Mortality Registers, and Hospital Admissions – were made available, subject to confidentiality safeguards, for the London Boroughs of Camden and Islington (Camden PCT and Islington PCT). These Boroughs have a combined population of approximately 400,000 people. However, it was decided to use these rich datasets in the evaluation and applications sections of this research, and not to build the name classification itself, in order that it could subsequently be independently validated. There is only one small exception with this in the patient register, as will be explained in Section 4.3.5. Other evaluations of the classification were conducted using similar datasets of the London Borough of Southwark (Southwark PCT), and the Electoral Register with nationality for the London Borough of Hammersmith and Fulham (see Chapter 8). Even though, the analysis of

names for these was performed by internal researchers and the results anonymised prior to release to the author.

Moreover, as regards the reference population, the research 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 or a proxy for it. Such universal and publicly accessible population registers, typically Electoral Registers or telephone directories, do not contain any explicit ethnicity information associated with people's names. Despite this, it was decided to go ahead with this approach and to explore other techniques to classify names using the computational and onomastic approaches as described in Section 3.5.1 of Chapter 3. A range of different data mining techniques was used for their classification, as explained in Section 4.3. Based on these techniques and the literature reviewed in Chapter 3, two components of additional information were seen as key to enhance the knowledge about the origin of names; the relationship between the forename and surname in a person, and the geographic location of the name. Therefore, such registers had to be accessed with the person's full forename and surname information, as well as full address or its fine aggregations, such as unit postcode or postal area.

Therefore, and marking a departure point from most of the literature in this area (except for Tucker's (2005) method), in this PhD thesis a reference population with total population coverage of individual names but without any 'true ethnicity' information was used. The names in this reference population were classified following a 'computational-onomastic' approach, in other words, names were

classified according to their intrinsic characteristics (forename-surname clustering, frequency, morphology, and geographic distribution) rather than the ethnicity reported by their bearers.

4.2.2. Description of data sources used

The data sources finally selected to build the name reference lists in this thesis were comprised of name frequency datasets with high population coverage and at various temporal and spatial resolutions for different countries (derived from the Electoral Register or telephone directories). These data sources are listed in Table 4.2, which also includes other characteristics such as the number of names included, and their temporal and geographic coverage. These datasets were obtained under a variety of use conditions from the data providers which restricted the level of disaggregation, as described in Table 4.2 ('resolution' columns), or the availability of location information.

Country/ Territory	Name of Dataset	Year	Nominal Resolution (finest record)	Spatial Resolution (smallest area)	Data Provider	Population included in Dataset	Total Population Enumerated	% of Country's Total Pop.	Country's Total Population (*)	No. Unique Surnames	Avg. People/ Surname
Great Britain	Electoral Register & Consumer Dynamics	2004	Individual person (Forename and Surname)	Postcode Unit	Experian UK	Residents registered to vote of age >17 (opt-in) + consumer database	46,336,087	77.5%	59,800,000	218,392	212
Great Britain	Electoral Register	1998	Surname	Postal Area	Experian UK	Surnames >100 occurrences (age >17)	37,278,477	63%	59,200,000	25,730	1,449
Great Britain	Census of population	1881	Surname	Postal Area (equivalent)	ESRC UK Data Archive	All census respondants	28,225,211	81%	35,026,108	44,545	634
Northern Ireland	Electoral Register	2003	Surname	Postal Area	Experian UK	Residents registered to vote of age >17	n/a	n/a	n/a	n/a	n/a
Ireland (Republic of)	Electoral Register	2003	Surname	County	Experian UK	Residents registered to vote of age >17	2,912,541	73%	4,015,676	n/a	n/a
Australia	Electoral Register	2002	Surname	Standard Statistical Division (SSD)	Pacific Micromarketing	Residents registered to vote	7,784,676	38%	20,264,082	12,266	635
New Zealand	Telephone directory	2002	Surname	Province	Pacific Micromarketing	Telephone subscribers	934,686	23%	4,076,140	n/a	n/a
United States	Telephone directory	1997	Surname	State	Ken Tucker	Names with >100 occurrences in the tel.directory	81,000,000	30%	266,490,000	145,242	558
Canada	Telephone directory	1996	Surname	National	Ken Tucker	Names with >100 occurrences in the tel.directory	9,150,000	28%	33,098,932	33,355	274
Spain	Telephone directory	2004	Individual person (Forename and Surname)	Full Address	Infobel	Telephone subscribers that have not opt-out	11,800,000	27%	43,200,000	292,512	148

Table 4.2: Sources of reference population data used to build the CEL classification
(*) Country's total population as per the official count for the year on which the dataset was drawn (year column)

The major source of data amongst those listed in Table 4.2 has been the Electoral Register for Great Britain, both in its 1998 and 2004 editions. The purpose of these registers is to record the names and addresses of British and foreign citizens entitled to vote in local or national elections in Great Britain (British, EU and Commonwealth citizens aged 18 or over, plus those that will attain age 18 during the year of the register's currency). Since 2002, UK residents have had the right to remove their records from of the public version of the Electoral Register, an option known as 'opt-out' (Electoral Commission, 2002). In the last few years the proportion of citizens who opt-out of the public version of the Electoral Register has risen as follows; 31.20% in 2004, 32.14% in 2005, and 36.71% in 2006, according to Equifax (2007). To compensate for 'opt-outs', private sector resellers of the Electoral Register, such as Experian, CACI or 192.com, supplement the public version of the Register, known as the 'edited version', with other data sources, such as public registers (company directors and shareholders registers) as well as commercial surveys or third party customer data, in order to compile population databases.

In the case of Experian (Nottingham, UK), this is now commercialised as a 'Consumer Dynamics' file that in 2004 contained 46,336,087 adults, a higher number than those in the unedited version of the Electoral Register (Sparks, 2005). Two versions of this dataset for the UK were kindly made available by Experian to University College London. One from 1998, which represents the full Electoral Register of that year, but that was made available to UCL in a form that only included surnames held by 100 people or more and their frequencies by postal area. A second version for 2004 included all surnames and forenames at unit postcode level. Full details of these datasets are given in Table 4.2.

A different type of dataset used for historical analysis was the distribution of surnames in the 19th century in Great Britain (i.e. excluding Northern Ireland), derived from the individual responses to the 1881 Census. This file was kindly supplied by Kevin Schürer, Director of the ESRC UK Data Archive at the University of Essex, and contained counts of surnames by Parish in the 1881 Census (Schürer, 2004). This file was aggregated to today's Postal Areas in a previous project at University College London (Surname Profiler, 2006). In that project, this dataset made it possible to trace internal migration movements in the changing geographic pattern of names over time, while in the research described in this PhD it has been used to screen out names that have arrived in the UK during the late 19th and 20th centuries.

An additional dataset used, that is not considered a 'name dataset' and therefore is not included in Table 4.2, is a Geodemographics neighbourhood classification system, Mosaic, provided by Experian, which classifies the UK's 1.6 million unit postcodes into 61 types according to the demographics of the immediate residential neighbourhood as at 2001. The neighbourhood types were clustered by Experian using both UK Census 2001 small area statistics as well as other publicly available and commercial datasets (Harris et al, 2005). In this research, the Mosaic dataset has made it possible to match the areas of highest concentration of certain names and relate them to neighbourhood types with higher presence of particular ethnic groups, religions, socioeconomic types, or urban/rural populations, as will be described in Section 4.3.3.

Besides the UK data, other less detailed population files for other countries or periods have been sourced from Electoral Registers or telephone directories from six other countries (Ireland, Australia, New Zealand, the US, Canada and Spain) containing surname and sometimes forename data at different levels of spatial disaggregation. The full details and characteristics of these datasets are also listed in Table 4.2. Other minor datasets were also used at different stages of the classification process, and will be mentioned in the context of the purpose for which they are applied.

The sourcing, compilation, cleansing and standardisation of the datasets listed in Table 4.2. consumed a considerable part of the research process and large amounts of computer time and power. These datasets were loaded in an Oracle database in order to process and link them in an efficient manner. The database contained over 100 million records pertaining to 225 million people and approximately 500,000 surnames and 200,000 forenames. Different types of aggregations and calculations were performed on the original registers using Structured Query Language (SQL) statements, creating multiple tables that were organised in a relational database management system, linking individual people with forename and surname tables, postcode and geographical units tables. The Oracle environment permitted efficient manipulation of the datasets and calculation of summary measures, not only through individual SQL statements but also through programming code in Oracle's proprietary language PL/SQL that makes it possible to link SQL statements with programme flow controls and to interact with the database in a dynamic way. These aspects of technical manipulation of the database will only be made explicit through the rest of thesis wherever is deemed relevant for the particular point being

explained, while elsewhere they will remain an implicit aspect of the overall techniques used. Examples of some of these SQL queries are offered in the Appendix 4.

After this phase of data sources evaluation and compilation process, a series of datasets were compiled and organised in a relational database management system for further processing. As described, these included universal name and addresses publicly available population registers, sourced from telephone directories and Electoral Registers at various spatial granularities. The next section provides a general description of the techniques to classify the names in this database system, while the explanation of how these datasets were mined in different ways to classify their names into cultural, ethnic and linguistic groups is covered in the following two chapters (5 and 6).

4.3. Name Classification Techniques

This section will set down the basic methodological framework that was developed during this research for the classification of names into the CEL Taxonomy. The different techniques and approaches identified from the literature have been summarised, grouped and renamed as a set of seven techniques. They are presented here with the benefit of hindsight after having been evaluated in the practice of classifying names. However, each of these seven techniques described here summarises the heuristics that were developed in the exploratory phase of the research, covered in Chapter 5, which guided the ('confirmatory') stage reported in Chapter 6. The automated approach described in Chapter 6 focuses on the extensive

use of one of these techniques, forename-surname clustering, as is explained in that chapter.

The task of classifying the 281,422 surnames and 114,169 forenames most commonly present in Britain in 2004 into cultural, ethnic and linguistic groups (CEL) is one that cannot be approached manually or following traditional etymological methods. The Dictionary of American Family Names (DAFN) includes the 70,000 most common surnames in the US and their etymological explanation, and comprises three bulky volumes of over 2,000 pages which took ten years and more than twelve experts to prepare (Hanks, 2003) even using as it did a semi-automated initial classification system to allocate groups of names to linguistic experts according to diagnostic forenames (Tucker, 2003). On the other hand, building name reference lists based upon pre-existing name-to-ethnicity information, applying the traditional approach used in the public health literature and reviewed in Chapter 3, is not possible since there are no such datasets available which cover the whole population, as pointed out in the previous data sources section (4.2). Therefore, given the number of names to be classified and the scarce resources available, a different type of approach was required for the current UK project.

Three main alternative approaches to classify names in ethnic groups were identified from the literature, and have already been covered in Chapter 3. The first one is the 'Forename-Surname Clustering' technique, which classifies surnames according to cross-occurrences of diagnostic forenames, and vice-versa, an early version of which was the CELG technique developed for DAFN and described in detail in section 3.5.2 (Tucker, 2005). The second one is a technique sometimes called 'geocoding'

(Fiscella and Fremont, 2006), ‘neighbourhood context’ (Abrahamse et al, 1994), or ‘geographical distribution’ (Passel and Word, 1980; Word et al, 1978), that basically consists of relating pre-known geographical concentrations of certain ethnic groups with concentrations of names in the same area, relative to national averages. This broad technique will be provisionally termed ‘geographical analysis’ here and will be further discussed later in this section under two different aspects; ‘spatio-temporal analysis’ and ‘geodemographic analysis’. A third recurrent technique found in the literature is termed here ‘text mining’, and relates to all the rules and text manipulation procedures that relate common name stems, endings, syllable patterns, character sequence and character presence or absence, to a particular cultural, ethnic or linguistic group (Bonaventura et al, 2003; Patman and Thompson, 2003; Perkins, 1993) or sometimes to a gender (Barry and Harper, 2000).

These three broad techniques together relate forename-surname clustering, geographical analysis of names, and name morphology, and as such constitute the core data mining methodology that will be explored in this thesis. The geographical analysis technique is divided here into two aspects: spatio-temporal and geodemographic analysis, bringing the number of techniques to be described to four. A fifth technique that will also briefly be described here is the use of lists of names with pre-known ethnicity to build name reference lists, the core method used in the literature reported in Chapter 3. Little use of this fifth technique was made because of the lack of data with universal coverage explained in Section 4.2.1. Finally, two other methods with a smaller relative effect than the rest are manual methods to research individual names, and international comparison of name frequencies, the latter practically not mentioned in the literature. This brings to seven the tally of name

classification techniques used in this research, and each will be described in the following subsections. All of the name classification methods found in the research literature relate either directly or indirectly to at least one of these seven techniques.

A summary of these seven major techniques used will be offered here, with the objective of giving meaning to the concepts employed in the next chapters and justifying their final selection after a preliminary evaluation.

4.3.1. Forename-Surname Clustering (FSC)

Forename-surname clustering (FSC) is the name given here to a technique that the present research has developed as an enhancement of the methodology originally termed the 'CELG technique' and introduced by Tucker (2003; 2005). FSC consists of identifying clusters of names grouped by high frequencies of cross-occurrences between forename and surname in individuals (e.g. surnames will be considered Chinese if a high proportion of their bearers also have Chinese forenames and vice versa). This technique is based on the underlying assumption that most cultural, ethnic and linguistic (CEL) groups; (a) adopt distinct forename naming practices that are passed from one generation to the next, even after the original language may have been lost in a family descending from migrants (Tucker, 2005), and (b) tend to marry members of the same CEL group (Lasker, 1985), with exogamy becoming prevalent only several generations after migration. These two processes, together with the geographical and social factors that underpin the relationship between marriages or procreation and naming practices, result in unique forename-surname clustering patterns and preserve the discontinuities between clusters.

The basic idea behind this technique is the same mechanism used in the DAFN to filter 70,000 surnames into 44 onomastic groups for etymology specialists to analyse, which was described in Section 3.5.2. The only difference being that in the DAFN the clustering was induced with a list of diagnostic forenames in a total list of 85,000 forenames pre-classified by CEL, while forename-surname clustering (FSC) can also be done automatically.

FSC can therefore be either user-induced or automatic. In the first case, the user selects a 'forename *A'* at random where the CEL is previously known, such as 'Pablo' (Spanish CEL), which will act as a 'seed' for building a new Spanish CEL. The user then finds the most common 'surnames *B'* that Pablos bear, such as Mateos, Garcia, Perez, etc, and then all the 'forenames *C'* associated with those 'surnames *B'* (e.g. Juan, Rosa, Javier, Marta, etc.). By repeating this process from the start for the 'forenames *C'* and conducting further iterations, the same forenames and surnames tend to be highly clustered around those individuals belonging to a same Spanish CEL category. This iterative process is illustrated in Figure 4.1, where only two of these cycles are shown (A-B and C-D).

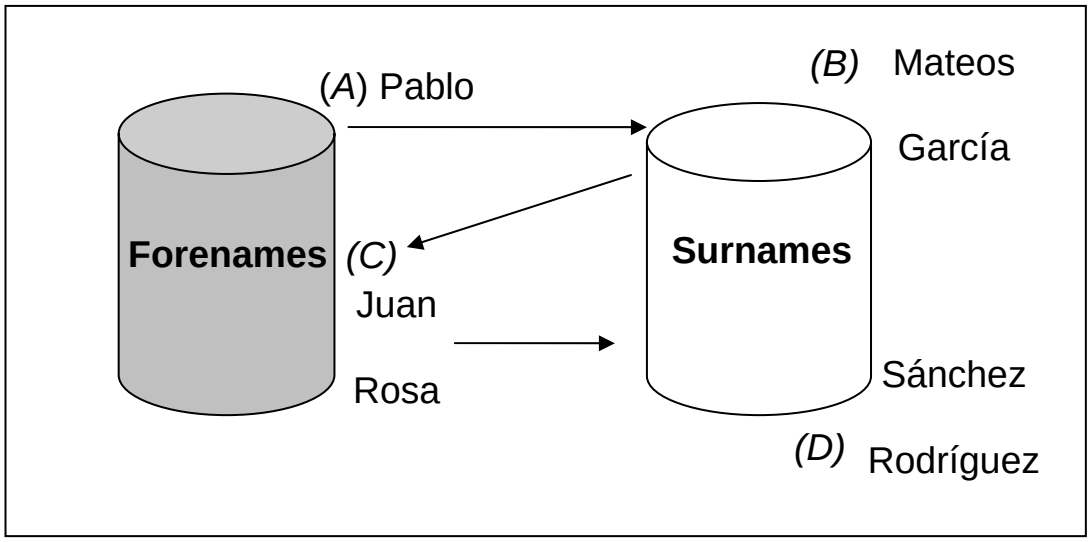

Figure 4.1: The Forename-Surname Clustering (FSC) technique, applied to a cluster of Spanish names

Therefore, after a few cycles one can find a set of several hundreds or even thousands of names belonging to a common CEL cluster, just by knowing a few forenames diagnostic of one CEL. In the automatic version of this technique, it is not even necessary to know the CEL of 'forename *A*'; the computer chooses a forename at random and automatically identifies clusters of common cross-occurrences through the same cycles described above. At the end of the automated process the user decides the most likely CEL of the whole cluster by looking up one or two names in a dictionary or through one of the other techniques described in this section. Another option is to attempt to cluster a large list of surnames all at once, measuring the 'forename distance' between them in terms of relative frequency of forenames found by common bearers in a pair of surnames. This option will be further analysed in Chapter 6, when explaining the automatic approach to classify names.

Returning to the DAFN example, manual classification of an initial set of approximately 3,000 forenames into CELs, and the application of FSC technique, allowed the authors to grow this list to 85,000 forenames preliminary coded by CEL

(Tucker, 2005). They used this list to automatically assign a CEL to over 100,000 surnames using this technique, proving its effectiveness.

This technique is perhaps the most useful method for the classification of large number of names into CELs, and it has indeed proved very reliable for classifying high frequency names in the heuristic approach presented in Chapter 5, and all of the names in the automatic approach reported in Chapter 6. It works best with CEL groups that are distinctive, such as Japanese names. Amongst its limitations are that it requires full forename and surname information, which is not always possible (many telephone directories just list initials instead of forenames). Furthermore, it has proved less appropriate for names that correspond to well-established immigrant groups that are very integrated with the 'host population' (e.g. Jewish or Huguenot names in Britain), since their overlap with 'host CEL' is high, and also for names with small frequencies since the pool of names for cross occurrences is too small to make inferences. However, as it will be explained in Chapter 6, further elaboration of the different clustering methods for this technique can yield much more productive results than all the other classification techniques taken together.

4.3.2. Spatio-temporal analysis

This technique is based on the analysis of spatio-temporal differences in the geographical distribution of name frequencies and rates across the locations and time periods available in the datasets. This implies the identification of significant differences in the total or relative frequencies of names between different areas within a country and between different countries or points in time. Once such significant differences are identified, there is a requirement for expert knowledge or

additional data on the geography and history of the countries and regions, and their internal and international migration and patterns. Through such human judgement or data analysis the CEL categories associated with such migration or differential distributions need to be explained. There are hundreds of different types of such spatio-temporal patterns and just a few different illustrative examples will be mentioned here.

For example, in the UK those names which are proportionally more common in postal area 'NW' (North West London) than in any other of Britain's 120 postal areas include large numbers of Jewish names, whilst most Greek or Greek Cypriot names have postal area 'N' (North London) as their most common. Figure 4.2 shows a map of Greek or Greek Cypriot names in London using the 2004 Electoral Roll classified by ethnicity during the first phase of the heuristic approach described in Chapter 5 and in Mateos et al (2007). It shows the area of concentration of Greek names in North London. Likewise, if a non-English name is more common pro rata in Wisconsin than in any other US state, this will support the contention that it is of German or Scandinavian origin, while other names that might appear to be Germanic yet are more common in New York than in any other US state are more likely to be Jewish than German. International comparisons are very useful, for example, most Chinese names are relatively more common as a proportion of the total population in the US than in the UK, whereas most South Asian names are proportionally less common in the US than in the UK. Moreover, names that are more common in Australia than in the US or the UK, and within Australia are even more common in New South Wales, are likely to be Vietnamese rather than Chinese.

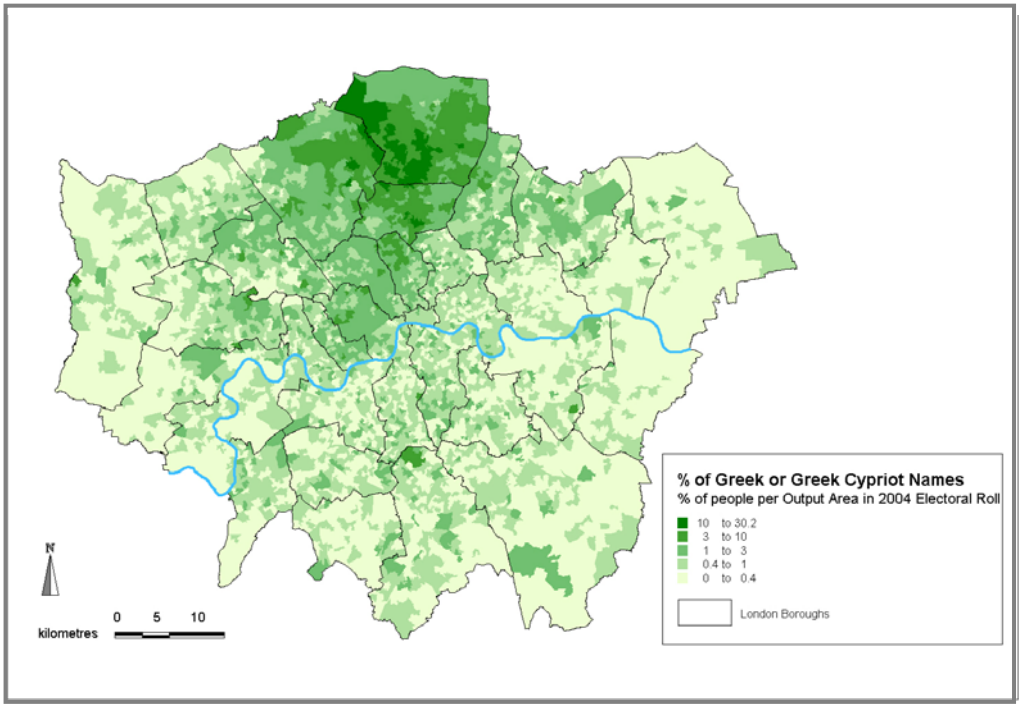

Figure 4.2: The distribution of Greek and Greek Cypriot names in London by Output Area (2004)
The map shows percentages of people in each Census Output Area classified into 5 intervals, using the 2004 Electoral register. Source: Mateos et al (2007)

With regard to the temporal dimension of this type of analysis, names present today in Britain that did not appear in 1881 are likely to be of foreign origin (or of more recent invention). Conversely, surnames of foreign origin that were present in Britain in 1881 are likely to have high numbers of British forenames today and therefore are unlikely to be clearly identified by the FSC technique.

The spatio-temporal analysis technique has been especially useful in the heuristic approach of this research (Chapter 5) to identify CEL groupings within regions or constituent countries of the UK, such as Scotland, Wales, Northern Ireland, Cornwall, or the Channel Islands. It has also been very useful to identify the most frequent names in the ethnic minority groups that are highly concentrated in a few

areas, such as South Asians in the UK. Its major limitation is that it requires detailed specialist knowledge or data of historic and current migrant settlement patterns by small area. Furthermore, it is applicable only to names with high frequencies, i.e. 100 or above because a sufficient number of them are required to be able to identify regional clusters (e.g. if a name has a national frequency of 10, of which 9 instances are found in one single area, it could be just because of a couple of related families from an ethnic group not belonging to the area's major ethnic groups).

4.3.3. Geodemographic analysis

Geodemographics is defined as *'the study of population types and their dynamics as they vary by geographical area'* (Birkin and Clarke, 1998: 88). The geodemographic analysis used in this research entails identifying the socioeconomic types of neighbourhoods where a name is most commonly concentrated, and making inferences about the population groups living in them. The analysis of the UK 2004 Electoral register data using the geodemographic neighbourhood classification *Mosaic* proved useful in identifying non-British names which are highly concentrated in a few geodemographic types.

There are 61 types of neighbourhoods in the 2001 Mosaic classification, numbered from 1 to 61, which are aggregated in 11 groups, identified by letters A to K (Experian Ltd., 2004). Some examples of geodemographic types in *Mosaic* that indicate the presence of certain ethnic minorities are; Mosaic Type C20 'Asian Enterprise' which has a particularly high proportion of residents classified by the Census as South Asian and of Hindu or Sikh religion. D26 'South Asian Industry' is an example of a Mosaic Type with a very high proportion of South Asian Muslim residents. F36 'Metro Multiculture' by contrast is a Mosaic Type with a high

concentration of more recent immigrant groups, only a small proportion of whom originate from South Asia. Other Mosaic Types with higher proportions of minority ethnic groups than the national average are A01 'Global Connections', with Jewish and Armenian names, E28 'Counter Cultural Mix', and D27 'Settled Minorities', which is characterised by a preponderance of Caribbeans, Greek Cypriots and Turks. Although these ethnicity and religious aspects could have been derived directly from Census data at Output Area level (average 100 households), the fact that Mosaic synthesises several other geographic and demographic factors into one variable and that it is available at the unit postcode level (average of 12 households), made it easier for the name analysis carried out here. However, a very similar result might have been achieved by just using UK Census data directly (Harris et al, 2005), combining the questions on ethnicity, religion, country of birth and household demographics.

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

Geodemographic analysis has proven useful for identifying certain non-British names that are typically very concentrated in a limited number of areas, particularly Jewish, South Asian and African names. However, this technique is less effective in

drawing distinctions *within* the major non-British CELs (intra-South Asian or intra-African divisions) or less residentially concentrated non-British CELs.

4.3.4. Text mining

Text mining embraces a series of techniques that seek to capture similarities in the morphology of names through text pattern analysis, in order to relate them to a particular language of origin and thus a CEL.

Most of the literature on text pattern mining of names comes from the computer linguistics field, where personal names, as well as other proper names, have arrived relatively late to a language modelling process in which computers are being adapted to dealing with natural (human) language. To illustrate this point it suffices to mention the problem of running a conventional spell checker software to any document that contains a number of proper names, to the frustration of the user who needs to skip over most of them since they are not found in the built-in dictionary. Names have been recently termed 'a new frontier in text mining' (Patman and Thompson, 2003: 27), and several research efforts seek to deal with the problems and exceptions of efficiently managing proper names from different languages in speech synthesis/recognition for automated voice systems (Bonaventura et al, 2003; Llitjos, 2002), in entity and semantic extraction (Leino et al, 2003), in identity linking across several databases and languages (Williams and Patman, 2005), in document searching and archiving, in data entry and cleansing operations, and so on. Other interesting approaches from statistics or biology have proposed a probabilistic approach for automatic discovery of character sequences patterns (e.g. Rigoutsos and Floratos, 1998). Most of these approaches start with a system to sort names into their

probable languages of origin, using amongst others text mining techniques, in order to apply subsequent rules to each language.

The text mining approach taken in this PhD thesis has been a heuristic one, 'letting the data speak for themselves' in respect of identifying commonly recurring character patterns in the names database that unequivocally lead to the identification of a CEL group in a set of names. There are two basic techniques to find forms of language commonalities between names; name stems and name endings on the one hand, and character sequences and character absences on the other.

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

1. Sort all names in alphabetic order.
2. For each unclassified name do steps 3 to 8.
3. Look at the 10 previous and 10 subsequent neighbouring names in the list (20 neighbours).
4. Identify which CEL these neighbours have been assigned to.

5. Assign a weight to those neighbouring CELs by inverse distance to the target name; a weight of 1 (farther) to 10 (closer) in each direction from the name (*a*).
6. For each CEL present in the 20 neighbouring names, sum up their weights ($\sum a$)
7. Re-sort all the names using the reverse of the name (termed *anti-alphabetic order*)
8. Repeat the process once from 3 to 8 ($\sum b$)
9. Create a total score per name and CEL as follows:

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

Where *S* is the score, *a* and *b* are the result of step 5 for the alphabetic and anti-alphabetic rounds, and the denominator (220) is the maximum possible total score for the 10 neighbouring names ($x=n*(n+1)/2$; which is in this case is 10*11/2 =55) in the 2 directions (55 x 2= 110) and 2 rounds (alphabetical and anti-alphabetical order, i.e. 110+110= 220)

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

An alternative method is to extract the first and last 2, 3, and 4 letters of a name, aggregate them and calculate their frequency in the name dataset, what makes it possible to locate the most common name stems and endings in a list of names. The CEL of each particular name stem or ending is then decided by using one of the other

techniques (i.e. non-text mining), and is then applied to all names with such forms that remained unclassified after using the previous techniques.

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

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

In the context of this research, text mining has proved useful for identifying non-British names which have been assimilated by the host community, for example those which, on the basis of FSC analysis, appear to be British but which are for example of Scandinavian origin. It is also a useful strategy for classifying large numbers of low frequency names, such as Spanish or Italian names in the UK, and this reduces the number of names that would otherwise remain 'unclassified'. It is

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

4.3.5. Name to ethnicity data

This is the method followed by most of the name studies in epidemiology, as reviewed in Chapter 3. As explained there, it is based on using population registers where the ethnicity (or a proxy) and the name of the person is already known in order to build appropriate name-to-ethnicity reference lists. Only two of the studies reviewed in Chapter 3 had access to a large enough population register to produce a reference list with a significant amount of unique surnames, in this case of over 20,000 surnames each (Lauderdale and Kestenbaum, 2000; Word and Perkins, 1996). These two studies satisfied the criterion established by Cook et al (1972) for minimum reference population size. Even then, these two major studies only aim to classify 7 ethnic groups, and not all of the possible ones.

As previously mentioned, it proved impossible to access a register of names in the UK which would include a large sample of the population surnames together with their ethnicities in sufficient numbers. Therefore, partial lists of names have been used in which a proxy for ethnicity was known, such as country of birth, or nationality.

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

This technique proved useful for 'seeding' new CELs out of pre-existing broader groups (specifically, of breaking Eastern European names down into Russian or Polish names), especially for 'rare' CELs, and thus needs to be used in combination with the FSC technique. Independently sourced name-to-ethnicity data is also very useful to validate the name classification, as it will be described in Chapter 7.

However, this technique does also present its limitations, the major one being the limited number of names for which ethnicity or place of birth is known. Even where these lists exist, two problems might be encountered; the availability of only a small number of ethnic groups (e.g. the 16 UK Census categories), and the differential distribution of names between; (a) periods of time; (b) receiving countries of immigration; and (c) regions within those countries, which all might introduce biases in the name to CEL attribution. These factors were explained in Chapter 3 Section 3.4.1.

4.3.6. Lists of international name frequencies and genealogy resources

This technique complements the others and essentially consists of accessing new and more anecdotal sources of name frequency data upon which all of the previous

techniques are based. It entails collating lists of names, and preferably their frequencies throughout as many countries as possible.

This has only been possible because of a significant surge of interest in genealogy and family history through the Internet in the last few years, which has only become apparent over the last ten years (see Rogers, 1995 for a striking account of the material available to a genealogist then). According to The Guardian, genealogy is now second only to pornography in generating internet traffic (The Guardian, 2004), and following this public thirst for information about ancestry, a range of data providers are willing to publish name data on the web, ranging from formal institutions such as national statistical offices, to amateur genealogist blogs. A previous project at University College London to show historic and current surname distributions in the UK has also leveraged this interest with over 3,000 daily visitors at www.spatial-literacy.org.

Amongst these types of lists, several resources are available from the official statistics offices or public registers in some countries, for example, lists of forenames and/or surnames and their frequencies in Belgium (Statbel, 2006), Denmark (Danmarks Statistik, 2006), Iceland (Statistics Iceland, 2006), Madrid and Catalonia in Spain (IDESCAT, 2006; Instituto de Estadística de la Comunidad de Madrid, 2006) and Germany (Gesellschaft für deutsche Sprache, 2006). These lists have been used to re-apply the techniques previously mentioned, so that broader CELs such as Scandinavian, Central Europe, or Hispanic, could be broken down into finer CELs, for example making it possible to distinguish between Wallonian and Flemish names in Belgium, or Catalan and Castilian in Spain.

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

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

4.3.7. Researching individual names

As a last resort, when names cannot be classified using any of the methods presented above the last resort is to search for a particular name. Such searches can be done either in name dictionaries, or in electronic telephone directories, or through a web search engine, such as Google. In this last one type of search, the objective is to find

particular associations between a name and a country or language through the contextual information in which they are found on the web. A similar technique to link geographic information found in miscellaneous web content, termed 'heuristics for geo-referencing web pages', has been developed by Silva *et al* (2006) to perform such associations automatically. The obvious limitations of this method are that it is time consuming, dictionaries are only available for a few names or countries, and different CELs have a very different presence on the Internet (e.g. African names are misrepresented in the web), or duplicate and competing CELs are presented for some of the same names. However, this method has proved very useful to seed new CELs into the FSC technique, where a forename or surname has a high proportion of corresponding names that are not classified – as, for example, to identify Fijian and Lao names in Australia or Ethiopian names in the UK.

4.3.8. The name pattern analysis toolbox

The set of seven classification techniques described in this section summarises the different approaches to name classifications found in the literature. They also comprise the preliminary outcomes of testing them on the UK data in this research, identifying the types of situations in which they are most or least useful. The underlying principles behind each of these techniques are consistent throughout the literature, and can be briefly summarised as finding patterns in the naming practices, geographic distribution, name morphology, and linguistic associations in names. These form the core set of tools available to the task of classifying a universal register of names into a comprehensive taxonomy of cultural, ethnic and linguistic origins.

4.4. Conclusion

This chapter has set down the three cornerstones upon which the name-based ethnicity classification developed in this thesis will be based; name taxonomy, data and techniques. These three components form the interwoven elements of a system that will be further developed in subsequent chapters, but that needs to be explicitly specified and discussed from the outset.

Although many possible classifications of human groups have been proposed, for the purpose of classifying them in terms of their name origins the use of a linguistic classification as a base for such taxonomy was amply justified here, following onomastic criteria. The Cultural, Ethnic and Linguistic classification, hereinafter the CEL classification, is ultimately based in Hanks (2003) DAFN dictionary's taxonomy, but is expanded here to many more groups and adapted to the British context and the purpose of this thesis. Its terminology and relationships are now based on the most commonly accepted language classification (Ethnologue- ISO 639-3), adapted to naming practices and regions according to religious or geographic factors. As a result, this research has proposed a taxonomy of 185 CEL types primarily built to classify ethnic groups as found in contemporary British society, and as such the hierarchy of CEL types are structured according to the relative size of each group and their names frequency distribution in Britain.

In order to fulfil the main objective of this thesis, i.e. to develop a name-based ethnicity classification to cover all potential groups in Britain, a population register with detailed name and origin data and universal coverage would have to be sourced. However, such registers with name and ethnicity data, or a proxy for it, are not

publicly available. After evaluating several alternatives, the strategy adopted was to collect a series of universal name and address publicly available population registers, namely telephone directories and Electoral Registers from different countries and different levels of disaggregation, and exploit them in innovative ways to classify their names by ethnicity following a series of data mining techniques found in the literature.

These techniques were used to put these two components together; the universal list of names to be classified into the CEL taxonomy. Three alternative broad methods were repeatedly found in the literature; forename-surname clustering, geographical analysis of names, and name morphology pattern analysis. These methods were subdivided into more specific techniques in combination with the more traditional name to ethnicity methodology reviewed in Chapter 3, to form seven techniques that have been described in detailed in this chapter; forename-surname clustering (FSC), spatio-temporal analysis, geodemographic analysis, text mining, name to ethnicity data, lists of international name frequencies and genealogy resources, and researching individual names.

Therefore, taxonomy, material and methods, as the title of this chapter suggests, form the set of tools with which an improved classification into cultural, ethnic and linguistic groups of all the names occurring three or more times in Britain has been attempted in this research, following two different approaches; a heuristic and an automatic one. Both approaches are described in the next two chapters.

Chapter 5. Heuristic Approaches to Creating a Name Classification

As discussed in the previous chapters, the end objective of this research is to classify into cultural, ethnic and linguistic groups (CELs) every surname and forename present in Britain that has a frequency of 3 or more people, and to assign a probability of association between each name and its CEL (here termed 'name score'). Using the main population register data source available to this research, the 2004 Electoral Register (and Consumer Dynamics enhancement), this means building and classifying a reference list of 281,422 surnames and 114,169 forenames. The datasets to be used as well as the potential classification techniques have been described in the previous chapter. It has been emphasised that the lack of available data linking name and ethnicity with extensive coverage of the UK population makes it necessary to develop alternative approaches for the purposes of this research. In methodological terms, there are many different ways in which the tools described in the previous chapter might be applied.

Two approaches to classify the lists of names mentioned above were followed during the PhD research: an initial exploratory approach, described in this chapter, during which heuristics were identified and developed; and a subsequent automated and integrated approach, described in the next chapter (6), that sought to distil the accumulated experience of the exploratory phase in a robust and transparent manner. The name classification resulting from the automated and integrated approach is then evaluated in Chapter 7.

The initial heuristic approach, covered in this chapter, laboriously classified different groups of names following the techniques described in the previous chapter (4), as they were first investigated, specifying them through different rules and applying them to different stages in a dynamic and iterative process. Because of the cumulative and exploratory nature of this approach, it is described as 'heuristic', fitting well the definition of this term in found in Oxford English Dictionary; '2- [computing] proceeding to a solution by trial and error or by rules that are only loosely defined' (Compact Oxford English Dictionary, 2005: 285)

The heuristic procedures are summarised in this chapter through a core set of these rules and processes and their sequence of use in the decision process, which have been substantially synthesised here for the purpose of clarity and brevity. The actual detailed process was in practice much more complex, in terms of number of exceptions and iterations, and unfortunately cannot be fully described in the space available in this thesis.

The reason to adopt this ad-hoc methodological approach to classifying name lists is because the research did not start with any pre-conceived notions of the optimal methods to classify all of the most frequent names according to their origins, and neither were all of the datasets available from the outset. Therefore a series of exploratory rules were tested and applied in a sequential process guided essentially by pragmatic considerations, not necessarily in the most logical order. This essentially heuristic approach shaped the way in which the first version of the final

CEL name classification was built, the techniques that were employed, and possibly the results that were obtained, as it will be evaluated later.

The chapter is structured in six sections. Section 5.1 describes how the heuristic approach was developed in a three stage design, how each of these was linked to availability of different tiers of names and how this presented a moving target in terms of the number of names to be classified. Sections 5.2 to 5.4 describe the detailed rules, decisions and processes for each of these three stages and tiers: section 5.2 considers the stage and tier 1 consisting of the top 25,630 surnames; Section 5.3 the stage and tier 2 consisting of the top 30,000 forenames; and Section 5.4 for stage 3 including the remaining surnames and forenames to reach a final reference list of 281,422 surnames and 114,169 forenames. In order to illustrate this sequence of stages, a comprehensive flow chart is included in each of these sections (Figure 5.1, Figure 5.2 and Figure 5.4). Finally Section 5.5 summarises how all of the classified names were put together in name-to-CEL tables.

5.1. Stages in the Creation of the Classification

A series of data sources concerning several reference populations were used in order to create a name reference list. These were initially described in Section 4.2.2 and were summarised in Table 4.2 as nine separate datasets (Great Britain- GB 2004, GB 1998, GB 1881, Northern Ireland, Republic of Ireland, United States, Canada; Australia, New Zealand, Spain). However, two of these datasets could not be sourced at the beginning of this research project; the GB Electoral Roll & Consumer Dynamics file and the Spanish Telephone directory, both for 2004. The former dataset will hereinafter be called ‘GB04’, and together with the Spanish directory is

the most detailed of the datasets in terms of the number of surnames and forenames that they contain and their resolution at the individual level. They only became available at a late stage in the research project. The other eight datasets were all available from the outset, but only included surname data (not forenames) and the records were aggregations of individuals to some coarse level of geography (i.e. no individuals or neighbourhoods, for details see column 'Nominal Resolution' in Table 4.2).

It must be also added that this PhD research built upon the cumulated experience of the UCL Surname Profiler project (Surname Profiler, 2006), which mapped the geographic distribution of the most frequent 25,630 surnames in Great Britain in 1881 and 1998 and used two of the datasets mentioned above. In this project, the 25,630 surnames were classified according to their etymological or regional origin, most of them within the British Isles. However, this list included a minority of 2,978 non-British surnames which were classified into 12 broad 'international groups'. This last list formed the initial motivation to start this PhD, and the description of the process to develop it is incorporated in the stage 1 described in Section 5.2.

The non-availability of some data sources had implications for the research design, and the way in which the resulting 'heuristic CEL name classification' was built. If all of the datasets had been available from the outset, and it had been understood that the ambitious objective of classifying such a large number of surnames and forenames was attainable within the lifespan of a single PhD, it is likely that other methodological paths would have been followed. In the event, this belated realisation, and the subsequent availability of all of the reference datasets provide the

motivation for developing the automatic classification described in Chapter 6. This arises from the accumulated experience of classifying 281,422 surnames and 114,169 forenames into cultural, ethnic and linguistic groups (CELs) in three stages, and the development of three distinct 'tiers' of names used as inputs into the classification. The characteristics of these three stages are summarised in the following paragraphs, and they are explained in detail in Sections 5.2 to 5.4.

5.1.1. Stage 1 and Tier 1 names

This stage formed part of the mentioned UCL Surname Profiler project (Surname Profiler, 2006) upon whose 'international surname list' this PhD was subsequently built. The first dataset to be sourced from *Experian* in stage 1 was an extraction of the GB Electoral Roll for 1998, consisting of 25,630 surnames with a frequency of 100 people or more nationally, together with their frequencies by postal area, a geographical division defined by the first two digits of the unit postcode (e.g. 'BR' for Bristol). There are 120 postal areas in Great Britain, with an average population of 500,000 people. This dataset will be referred to in this thesis as the 'GB98 dataset', and formed the backbone of the UCL Surname Profiler project (Surname Profiler, 2006). The aim of the first phase of the heuristic approach was to classify these 25,630 unique surnames into British regions and major ethnic minority groups of origin. This table of most frequent surnames will hereafter be termed 'Tier 1' names. Moreover, the other seven datasets that were obtained at approximately the same time as GB98 were used primarily to support the classification of the surnames in 'Tier 1', as will be detailed in Section 5.2.2.

5.1.2. Stage 2 and Tier 2 names

At the beginning of stage 2 the 'GB04' file was received from *Experian* (Electoral Roll & Consumer Dynamics for 2004), which included the forename, surname and unit postcode for each of 46.3 million electors/residents. At this stage only, it was considered appropriate to add to the existing CEL name classification the capability to classify forenames as well as surnames. Therefore, a list of the most common forenames, defined as those with a frequency of at least 9 people, was extracted from 'GB04' and produced a file with 29,979 forenames, hereinafter termed 'Tier 2' names. The methodology to link Tier 2 to Tier 1 and classify it into CELs is detailed in Section 5.3.

5.1.3. Stage 3 and Tier 3 names

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

5.1.4. Classification of Tiers 1, 2 and 3

As a result, Tier 1 names contained the top 25,630 surnames, Tier 2 names the top 29,979 forenames, and Tier 3 names the rest of both surnames and forenames. The

three Tiers combined comprise a total of 281,422 surnames and 114,169 forenames. The seven classification techniques described in Section 4.3 were applied to each of the three Tiers of names described in the previous Sections 5.1.1 to 5.1.3, according to a set of rules which will be summarised in the following three sections.

5.2. Tier 1 Names: 'Top' Surnames

5.2.1. Data preparation

The names in the Tier 1 file were initially processed to eliminate data errors, such as inconsistent geographic indicators or invalid entries (e.g. 'N/K'), and to standardise the format by trimming spaces, and unifying different dashes, apostrophes and other special characters used. These data cleansing steps were required to make sure that there was a common entry for each unique name across the seven datasets that were to be compared (i.e. the nine datasets in Table 4.2, excluding GB2004 and the Spanish Directory). Other known errors in the data, such as the presence of name initials or honorifics (e.g. Prof., Ms., Dr., Sir), were kept in a separate field since they were judged to be able to provide some valuable information for later classification tasks.

Each surname frequency per postal area was converted to a rate per 1 million people, in order to be able to compare the names in a consistent way across all the geographies studied. Two additional geographies were created for the purposes of calculating additional name frequencies and rates, through the aggregation of some countries into bigger regions; *All Ireland*, including the Republic of Ireland and Northern Ireland, and *British Isles*, including the former plus Great Britain.

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

Table 5.1: List of attributes associated with each name in 'Tier 1'
Country codes used: UK= United Kingdom; GB = Great Britain (ex. NI); NI= Northern Ireland; IE= Republic of Ireland; All Ireland = IE+NI; British Isles= IE+UK; US= United States, CA= Canada; AU= Australia; NZ= New Zealand

In total, the surname frequency datasets at this stage included a total of 9 'countries or territories': Great Britain (GB 2004, GB 1998, GB 1881), Northern Ireland, Republic of Ireland, United States, Canada; Australia, New Zealand, All Ireland and

British Isles. These were linked together through the common surnames between them. For each surname in Tier 1, and for each of these 9 geographies, a set of 10 different statistical and geographic variables were calculated, the details of which are offered in Table 5.1 (items ***a*** to ***j***).

After these two steps of data cleansing and name frequencies calculation, each name was linked to both a UK and a Worldwide place name gazetteer (Edina, 2006; National Geospatial Agency, 2006). This made it possible to evaluate whether there were any gazetteer entries for each name, and if so to count them by region of occurrence. The results of this search were stored for the UK gazetteer and for entries in Africa or Asia in the Worldwide gazetteer (items ***l*** to ***o*** in Table 5.1), since America, Europe and Oceania had substantial numbers of names in countries not corresponding to their apparent region of language origin (i.e. English or Spanish place names).

Finally, the *Mosaic* neighbourhood classification was included, using a separate table provided by Experian with the distribution per *Mosaic* Type for each surname. Using this table and for each surname in Tier 1, the *Mosaic* Type with the highest rate of names per million people was selected (item ***p*** in Table 5.1). This made it possible to calculate the percentage of names in rural versus urban areas, as well as in high socioeconomic status neighbourhoods (items ***r*** and ***s*** in Table 5.1)

As a result, the Tier 1 names table included for any name, between 8 and 19 attributes for each of the 9 geographies studied, resulting in over 100 attribute

combinations for some of the most common names present across all the geographies. These attributes are fully described in Table 5.1.

5.2.2. Classification rules applied to Tier 1 names

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

Non-British names, by exclusion, are defined in this research as those which arrived in the British isles after 1700. These were originally divided into the following 12

major CEL groups present in the UK, and were reasonably straightforward to isolate; African, Scandinavian, Greek or Cypriot, Jewish and Armenian, Hispanic, Rest of Europe, East Asian, Japanese, Muslim, Sikh, South Asian, and 'non-British Unclassified'. In subsequent Tiers 2 and 3 these 12 *CEL groups* were further subdivided into much finer CELs denominated *CEL types* (e.g. the Hispanic CEL group into Spanish, Portuguese, Catalan, Basque, and Galician CEL types).

Offering a detailed account of all the variables and thresholds considered in each decision to assign a CEL to a name is much beyond the scope of this thesis. A summary of the main eight rules and decisions taken is offered here, in order to help to understand how the classification was created and the decisions and assumptions taken. These eight rules have been numbered A.1 to A.8. A chart with the decision tree of these rules is offered in Figure 5.1 and should be used to accompany the text.

Rule A.1

In order to split between British and non-British names, the main rule applied was to check whether a surname was present in the 1881 Census (potentially British) or not (potentially non-British, of which there were 2,374 surnames). Even if the surname was present in 1881, when the increase in its rate per million names between 1881 and 1996 was judged to have been high (over 100% on average) it was considered potentially non-British, adding over 1,000 more surnames (mainly European) to this list. This rule A1 was developed as part of the UCL Surname Profiler project, and as a result 22,652 names were classified as British and 2,978 as non-British out of a total of 25,630 surnames.

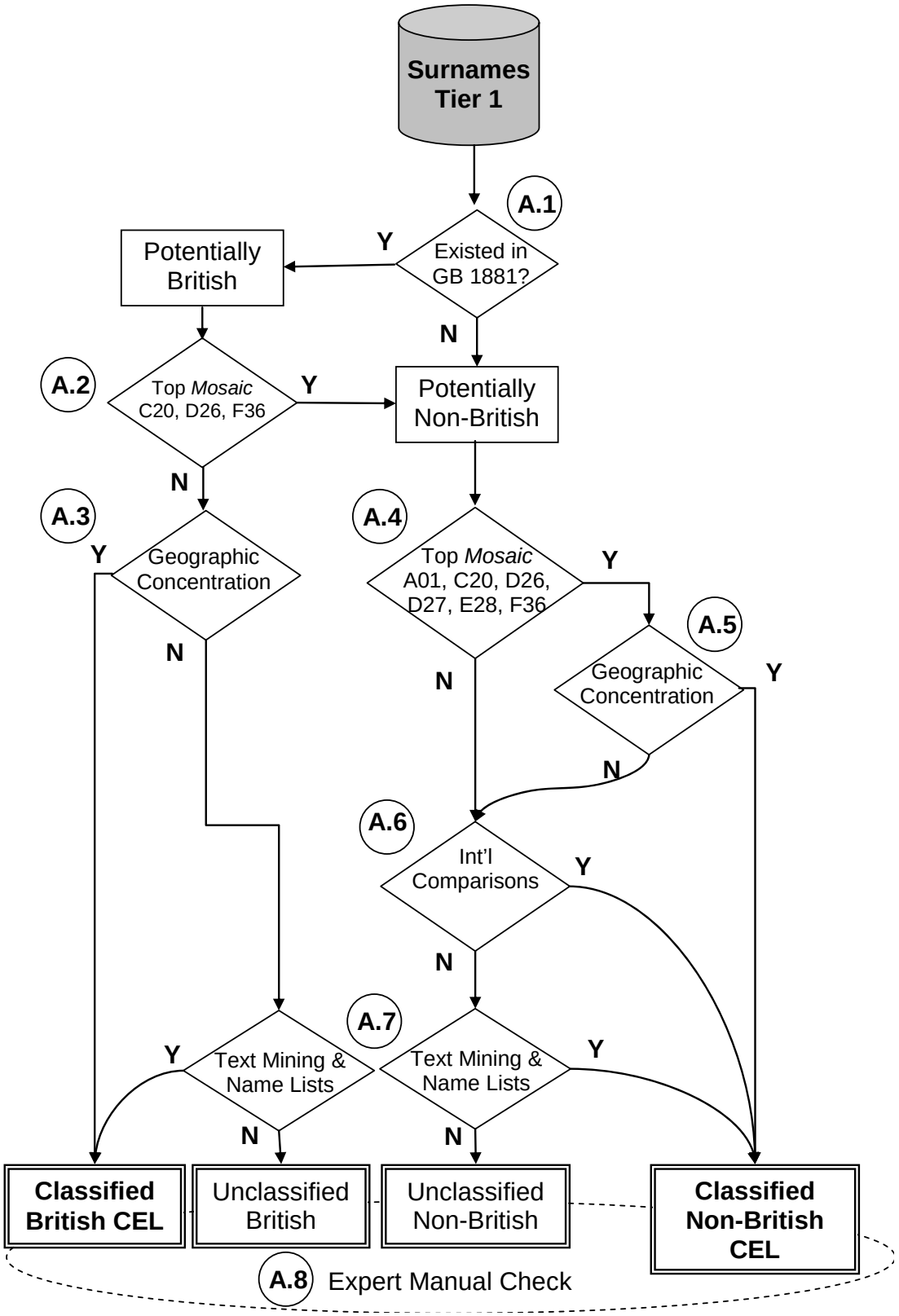

Figure 5.1: Classification Decision Tree for surnames Tier 1
The code in a circle relates to the reference of each rule in Tier 1, A.1 to A.8 and is described in the text.

Rule A.2

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

Rule A.3

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

Rule A.4

The Top *Mosaic* Type of a potential non-British surname was used to identify specific CELs (letter ***p*** in Table 5.1). Those with C20 'Asian Enterprise' were pre-assigned the CEL 'South Asian', since this type has a particularly high proportion of residents classified by the census as South Asian and of Hindu or Sikh religion. Those surnames with D26 'South Asian Industry' were assigned to the CEL 'Muslim', since the majority of residents are of South Asian ethnicity and Muslim religion. Finally, *Mosaic* type F36 'Metro Multiculture' representing more recent immigrant groups, especially Black Africans, was provisionally assigned to the CEL 'African' when Top Areas were in South London, since these present a high concentration of this CEL. Other *Mosaic* types with high proportions of minority

ethnic groups are A01 'Global Connections', with Jewish names, E28 'Counter Cultural Mix', and D27 'Settled Minorities', which contains mostly Caribbeans, Greek Cypriots and Turks. However, to avoid mistakes in the assignment due to the ecological fallacy, this rule was best used in combination of rules A5, A6 and A7.

Rule A.5

Through spatial analysis of the top areas of potentially non-British surnames, groups of postal areas with much higher concentrations of non-British names were easily identified. Local knowledge of the postal geography of the UK, and specifically of London, allowed these groups of surnames to be provisionally assigned into CELs, for example Greek or Turkish names in postal area 'N', Jewish in 'NW', or South Asian in 'LE'. However, the socioeconomic characteristics of the *Mosaic* classification were also taken into account and overrode the spatial considerations. For example, more affluent *Mosaic* Types are identified with certain ethnic groups (e.g. Japanese, Scandinavian, or Jewish), and some groups are mostly only present in cities, and thus their index of rurality would be expected to be very low (e.g. Jewish names are highly urban).

Rule A.6

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

Rule A.7

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

Rule A.8

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

Outcome

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

5.3. Tier 2 Names: 'Top' Forenames

5.3.1. Data preparation

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

However, since no other forenames dataset was available for any other geography, a new strategy was required to provide sufficient variables to classify these forenames. Use of the Forename-Surname Clustering (FSC) method, based on Tucker's (2005) CELG technique and described in Section 4.3.1 in the previous chapter, proved to be the most efficient option using the existing surnames classified in Tier 1. The key was then to use the existing CEL Group assigned to surnames in Tier 1, hereafter called SCEL Group (for Surname CEL Group), to cluster forenames into the same groups. For each forename in Tier 2, the proportion of its bearers by SCEL was

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

Forename: ***Pedro*** *Total Frequency: 3,435*

SCEL Groups: British 245 (7.1%), Hispanic 2,410 (70.2%), European 35 (1.0%), 'None' 745 (21.7%)

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

5.3.2. Classification rules applied to Tier 2 names

As becomes obvious by now, the main method used to classify Tier 2 forenames into CELs was the FSC method, which was applied in a series of steps in combination with some of the other classification techniques described in Section 4.3. The decision tree of the eleven rules applied to classify the forenames in Tier 2 file is illustrated in Figure 5.2, and the explanation of each rule is offered in the next paragraphs. As a result of this process, 29,979 forenames were classified into CELs, which hereinafter will be termed FCELs (for Forename CELs).

Rule B.1 – Unclassified forenames

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

Rule B.2 – Void forenames

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

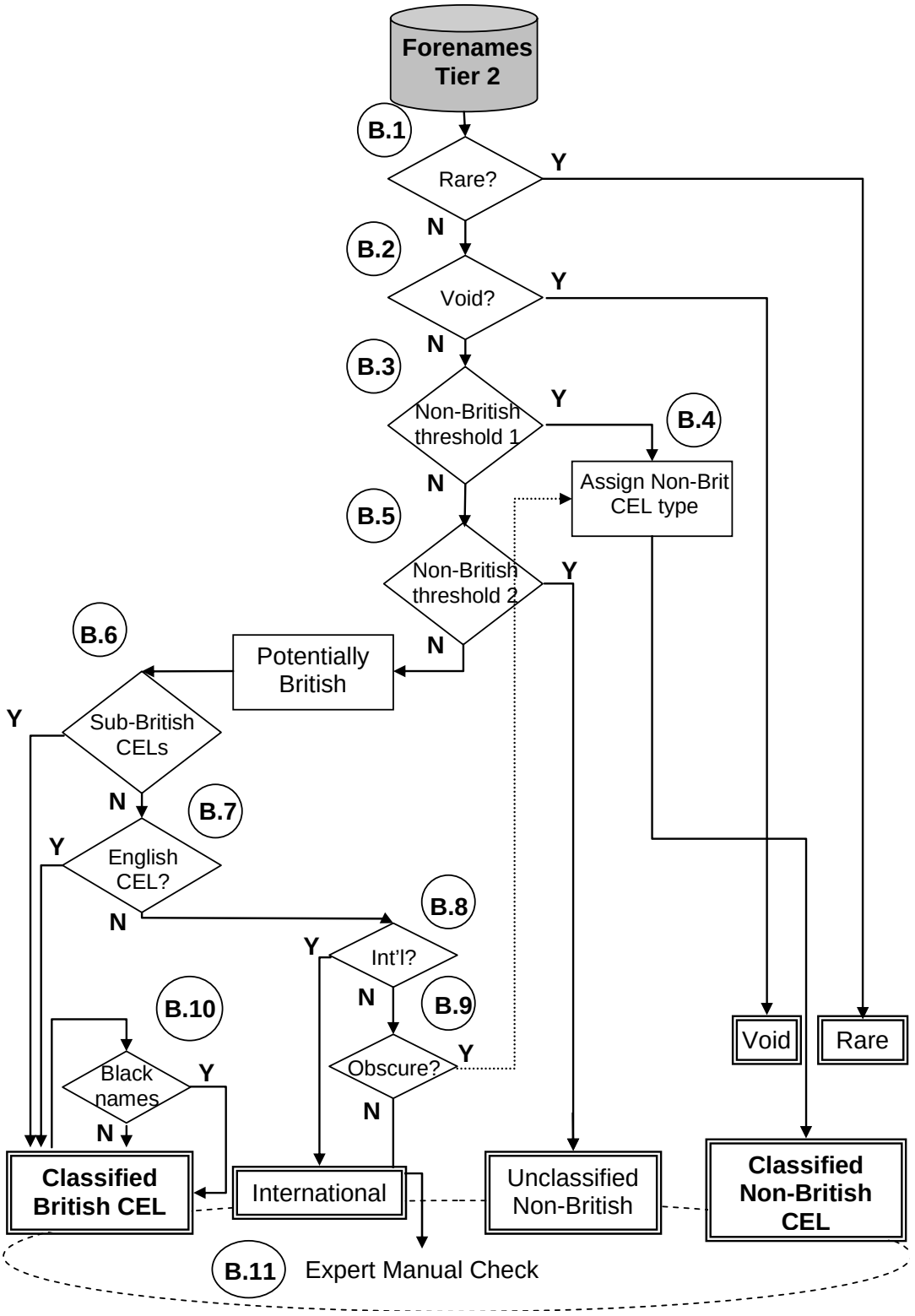

Figure 5.2: Classification Decision Tree for forenames Tier 2
The circled codes refer to each rule in Tier 1, B.1 to B.11 as described in the text.

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

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

If a forename had at least 2 occurrences in one of the non-British SCEL Groups (excluding Jewish), and if this frequency represented 5% or more of the occurrences in the combined British and Jewish SCEL Groups, the FCEL Group was made equal to that of the SCEL Group. If more than one SCEL Group met this threshold criterion the FCEL Group was assigned to the SCEL Group with the largest number of occurrences. If two SCEL Groups had an equal number of forename occurrences, the FCEL Group was then assigned to the SCEL Group with the smaller or smallest total number of occurrences on the entire Electoral Register file. In all of the above situations the SCEL Group 'None' was not taken into account. For example:

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

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

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

Rule B.4 – non-British FCEL Types

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

Taking the example of the forename 'Ourania', the following calculation was performed:

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

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

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

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

Rule B.5 – non-British Unclassified

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

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

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

In all likelihood names such as 'Joris' that met this criteria were surnames from a CEL which were not included in the list of CELs used in the analysis. In the case of 'Joris' this name, according to the Oxford Dictionary of First Names (Hanks et al,

2003), originated from Frisia. However, at this stage it was decided to leave these rather obscure non-British forenames in the 'non-British Unclassified' FCEL for further work on their surnames in Tier 3.

Rule B.6 - British and Jewish FCEL Types

Where on the basis of rules B.3, B.4, and B.5, a forename had not been assigned to a non-British or Jewish FCEL Type, it was assumed that the forename was likely to be of British or Jewish origin. The reason why the Jewish CEL Type was treated together with British CEL Types is because although Jewish surnames are quite distinct, they carry a high proportion of British forenames because of a long history of integration into British society, compared to the other non-British CELs (bearing in mind that, as stated before, the 'British' CEL includes all names originating in the British Isles or that immigrated up to around 1700).

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

value of *ratio 'c'* was assigned to that forename's FCEL Type. This is illustrated in the following example:

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

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

Those forenames which did not reach a *ratio c* of a least 2, were deemed most likely to be English, since the average of 69.4% SCEL Types prevented their meeting rule B.6, and thus they were dealt with in the next rule B.7. Therefore B.6 identified Irish, Scottish, Welsh and Jewish FCEL Types.

Rule B.7 – English FCEL Type

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

Rule B.8 – International FCEL Type

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

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

SCEL Groups [SCEL Types]:

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

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

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

Rule B.9 – Obscure non-British FCELs

A further test was applied to the set of forenames that at this stage still remained unclassified, and with a minimum of two occurrences in any one non-British SCEL Type. At this stage, these forenames were subject to rules B.3 and B.4, except for the criterion of a having a non-British frequency of 5% or higher of the occurrences in the combined British and Jewish SCEL Groups. This is illustrated with the example *Nelson*:

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

SCEL Groups [SCEL Types]:

British or Jewish 1,139 (69.96%) [English 756, Welsh 143, Scottish 135, Irish 95, Jewish 2]

non-British 489 (30.04%):

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

From these figures it can be seen that the largest 'non-British' SCEL Group was Hispanic, with 55 occurrences. This number, as a proportion of those with a British or Jewish SCEL (1,139), was just under the 5% threshold, so *Nelson* did not qualify in the initial Rule B.3 as a non-British forename. However, the proportion of non-British SCELs (30.04%) was too low for the name to have been considered a 'non-British Unclassified' in Rule B.5. The name was not especially associated with Irish, Scottish, Welsh or Jewish SCEL Types in rule B.6, and, because the proportion of occurrences with an English SCEL was slightly below average and non-British SCELs just above 30%, the name did not qualify as English either. However with 42 of the 55 Hispanic SCEL occurrences classified as Portuguese, the name Nelson qualifies as Portuguese. This assignment was also corroborated by some anecdotal historic research.

Rule B. 10 – Black British forenames

Finally, those forenames which by this stage have been assigned a FCEL Type English, Irish, Scottish, Welsh or 'International' (but not Jewish) were selected. For each of these forenames, the proportion of people that were resident in postcodes classified by the *Mosaic* geodemographic classification as being predominantly Non-White British based on Census data. These are *Mosaic* Type codes: A01 – Global Connections; C20 – Asian Enterprise; D26 – South Asian Industry; D27 – Settled Minorities; E28 – Counter Cultural Mix; F36 – Metro Multiculture, where overall, 7.2% of British households lived in such types of neighbourhood at the time of the 2001 Census. When the proportion of occurrences of a forename in such neighbourhoods exceeded four times the national average, i.e. exceeds 28.8% of all

occurrences, and the proportion of occurrences with British or Jewish SCEL Types was equal to 80% or more, the name was assigned to the FCEL 'Black British'. Note that there was no corresponding SCEL for 'Black' names, since most of these names are associated with Caribbean immigrants or their descendants, most of whom hold British surnames. However there are many forenames that are highly frequent amongst Blacks, a fact also found in the US by Levitt and Dubner (2005).

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

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

SCEL Groups [SCEL Types]:

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

-Non-British 201 (18.9%):Hispanic 7,'Non-British Unclassified' 169, Others [Lithuania 5, Portuguese 4, Spanish 3, Others]

Applying Rules B.1 to B.9 *Hyacinth* was assigned the FCEL Type 'English'. However 40.8% of all occurrences of 'Hyacinth' are resident in one of the six disproportionately Non-White British *Mosaic* neighbourhood types, significantly above the threshold of 28.8% required under rule B.10. The name 'Hyacinth' therefore is one example of many forenames found among people with British surnames but who live predominantly in residential neighbourhoods with a high proportion of Black British or Black Caribbean population in the Census.

Outcome

At the end of these 10 rules described for the Tier 2 file, the 29,979 forenames were classified into FCEL Groups and Types (13,600 British Group and 16,379 non-British) including a small proportion of them being assigned to the 'Unclassified'

and 'non-British Unclassified' FCELs (434). These forenames covered 45.3 million people out of the 46.3 million in the GB04 file.

5.4. Tier 3: Rest of Names

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

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

5.4.1. Classification by Forename-Surname Clustering (FSC)

There is a known difference in the frequency distribution of surnames and forenames, the latter with a higher average number of people than the former (Tucker, 2007b). This is explained by a relatively smaller pool of names from which a society selects children's forenames, together with the temporal effects in their

naming fashions. This contrasts with the fixed nature of surnames, a proportion of which disappear due to a process of 'natural selection' (Manni et al, 2005). This feature of names has been noted in different countries, for example in the U.S. (Tucker, 2003) and Spain (Mateos, 2007).

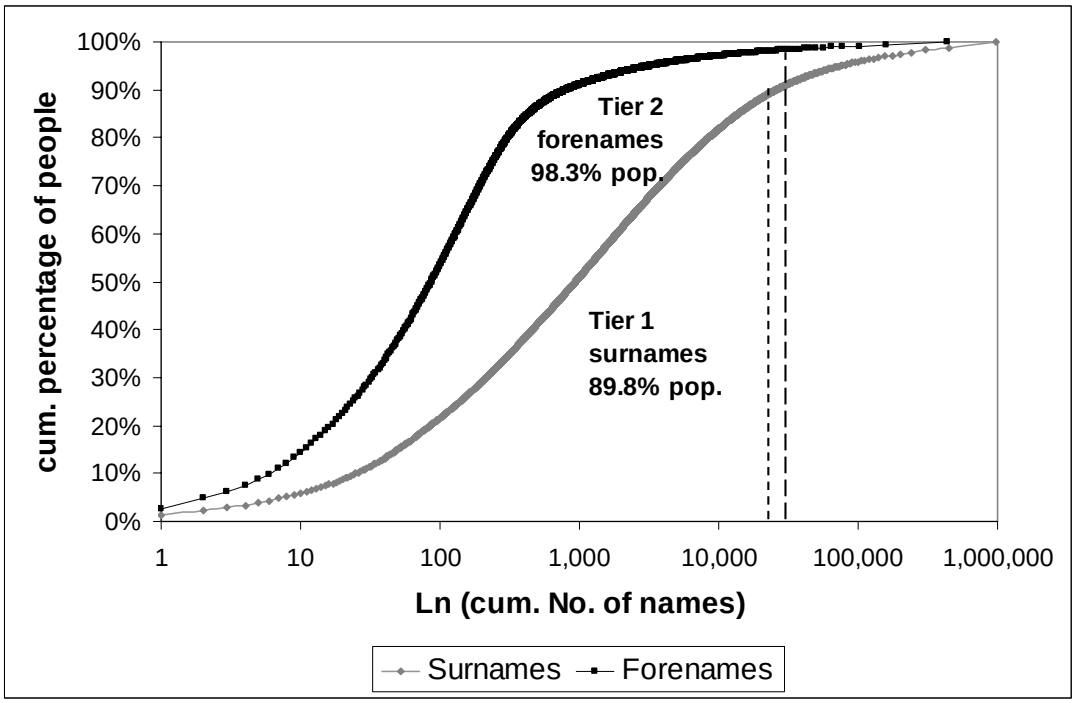

Figure 5.3: Graph of cumulative number of surnames and forenames (log scale) against cumulative percentage of population in the GB 2004 Electoral Roll

Figure 5.3 illustrates this difference in the frequency distribution of forenames and surnames for the UK Electoral Roll. The graph shows the cumulative number of surnames or forenames, on a logarithmic scale on the x-axis, against the percentage of population covered, in the y-axis. If the logarithmic scale is not used both curves are so highly positively skewed that no difference is appreciated. The vertical dotted lines represent the cut-off points of both Tier 1 surnames (25,630) and Tier 2 forenames (29,979), and hence the area to the left of these dotted lines represent the total population classified by Tier 1 (89.8% of the total Electoral Register) and Tier 2 (98.3%). The area between the two curves actually represents the number of people

in the Electoral Register for whom their forename is classified but the surname is not, although it must be stated that the forenames and surnames are not individually paired in this chart.

This difference between the degree of skewness in the frequency distribution of surnames and that of forenames actually permitted the classification of names in Tier 3 in a relatively effortless way, by using the Forname-Surname Clustering (FSC) technique described in Section 4.3.1. The 25,630 surnames in Tier 1 covered 37.2 million residents, while the 29,979 forenames in Tier 2 covered 45.3 million residents, and there was obvious potential to find out more information about the surnames of the 8.1 million people whose forenames appeared in Tier 2 (known FCEL) but whose surnames remained unclassified and thus were present in Tier 3 (unknown SCEL). These (several thousands) surnames were classified into SCELs using the FCEL distribution of those 8.1 million people, that is, the FSC technique using Tier 1 and Tier 2 SCELs and FCELs. Such classification was performed in a step-wise approach through a series of iterations of the same process that is summarised in Figure 5.4.

Therefore, each iteration or cycle, which are termed ‘C.a.’ to ‘C.d’ in Figure 5.4, aimed to expand the number of names classified, leveraging upon the mentioned difference between the frequency of forenames and surnames, and using all of the names whose CEL were already known at each step (i.e. using Tier 1, Tier 2 and Tier 3 files). Unlike the classification processes of Tier 1 and 2 previously described, in Tier 3, the cross-CEL distributions for the whole dataset of names (i.e. the count and

percentage of each SCEL and FCEL Group and Type associated with each name) were re-calculated at the end of each cycle, thus making the process a dynamic one.

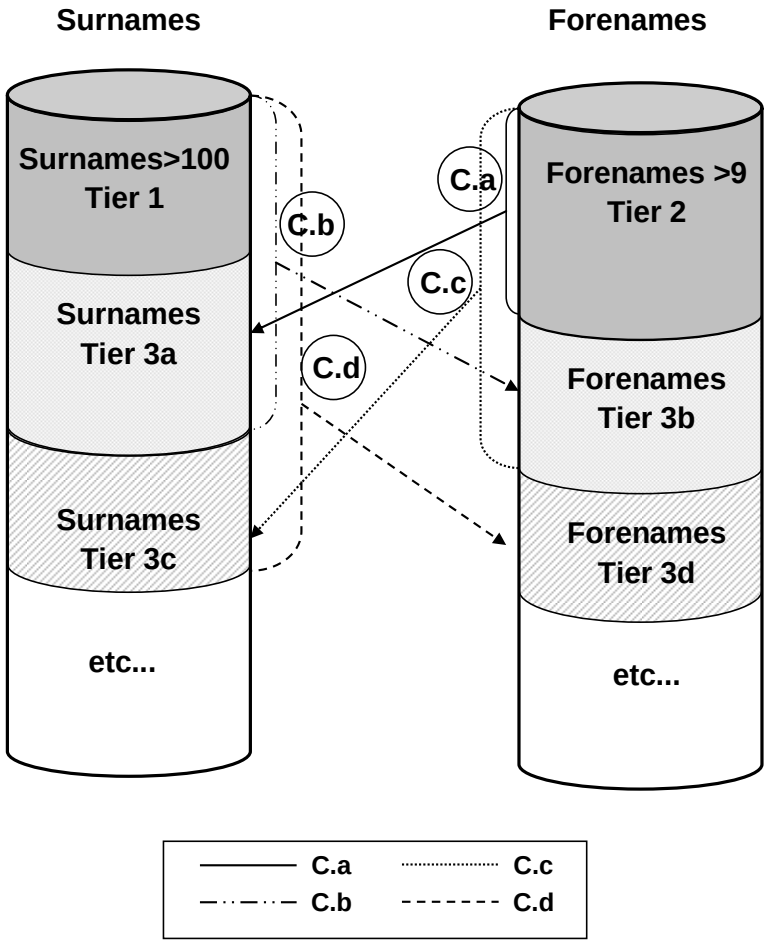

Figure 5.4: Iterative processing of name classification cycles in Tier 3
Cycles start at 'C.a' which runs from right to left producing set 'Tier 3a', then 'C.b' from left to right producing 'Tier 3b', etc.

These cycles were run several times, as shown in Figure 5.4, and as the volume of both surnames and forenames classified grew, it shed new light on previously unclassified names in Tier 3. Finally, the process stopped when most of the names in Tier 3 were classified and a few remained unconnected with the rest of the names. This process could even have resulted in the change of a CEL allocation in Tier 1 or Tier 2, when the CEL distributions of the names in Tier 3 pointed to errors in the CEL allocations previously made.

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

5.5. Name-to-CEL Tables

At the end of the Tier 1, 2 and 3 processes, all the classified name files were merged into two separate tables: a surname-to-CEL table with 281,422 surnames, and a forename-to-CEL table with 114,169 forenames. For each name in these two tables the following fields were available: the name's frequency in the GB04 file, the CEL Group, and the CEL Type. These two tables will hereinafter be referred to as Name-to-CEL tables. See Table 5.2 for a summary of the classification results.

These tables were then applied to classify the names on the patient register of Camden PCT, with approximately 212,000 records. This exercise proved useful to explore a real scenario of application and start to test the methodology to classify populations. In doing so, it became apparent that there was a need for a further algorithm to arbitrate in potential situations when there was a conflict between two CELs for the forename and surname (FCEL and SCEL) of the same person. As a result, a new phase was initiated to create the desired name-to-CEL strength scores to be used in such arbitration. These efforts came to fruition in the development of the automated approach, which is described in the next chapter.

CEL Type	Nr Fore-names	Nr Sur-names
AFRICAN	1496	466
BLACK SOUTHERN AFRICAN	258	319
BURUNDIAN	0	6
CAMEROONESE	0	12
CONGOLESE	35	289
ETHIOPIAN	54	249
GAMBIAN	0	4
GHANAIAN	486	2725
IVORIAN	0	15
KENYAN	0	38
NIGERIAN	2365	4745
RWANDAN	0	7
SENEGALESE	0	3
SIERRA LEONIAN	200	170
TANZANIAN	3	6
UGANDAN	3	65
ZAMBIAN	0	22
ZIMBABWEAN	0	41
CELTIC	1270	2599
IRISH	2493	8447
NORTHERN IRISH	0	257
SCOTTISH	2517	13176
WELSH	3136	6466
BURMESE	61	144
CHINESE	831	411
EAST ASIAN & PACIFIC	456	32
HONG KONGESE	2048	915
INDONESIAN	0	7
MALAYSIAN	95	27
MALAYSIAN CHINESE	3	50
MAURITIAN	53	391
POLYNESIAN	6	3
SINGAPOREAN	40	11
SOUTH KOREAN	78	43
THAI	48	25
VIETNAMESE	393	205
BLACK CARIBBEAN	1243	364
CHANNEL ISLANDER	0	172
CORNISH	1	413
ENGLISH	12658	119086
AFRIKAANS	10	203
ALBANIAN	105	162
AZERBAIJANI	0	6
BALKAN	262	4611
BELGIAN	2	95
BELGIAN FLEMISH	0	314
BELGIAN WALLON	0	24
BOSNIAN AND HERZEGOVIAN	3	69
BRETON	12	0
BULGARIAN	0	53
CANADIAN	0	4
CROATIAN	59	89
CZECH	89	247
DUTCH	159	1932
ESTONIAN	0	61
EUROPEAN	1437	3382
FRENCH	537	4734
GEORGIAN	0	131
GERMAN	658	6690
HUNGARIAN	116	560
ITALIAN	1946	11001
LATVIAN	39	301
LITHUANIAN	48	595
MACEDONIAN	16	84
MALTESE	3	72
MONTENEGRIN	0	6
POLISH	838	8842
ROMANIAN	10	287
ROMANIAN DOBREGA	0	4
ROMANIAN MANAMURESCRIANA	0	16
ROMANIAN MOLDOVA	0	10
ROMANIAN MUNTENIA	0	26
ROMANIAN TRANSILVANIA	0	21
RUSSIAN	233	2029
SERBIAN	194	118
SLOVAKIAN	10	259
SLOVENIAN	0	86
SWISS	0	12
UKRANIAN	53	1651
GREEK	1139	1326
GREEK CYPRIOT	1281	5502
ANGOLAN	0	16
BASQUE	4	255
BRAZILIAN	1	92
CASTILLIAN	0	78
CATALAN	0	192
FILIPINO	12	167
GALICIAN	0	19
GOAN	4	7
HISPANIC	269	1191
LATIN AMERICAN	11	226
PORTUGUESE	1145	2147
SPANISH	557	4514
INTERNATIONAL	4104	259
JAPANESE	567	1536
ARMENIAN	148	669
JEWISH	1249	1417
JEWISH AND ARMENIAN	141	0
SEPHARDIC JEWISH	0	6
AFGHANISTANI	0	59
ALGERIAN	49	164
BALKAN MUSLIM	0	3
BANGLADESHI	5831	2069
EGYPTIAN	0	32
ERITREAN	59	120
IRANIAN	367	754
IRAQI	48	32
JORDANIAN	0	5
KAZAKHSTANI	0	3
LEBANESE	6	188
LIBYAN	0	7
MALAYSIAN MUSLIM	0	12
MOROCCAN	0	58
MUSLIM	12952	9634
MUSLIM INDIAN	1062	643
MUSLIM MIDDLE EAST	133	0
PAKISTANI	8381	5190
PAKISTANI KASHMIR	2001	1491
SAUDI ARABIAN	0	9
SOMALIAN	1336	1052
SUDANESE	0	29
SYRIAN	0	8
TUNISIAN	0	8
TURKISH	1437	2319
TURKISH CYPRIOT	0	22
WEST AFRICAN MUSLIM	26	95
DANISH	235	697
FINNISH	209	1457
ICELAND	5	25
NORDIC	126	149
NORWEGIAN	13	835
SWEDISH	182	631
SIKH	7288	3533
ASIAN CARIBBEAN	3	30
BANGLADESHI HINDI	23	119
GUYANESE	0	4
HINDI NOT INDIAN	898	1543
INDIA NORTH	2398	1939
INDIAN HINDI	6251	2637
INDIAN SOUTH	8	27
KENYAN ASIAN	0	17
NEPALESE	28	1
SEYCHELLOIS	2	4
SOUTH ASIAN	8578	1048
SRI LANKAN	3172	2149
UNCLASSIFIED	310	8897
VOID	238	541
VOID - FORENAME	0	246
VOID - SURNAME	185	0
VOID INITIAL	444	18
VOID OTHER	6	0
VOID TITLE	62	1

Table 5.2 Summary of the heuristic classification's results by CEL Type

5.6. Conclusion

This chapter has offered an account of the exploratory processes developed to classify the names with a frequency of three or greater in the 2004 GB Electoral Roll (representing 281,422 surnames and 114,169 forenames). This heuristic approach has classified different groups of names into CELs following the techniques described in Chapter 4, through different stages and rules described here in the order in which they were first investigated. These techniques were specified through different rules and applied to different stages in a dynamic and iterative process.

At the end of this process there was a double sentiment of achievement and pessimism. On the one hand, the fact that the objective to classify the most frequent names in Britain, had been achieved, brought with it a high degree of optimism in the potential applicability of the results. However, on the other hand, it was clear that the process to get there had been very convoluted, painstakingly long, and full of manual interventions which were difficult to recall and explain, hence opening up a long list of limitations with the heuristic approach taken. The main problem with this approach was that its results were not easily reproducible, due to the ad-hoc considerations that had accumulated through the classification process. In order to evaluate the experience gained through the heuristic approach, a thorough evaluation of its issues was carried out. This evaluation is presented in the next chapter, forming the basis for building a new automated and integrated approach, which aims to overcome this limitations and learn from the experience gained through this exploratory phase.

Chapter 6. An Automated and Integrated Approach to Name Classification

The previous two chapters described the outcome of what in essence was a thorough learning exercise in building a new name-based ethnicity classification, identifying the most appropriate data and techniques (described in Chapter 4) and applying them through different steps and rules following a heuristic approach (described in Chapter 5). Through these steps, set in the context of the specialised literature (described in Chapter 3), valuable experience was acquired and a substantive body of knowledge was developed during this research. At the end of this first phase of the research, the objective of building a classification of most of the forenames and surnames present in Britain into the kinds of cultural, ethnic and linguistic groups described in Chapter 4 was achieved. As such, this work established an important provisional ethnicity classification that successfully covered a high number of names.

However, on reflection, the way this objective was achieved entailed work practices that were not always wholly systematic, and a resulting classification that could not be documented in a wholly transparent manner. This seriously limits the scientific reproducibility of the research, and hence some of its potential usefulness to the scientific and policy communities – particularly given that, as reviewed in Chapter 2, ethnicity classifications are a contested and sensitive process from the standpoint of government and policy. Although the outcome of this work was a functioning classification, its future adoption and improvements would be substantially hindered by the inability to explain in a systematic and transparent way how it was built, in order for others to evaluate it, reproduce it and enhance it.

This chapter begins by analysing the limitations recognised in the heuristic approach presented in the previous chapter, and sets out the requirement to develop a transparent and reproducible method. Both its limitations and achievements are taken into account in designing a new strategy to produce an enhanced methodology that builds upon the strengths identified in the previous chapters, while overcoming most of the previous limitations. This chapter will describe how such an enhanced methodology was developed in what constitutes an automated approach to classification of names into CELs. This automated approach has the objective of providing a simple and systematic method that can be easily explained and understood, and allows third parties to understand the explicit procedures that were used to develop the classification.

The chapter is structured in five sections. Section 6.1 summarises the practical limitations of the heuristic classification in ten major points, evaluating it against the basic scientific test of reproducibility. Section 6.2 explores the possible routes to building a new an automated classification, using the strengths of the techniques developed in the heuristic approach and introducing enhancements to the CEL taxonomy and the forename-surname clustering (FSC) technique. Section 6.3 describes the first part of the automated approach, specifically the development of a seed list of forenames that will form the input to the second part, while Section 6.4 describes this second part, in which two main cycles of the FSC technique are used to assign surnames and forenames to CELs. Finally, Section 6.5 reflects upon the achievements and limitations of the automated approach presented here and points out avenues for further research in this area.

6.1. Practical Limitations of the Heuristic Approach

The heuristic approach described in the previous chapter achieved the objective of classifying into cultural, ethnic and linguistic groups all forenames and surnames with a frequency of three or greater in Britain, subject to a series of obvious limitations. These will be discussed in this section with a view to informing the development of an automated and integrated approach in the rest of the chapter.

6.1.1. Simplicity and reproducibility

As described in the introduction to the previous chapter, the heuristic approach to classifying names was based on a collection of rules and practices that aimed progressively to proceed to a solution by trial and error as more data, knowledge and experience became available to the research. As such, it was at the same time both a learning and an exploratory exercise. It was a major classificatory effort, combining the investigation of techniques described in Chapter 4 with their application to different stages in a dynamic process. After completing and attempting to document the heuristic phase of the classification, it became obvious that the approach had a number of very serious flaws, and that the cumulative and ad hoc application of heuristics amounted to a failure in terms of a basic tenet of scientific inquiry; namely that the derivation of results should be independent of the experimenter and the method should be reproducible by other researchers (Longley and Goodchild, 2008).

Kuhn (1977) identified the five characteristics of a good scientific theory; accuracy, consistency, scope, simplicity, and fruitfulness. Although this PhD thesis does not attempt to propose a new scientific theory, it does intend to propose a new ontology of ethnicity based on personal names for other researchers to use and expand, and

therefore its methodology should comply with the minimum standards of the scientific method. It is useful here to compare its achievements against Kunh's (1977) five characteristics of good scientific theory:

> *What, I ask to begin with, are the characteristics of a good scientific theory? (…) First, a theory should be* ***accurate****: within its domain, that is, consequences deducible from a theory should be in demonstrated agreement with the results of existing experiments and observations. Second, a theory should be* ***consistent****, not only internally or with itself, but also with other currently accepted theories applicable to related aspects of nature. Third, it should have broad* ***scope****: in particular, a theory's consequences should extend far beyond the particular observations, laws, or subtheories it was initially designed to explain. Fourth, and closely related, it should be* ***simple****, bringing order to phenomena that in its absence would be individually isolated and, as a set, confused. Fifth (…) a theory should be* ***fruitful*** *of new research findings: it should, that is, disclose new phenomena or previously unnoted relationships among those already known. These five characteristics - accuracy, consistency, scope, simplicity, and fruitfulness - are all standard criteria for evaluating the adequacy of a theory.*
>
> (Kuhn, 1977: 321, emphasis not in original)

Making the necessary translation from theory to methodology, the heuristic CEL name classification as a method to develop an ontology of ethnicity based on names, does not meet all of these characteristics. It seems to meet the requirements of fruitfulness and scope, and possibly of accuracy as well, within a rather narrow

domain, but it performs badly against the internal consistency aspects, and it definitely does not comply at all with the simplicity requirement.

Furthermore, taken as a test or an experiment, the methodological approach described in the previous chapter does not comply with a commonly agreed principle in the scientific method of inquiry; an experiment's ability to be accurately reproduced, or replicated, by someone else working independently. Therefore, lack of reproducibility, internal consistency and simplicity are the three main weaknesses of the heuristic approach described in the previous chapter.

6.1.2. Ten major limitations of the heuristic approach

The weaknesses pointed out above relate to a series of problems with the heuristic approach that are explained here, and summarised as ten apparent limitations.

a) *Moving target.* The heuristic approach first aimed to classify the top 25,630 surnames into CELs, before moving on to classify the top 30,000 forenames and finally the all other names with a frequency greater than three in the GB04 file, to reach a combined total of 281,422 surnames and 114,169 forenames. As such, the objectives and methodological considerations changed during the process. Furthermore, as the number and type of names increased, so the number and detail of CELs was also expanded, from an initial 12 broad groups to around 50 CEL types which were then further split up into a total of 185 CEL types. Therefore the resulting method suffers from a moving target problem, closely linked to the following data availability limitation.

b) *Progressive changes in data availability*. As explained in the previous chapter, the detailed GB04 dataset including forenames, surnames and postcode unit at the individual level only became available half way through the research. Because of this, the top 25,630 surnames in Tier 1 had to be classified without any forename information, and thus without using the FSC technique, and without the fine geographic detail that allowed its linkage to Census data. Had the GB04 dataset been available from the start, and the research had, as a consequence, set the ambitious objective of classifying such a large number of surnames and forenames, other, simpler methodological paths would have been followed. Such paths will be pursued later in this chapter.

c) *Lack of pre-conceived notions on optimal name classification methods*. The literature available to tackle similar problems is relatively small in size and rather obscure in character. Therefore, the heuristic approach was developed through a lengthy and cumulative learning process, in order to understand how to best approach the classification of names and to pre-evaluate the efficiency of the different techniques. This fact, together with the multidisciplinary nature of the topic, and the large sizes of datasets to be handled, made the process very slow and its learning curve very steep. After having used all the methods described in Section 4.3, it can be concluded with hindsight that the approach adopted was not necessarily the most effective.

d) *Arbitrariness in the sequencing of rules*. The sequence in which the techniques and rules were implemented, as portrayed in the flow charts shown in figures 5.1, 5.2 and 5.4, was guided essentially by pragmatic

considerations, and not necessarily in the most logical order when taken as a whole.

e) *Ad-hoc variables, thresholds and decisions*. The amount of ad-hoc information used (e.g. knowledge that Jewish names are more concentrated in postal area 'NW' in London, and the State of New York in the US, but if a particular name was concentrated in a rural area then it would not be Jewish), and the arbitrariness of decisions invoked and thresholds applied to classify certain names at each of the steps shown in figures 5.1, 5.2 and 5.4, was very difficult to systematise. This was in part because of the rather subjective and partial local knowledge of geographic concentrations in Britain and internationally that was gathered during the research process. Moreover, this was also because of the number of thresholds applied to filter names in each of the 23 rules described in Chapter 5 (A.1 to A.8, B.1 to B.11 and C.a to C.d). Such thresholds were introduced to select or reject a name for subsequent assignment to a CEL within a rule, for example to decide when a name was highly concentrated in an area, deemed to be British or non-British, or associated to a particular CEL. Furthermore, failure to adopt in an early stage a systematic approach to documenting each decision taken, made this ad-hoc knowledge even more obscure when it came to write-up the methodology.

f) *Ecological Fallacy*. Many of the early decisions on surname to CEL assignment were based on geographic concentrations in Britain and internationally, and assumptions about the types of areas where they are most frequent, as explained in Section 4.3.2 and 4.3.3. This method could have led to the incorrect allocation of surnames to a certain CEL because of the

ecological fallacy, defined as 'the false assumption that knowledge of the general characteristics of a neighbourhood will always yield accurate and precise information about specific individuals' (Harris et al, 2005: 33). Furthermore, the decision to use the *Mosaic* geodemographic classification rather than just the raw Census ethnicity data at Output Area level, contributed to this ecological fallacy by confounding the association between ethnic groups and residential areas at finer postcode units with other socio-economic profiling.

g) *Number of exceptions*. A number of exceptions to the rules described in the previous chapter were found, but usually there were no general patterns and each of them was dealt with individually. As a result, they were too numerous and complex as to be explained as part of the methodology. Many of them arose when one rule contradicted a previous or posterior rule, or when the manual check performed by experts led to reallocation of a name to a different CEL.

h) *Variability of iterations*. The classification of Tiers 2 and 3 made ample use of the Forename-Surname Clustering (FSC) technique, and different iterations of the technique were run as the number of forenames and surnames classified by CEL grew or name to CEL assignments changed. There are obvious circular implications in this procedure, that is, the same results would have not been obtained for the same names in two different iterations of FSC, since the cross-occurrences between forenames and surnames within a CEL would have changed at each step.

i) *CEL overlap and inconsistencies with the main classification*. The numerous steps and rules applied through the heuristic process led to some names

having contradicting CELs depending on the sequence in which certain rules were applied or sets of names used within a rule. The text mining technique produced many of these potential CEL overlaps. A name was reallocated to a new CEL if after a quick check through the FSC technique or with manual expert knowledge, it was judged that the name would be better moved between CELs and that the CEL would also benefit over-all from the change. This involved many individual decisions that were not documented in the process.

j) *Manual check and reclassification*. This problem is similar to the previous one, but relates to the process by which many different people with knowledge in each language culture were consulted at the end of each phase. Most of their ad-hoc decisions were taken as valid if the final results did not contradict the FSC technique outcomes. However, their re-allocations of names to a different CEL were not justified systematically, and no log of changes was kept for subsequent evaluation of the rule or step that might have created an incorrect classification in the first place. However, most of the manual reallocations occurred between similar CEL Types, always within the same CEL Groups, and in fact it was in this phase that many of the finer CEL Types were actually subdivided (such as, for example, the regional Romanian CEL Types).

Together, these ten issues prevented the future reproducibility of the methodology, led to a lack of internal consistency in the approach and made it very difficult to explain the method in a straightforward and transparent way. Drawing on them, it was clear that in order to be able to publish the results of the new classification, and

promote its adoption amongst researchers, a new approach was required. In order to be accommodated within the research resources of a single Ph.D. thesis, it was recognised that such an approach would have to develop a much simpler, holistic and automated strategy, in a single major phase or step, with minimal manual intervention, and through the imposition of a simple set of rules. This was the objective of the automated approach described in the rest of this chapter.

6.2. Exploring Alternative Automated Approaches

'Simpler is better' (Popular saying)

After going through the learning exercise of identifying the data and techniques described in Chapter 4 and applying them through different steps and rules in the heuristic approach described in Chapter 5, a great deal was learnt about the effectiveness of such classification techniques and their usefulness alongside name datasets. The limitations found in the heuristic approach were summarised in the previous section, in order to justify the need for a simpler and reproducible method. However, there were obviously a large number of benefits associated with the heuristic approach, and it did achieve the objective of classifying all of the most frequently occurring names in Britain. These benefits were implicit in the description of the techniques and rules given in the two previous chapters and will not be repeated here. Nevertheless, the overall positive conclusions about these techniques and rules will be summarised here to offer light and set a starting point in the investigations to tackle the problem of how to build a new automated and simpler approach to classify names into cultural, ethnic and linguistic groups of origin.

6.2.1. Coarser CEL Subgroups

One of the problems with the heuristic approach was the proliferation of CEL Types, resulting in an unwieldy classification overall. The process of developing the CEL classification became a major taxonomic exercise requiring ever more specialised decisions, as part of a continuous effort to further subdivide CELs into smaller groups. New CEL Types were opened up after using text mining, receiving new data on new names or countries of birth, or through the manual revision of lists using the knowledge of students from different countries. At the end of the heuristic approach, there were 185 CEL Types, a complete list of which appears in the Appendix 3.

It is worth remembering that the ontology of ethnicity proposed in this research was designed for Great Britain. This is an important point since it is relevant to the structure, size and composition of the CEL categories, and the 185 CEL Types exceeded the number of groups that were actually relevant for the GB context. As can be seen from the list of CEL Types in the Appendix 3, when the final heuristic classification was applied to the whole GB Electoral Register (GB 04 file) the reality was that a large number of these very small CEL Types, such as 'Kyrgyzstan', 'Namibia', 'Bhutan' or 'Madagascar' had fewer than 10 individuals within them in Great Britain, according to the GB04 dataset. Such CEL Types often only had surnames associated with them (SCEL), and no forenames (FCEL). Consequently, when compiling the person's overall CEL, they would be more likely to be assigned to the FCEL Type if it had a higher score than the SCEL (the way the personal allocation system works is explained in Chapter 7). Out of the 185 CEL Types, 21 had fewer than 10 people, 49 less than 100 people, and 93 less than 1,000 people in

the GB04 Electoral Register. Therefore, when analysing the results of the classification, or devising the applications described in Chapter 8, the large number of very small groups was unwieldy (not least when mapping the CEL Types). Moreover, when trying to triangulate between forenames and surnames through FSC, in order to build the automated classification discussed here, these very small CEL types lead to a very large number of dead-ends, where insufficient forename-surname connexions were found within the same CEL Type, and indeed sometimes none at all.

Therefore, it became obvious that a new set of coarser CELs would be needed for the FSC technique to be more effective when applied from scratch in the automated approach. The groupings of CEL Types that were used before in the heuristic approach were too coarse for this purpose, with only 16 CEL Groups. Hence the structure of the CEL taxonomy was rather imbalanced, with a top level in the hierarchy of 16 CEL Groups subdivided into 185 CEL Types at the bottom level. It was thus considered advisable to devise an intermediate level, grouping the CEL Types into more manageable units but still retaining sufficient variety as to retain the usefulness of the name analysis method.

This could be easily done by merging smaller CEL Types until a minimum 'population threshold' could be met, so that all CEL Types would all have a minimum size. This was of course only possible in a second phase, once the heuristic classification had been built and the whole Electoral Register could be classified. It was decided to build aggregations of CEL Types of a minimum population size of 1,000 people, which initially reduced the taxonomy from 185 CEL Types to 92

aggregations. These aggregations were called CEL Subgroups, since they were in fact a subdivision of the top 16 CEL Groups. Further merging of CEL Types was required of CEL Types if, although having more than 1,000 people assigned to them, they were deemed culturally too close to others as to present a distinguishable pattern of forename-surname clustering. Examples of these CEL Types were the Spanish ones; Castilian, Galician, Catalan, Basque, Latin-American, Philippino and Spanish Other, which were all merged into a single Spanish CEL Subgroup. Other examples were the Romanian CEL Types, the Greek and Greek Cypriot, North African CEL Types, the different 'Void' names, and so on.

A total of 66 CEL Subgroups were finally compiled, which all had sufficient internal consistency as to make the automated classification using FSC more successful and effective for the Great Britain application. A list of these 66 CEL Subgroups is included in Table 6.1 alongside the CEL Group they belong to, and its total frequency in the GB 2004 Electoral Register, as a result of aggregating the individual counts per CEL Type. A full mapping between the original 185 CEL Types and the final 66 CEL Subgroups is offered in Appendix 3 within the complete list of CEL Types (CEL Subgroup column). Henceforth the term CEL Subgroup will refer to these 66 units as a middle layer between the 185 fine CEL Types and the 16 coarse CEL Groups. When the term CEL is used on its own it will usually refer to any of these three levels, but in the context of this chapter the terms CEL and CEL Subgroups are sometimes used interchangeably, for ease of reading.

It is important at this stage to adopt a convention regarding the terms used to describe the number of names and their frequencies, in order to make the definitions more

explicit. A commonly accepted convention in the linguistics, metaphysics, statistics and infometrics literature, is the definition of the 'type-token' problem. This refers to a distinction, first proposed by the philosopher Charles Peirce (1839-1914), between signs considered as abstract things (types) or as particular instances (tokens), as that 'the class of all tokens of a given world is called a type' (Hartshorne and Weiss, 1931: 4537 cited in Burks, 1949: 681). Applications of the type-token problem usually refer to words, where one text will have a number of unique words (types) and a much larger number of instances of them (tokens). Applied to names, a unique instance of a name is a type (e.g. Pablo) of which there are a number of tokens in a given population or context (379 tokens of Pablo in GB04). Therefore, in the Great Britain Electoral Register (GB04) there are 46 million tokens of both forenames and surnames, but there are 437,639 forename types and 992,603 surname types. The type-token distinction is important especially when explaining the combinations between forenames and surnames and their cross-occurrences in the next sections.

CEL Group	CEL Subgroup	Frequency
AFRICAN	AFRICAN	7,842
AFRICAN	BLACK SOUTHERN AFRICA	5,198
AFRICAN	CONGOLESE	1,164
AFRICAN	ETHIOPIAN	1,238
AFRICAN	GHANAIAN	46,095
AFRICAN	NIGERIAN	88,243
AFRICAN	SIERRA LEONEAN	6,155
AFRICAN	UGANDAN	1,018
CELTIC	IRISH	3,442,517
CELTIC	SCOTTISH	4,749,864
CELTIC	WELSH	3,065,041
EAST ASIAN	CHINESE	24,423
EAST ASIAN	EAST ASIAN	3,997
EAST ASIAN	HONG KONGESE	119,566
EAST ASIAN	KOREAN	2,315
EAST ASIAN	MALAYSIA	2,092
EAST ASIAN	VIETNAMESE	15,723
ENGLISH	BLACK CARIBBEAN	23,665
ENGLISH	ENGLISH	31,258,100
EUROPEAN	AFRIKAANS	7,805
EUROPEAN	ALBANIAN	3,440
EUROPEAN	BALKAN	24,721
EUROPEAN	BALTIC	4,127
EUROPEAN	CZECH & SLOVAKIAN	4,881
EUROPEAN	DUTCH	24,912
EUROPEAN	EUROPEAN OTHER	31,341
EUROPEAN	FRENCH	128,129
EUROPEAN	GERMAN	129,318
EUROPEAN	HUNGARIAN	11,768
EUROPEAN	ITALIAN	229,931
EUROPEAN	MUSLIM	1,034
EUROPEAN	POLISH	155,743
EUROPEAN	ROMANIAN	2,531
EUROPEAN	RUSSIAN	11,342
EUROPEAN	UKRANIAN	3,948
GREEK	GREEK	109,370
HISPANIC	PORTUGUESE	90,327
HISPANIC	SPANISH	107,843
INTERNATIONAL	INTERNATIONAL	15,799
JAPANESE	JAPANESE	6,335
JEWISH AND ARMENIAN	ARMENIAN	4,353
JEWISH AND ARMENIAN	JEWISH	81,415
MUSLIM	BANGLADESHI	179,401
MUSLIM	ERITREAN	1,397
MUSLIM	IRANIAN	10,312
MUSLIM	LEBANESE	3,107
MUSLIM	MUSLIM	6,808
MUSLIM	MUSLIM MIDDLE EAST	104,859
MUSLIM	MUSLIM NORTHAFRICAN	3,713
MUSLIM	MUSLIM SOUTH ASIAN	25,704
MUSLIM	PAKISTANI	508,699
MUSLIM	PAKISTANI KASHMIR	91,472
MUSLIM	SOMALIAN	33,260
MUSLIM	TURKISH	51,911
NORDIC	DANISH	20,561
NORDIC	FINNISH	5,685
NORDIC	NORDIC	6,492
NORDIC	NORWEGIAN	186,375
NORDIC	SWEDISH	19,090
SIKH	SIKH	283,657
SOUTH ASIAN	HINDI INDIAN	319,979
SOUTH ASIAN	HINDI NOT INDIAN	25,080
SOUTH ASIAN	INDIA NORTH	75,282
SOUTH ASIAN	SOUTH ASIAN OTHER	15,536
SOUTH ASIAN	SRI LANKAN	53,919
VOID	VOID	246,811
TOTAL		**43,639,2**

Table 6.1: List of 66 CEL Subgroups
The list includes the 66 CEL Subgroups as a result of aggregations of finer CEL Types. For each CEL Subgroup its overarching CEL Group is given alongside its total frequency in the GB 2004 Electoral Register, as a result of aggregating the individual counts per CEL Type. For a full list of correspondence between 185 CEL Types and 66 CEL Subgroups see full CEL Types list in Appendix 3.

6.2.2. Positive aspects of the seven classification techniques used in the heuristic approach

Of the seven classification techniques explained in Section 4.3 there was one whose efficiency stood out from the rest in terms of simplicity and classificatory power. This was the Forename - Surname Clustering (FSC) technique, based on Tucker's (2003; 2005) CELG method. Because of this, the main focus of the exploratory phase of alternative methods discussed in this section will be the FSC technique and how to adapt it to contribute to an optimal automated classification. After reviewing the role of the other techniques the chapter will focus primarily on development and application of FSC.

Despite the weight and importance of the FSC technique, its success would have not been possible without leveraging on the results of two other techniques, originally used to form the initial CEL clusters of surnames, and later exploited by FSC to triangulate between surname CEL to forename CELs and then back again to surname CEL (depicted as; SCEL=> FCEL => SCEL and so on). The two techniques that substantially contributed to initial sorting of surnames into CELs were spatiotemporal and geodemographic analyses.

Aspects of spatiotemporal change, when comparing the different name distributions between 1881 and 1998, were crucial to determine whether the origin of a surname was British or non-British, and to identify regions of origin within Britain. Regarding spatial and geodemographic analysis aspects, geographical concentration of relative name frequency at different scales (locally, regionally, or internationally), helped to determine the most probable CEL of a surname's origin, albeit while having to rely

on obscure local knowledge, as pointed out in the previous section under the limitations of this method. Furthermore, these two techniques were always useful, even after FSC was applied to Tier 2 and Tier 3, to distinguish British from non-British name origins. This helped to improve FSC's reliability since it made it possible to remove the substantial overlap occurring between British forenames and non-British surnames as cultures adopt the naming practices of the 'host society'. This fact also led Tucker (2003, 2005) to ignore the percentage of 'English' forenames when applying his CELG technique to DAFN.

The rest of the techniques described in Section 4.3 – text mining, name to ethnicity data, analysis of lists of international name frequencies and genealogical resources, and researching individual names – were most useful in seeding new CEL Types, splitting up coarse ones (such as Greek into Greek and Greek Cypriot), or classifying rarer names (typically with frequencies below 10). FSC does not work so well with rarer names since it does not pick up enough significant differences in the cross-occurrences between FCELs and SCELs (i.e. most cross-occurrences were evenly spread between CELs or had just one or two occurrences per CEL), or because of small numbers it was easily influenced by exceptions in the population (e.g. a single family with a rare Italian surname that might have given all their children Greek forenames). These types of situations were picked up well by the other marginal techniques, complementing the FSC technique and spatiotemporal analysis.

6.2.3. Benefits and limitations of Forename-Surname Clustering (FSC)

As shown in the previous chapter, the FSC technique allowed the quasi-automatic classification of thousands of names in Tiers 2 and especially in Tier 3, in which 255,792 surname types and 84,192 forename types were classified into CELs,

respectively representing 91% and 74% of the final total number of names classified. What is more, almost all of the 114,169 forename types were solely classified using FSC, once they were linked with previously classified surnames. The power of this technique demonstrates the vigour with which distinctive forename-surname practices are still unique within CELs in the contemporary population of Britain.

The detailed explanation of the FSC technique was given in Section 4.3.1, and its application to classify names in Tier 2 and 3 was respectively described in sections 5.3.2 and 5.4.1. Therefore, repetition of the 'ins and outs' of the FSC technique is avoided, and only the enhancements to the version already described and applied in previous chapters will be offered here. The objective of this subsection is therefore to explore potential enhancements to the FSC technique so that it may be used on its own to classify names from scratch in a new automated approach.

As previously mentioned, FSC is based on two assumptions; (a) that most cultural, ethnic and linguistic (CEL) groups adopt distinct forename naming practices that are passed from one generation to the next, even if they are settled immigrants; and (b) that most people tend to marry or procreate with members of the same CEL group. If proved true, these two assumptions would preserve the unique cultural, ethnic and linguistic character of forenames within CEL groups, in the same way that surnames are more or less preserved intact by the strict rules that govern naming practices in civil registration of births in most countries. Therefore, two factors are key to the preservation of a cultural link between forenames and surnames; continuation of CEL group's naming tradition and a high degree of within group marriages or procreation. Conversely, the forces eroding such assumed linkages are cultural and

linguistic integration in the 'host society', and intermarriage between ethnic groups, both of which are key indicators of the so called 'assimilation' of ethnic minorities. Both group-preserving and group-dispersion forces are highly shaped by spatial processes (Peach, 1980).

Most of the misclassification problems found in the application of the FSC technique arose because of high overlap between a CEL's and host society's forenaming practices. This was especially acute for groups with a relatively long history of migration and which have now become culturally integrated; in Britain typically pre-twentieth century migrants such as Jewish, Huguenots or Irish, but increasingly also early twentieth century migrants, such as Italians, Polish, and Indians. A second problem with FSC arose when considering names with very low frequencies of occurrence in Britain (less than ten or even less than five), since their small numbers made them very sensitive to incorrect assignment of CEL based on the forenames of sometimes just one or two individuals or families.

Despite these two potential misclassification problems, which will be dealt with later, the major limitation of the FSC technique, in its current form previously applied in Chapter 5 as well as by Tucker (2003) in the DAFN, is that it requires pre-classification of a large list of names in order to initiate the triangulation between forenames and surnames. Tucker (2003) used a list of 80,000 forenames manually coded by CEL, gender and with a flag indicating whether they were deemed to be diagnostic of the CEL or not. This was only possible through the expert knowledge of the dictionary's editor, Patrick Hanks, to code an initial set of 3,000 forenames and then expand it through manual revisions and FSC to 80,000 (Hanks and Tucker,

2000). The application of FSC in the heuristic approach described in the previous chapter used a list of 25,630 surnames previously classified by CEL in Tier 1 using the techniques and rules described in Section 5.2. These two types of 'ignition list' of names will be termed here the 'seed names list'.

Furthermore, both the DAFN and the heuristic approach required a good starting knowledge of the 'host society's' names in order to treat them separately in the FSC technique computations, and hence account for the observed overlaps in naming practices. Such 'host society' was considered here as 'British Isles's names', which included English, Cornish, Welsh, Scottish, and Irish names found in the Great Britain 2004 Electoral Register. In the DAFN example, a similar definition of 'Anglo-saxon names' was taken as the 'default CEL' for the US population. This list will be called here 'host names list'.

Taken together the main obstacles to automated classification are: FSC misclassifications arising out of overlaps in naming practices, intermarriage between CEL groups and low frequencies of occurrence; and the need to provide seed and host names lists to initiate the triangulation. If most of the effects of these issues can be overcome, the FSC technique offers great potential to classify all of the names using an automated approach, and can be observed to be based on a single set of simple and transparent rules rather than the longwinded version of the heuristic approach. In achieving such an objective, FSC would play the core role, the details of which will be described in the next section.

6.3. Building a Forename Seed List

6.3.1. Alternative options to the 'seed' and 'host' name lists

Having a good quality set of 'seed' and 'host' names lists is an inherent requirement for the successful classification of names into CELs using FSC, as pointed out in Section 6.2.3. Alternative approaches to eliminate the need for such seed and host names lists in the application of the FSC technique were initially explored but discarded in this research because of technical constraints, as will be explained later in Section 6.5 when describing enhancements that were attempted but subsequently abandoned. Therefore, at this stage, the objective was to develop a new set of seed and host lists in the most automated way possible in order to facilitate the initiation a new classification and its reproducibility by third parties.

It was decided to develop a seed names list based on forenames rather than surnames, following the DAFN example. This is because, as suggested above, the frequency distribution of forenames in most countries has been proven more positively skewed than that of surnames (see Figure 5.3), or in other words, that a substantially smaller number of forenames than surnames is required to classify the same population (e.g. 95% of the population –tokens– in the US are covered by just 14% of surname types, while only 1% of forename types suffice to cover the same population. See Section 5.4.1 for more detail). Examples of this asymmetry have been found in the UK (Tucker, 2004a), Ireland (Tucker, 2006), US, (Tucker, 2001), Canada (Tucker, 2002), and Spain (Mateos, 2007; Mateos and Tucker, 2008). Therefore, using the information gathered through the heuristic approach and Great Britain's Electoral Register (GB04) it was estimated that with an initial seed list of say the most

frequent 20,000 non-British forename types one could classify at least 90,000 non-British surname types in Britain.

Several ideas were initially followed up to develop a comprehensive initial seed list of forenames. One first idea was to collect lists of forenames from different publications or the increasingly popular government statistics on forenames, available in Belgium (Statbel, 2006), Denmark (Danmarks Statistik, 2006), Iceland (Statistics Iceland, 2006), Madrid and Catalonia in Spain (IDESCAT, 2006; Instituto de Estadística de la Comunidad de Madrid, 2006) and Germany (Gesellschaft für deutsche Sprache, 2006). However the worldwide coverage of these lists was not large enough for the purpose stated here, and even when available their length and quality tend to vary significantly. Another idea was to set up a major survey to international students at UCL asking them to manually classify lists of forenames, picking up the forenames associated with their mother language. Preliminary exercises in this respect resulted in a high degree of overlap between similar CELs (e.g. Portuguese, Spanish and Italian forenames), the task was very tiring and time consuming, and only a sub-set of languages and nationalities that could be covered. A third approach was to build such a list from existing dictionaries of forenames (see for example, Parekh and Parekh (2003) for South Asian forenames), which might have infringed copyright as well as have taken a long time to select and digitize thousands of names from at least 50 languages. Furthermore, none of the above approaches provided a direct way of assigning a probability of a forename belonging to a CEL, or even a method of external validation of the quality of the assignment. This probability or strength of association between a name and a CEL was an important requirement, making it possible to discriminate between highly diagnostic

and low diagnostic forenames, making the FSC technique more efficient, as demonstrated by Tucker (2005).

Lacking any other feasible alternative, it was finally decided to build a seed list of forenames based on the pre-existing information gathered through the heuristic classification. This option was not ideal in that it did not meet the independence requirement set at the beginning of this chapter, viz. to build a new automated classification from scratch. However, it was recognised that this was the best resource available for the task, and that the subsequent list of seed forenames derived through this process might be made available to any researcher interested in validating this approach or developing a new one, hence meeting the requirement for reproducibility of scientific inquiry mentioned before. The construction of a forename seed list following this approach will be described in Section 6.3.2.

A straightforward approach to creating the 'host' names list was adopted, by using the spatiotemporal analysis technique and comparing the GB 1881 and GB04 registers as explained in Section 4.3.3. This entailed identifying names present today in Britain that did not appear in 1881 and which are therefore likely to be of foreign origin (or of more recent creation). Furthermore, some foreign names that were present in 1881 have experienced dramatic growth relative to the rest of the population, and hence were also identified as 'foreign' when comparing the two datasets (using a threshold of growth of over 2,000% in relative terms). Through this technique a list of 22,078 British-Irish forename types was compiled, including the following CEL Types: English, Cornish, Welsh, Scottish, Irish, Northern Irish, and 'Channel Islands'.

6.3.2. 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 FSC technique.

This was achieved through the series of steps summarised below. The full set of structured query language (SQL) queries built in the Oracle database to automate and replicate this process is included in Appendix 4. This illustrates an important feature that differentiates the automated approach presented here from the heuristic approach. These SQL queries could be applied to any other set of names and ethnicity datasets, such as patient registers with country of birth, in order to automatically build other possible forename seed lists for different populations.

What follows is a sequence of the ten steps taken to obtain a forename seed list, which is itself the main ingredient in the Forename-Surname Clustering 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 Section

6.3.1, 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 6.2, 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 6.2: 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 6.3.

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

Table 6.3: 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 6.3 (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); μ and σ 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 6.1, 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.

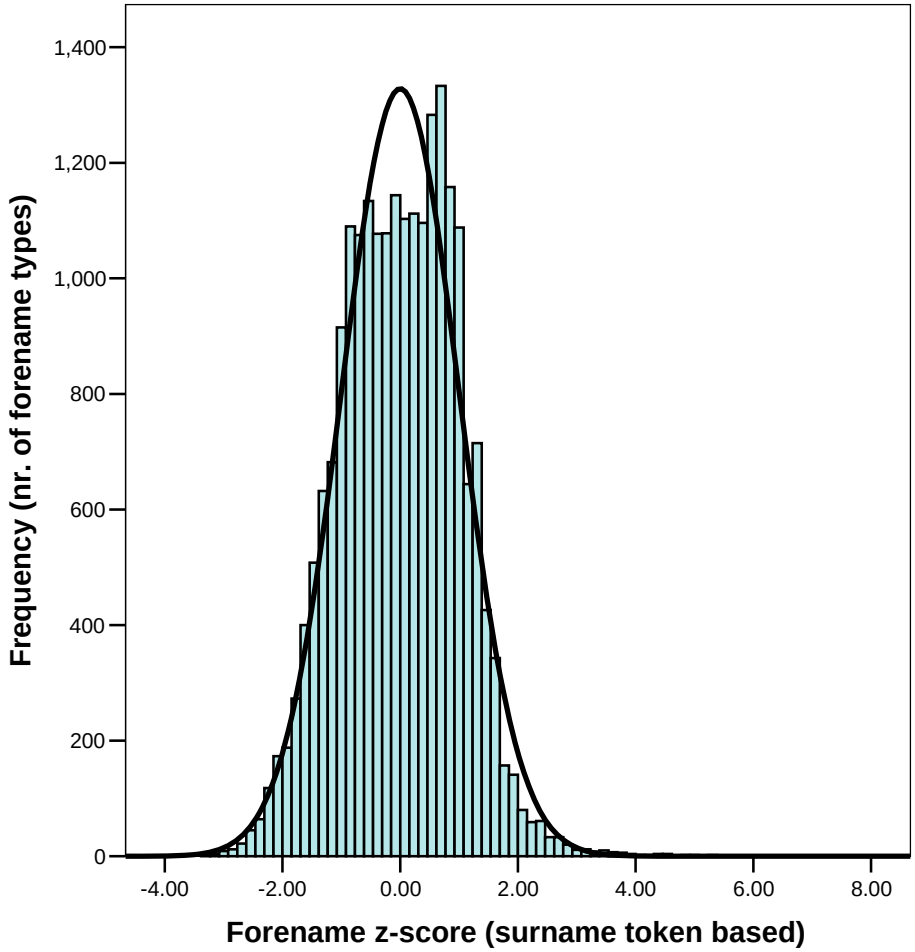

Figure 6.1: Histogram of forename z-scores distribution (based on surname tokens)
This histogram shows the forename type frequency distribution of the z-score value for each forename type (n=20,939). Those z-scores standardised by CEL Subgroup the percentage of surname tokens per forename type (see steps 6, 7 and 8 in the text and Table 6.3). 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.

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

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 6.2. 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.

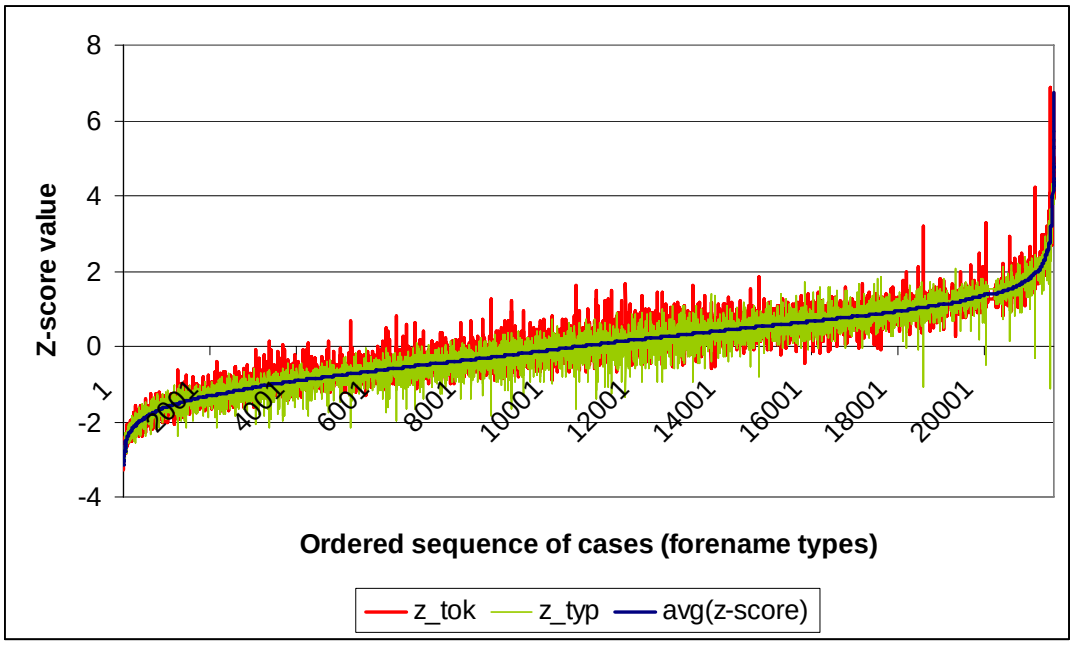

Figure 6.2: Distribution of forename z-score values based on surname tokens, types and their average
This graph shows the z-score values on the y-axis and the sequence of each of the 20,939 forename types on the x-axis ordered by ascending average z-score. The three lines show; z-score values based on 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.

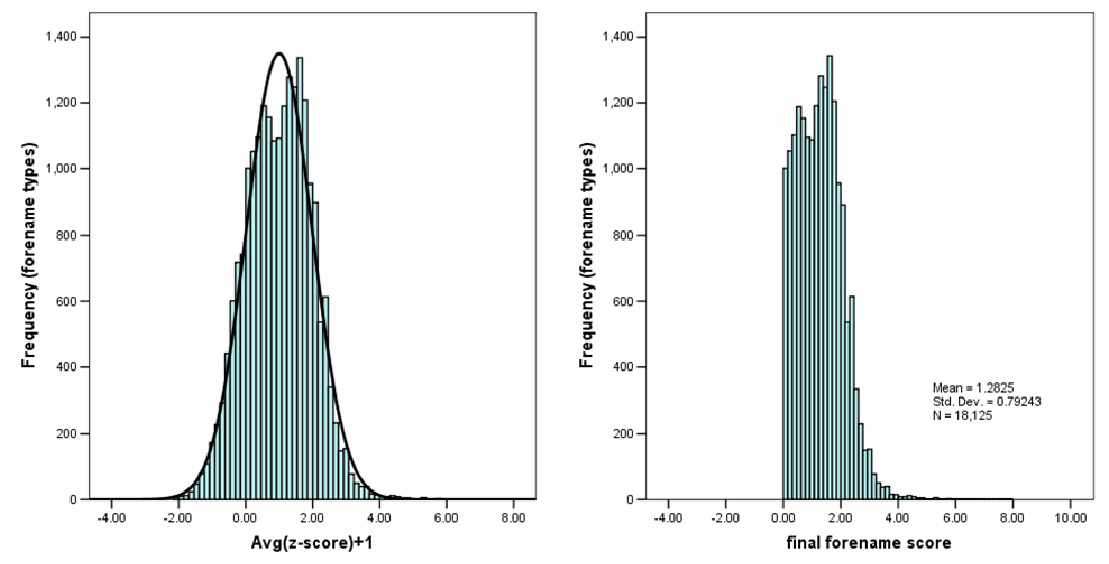

Figure 6.3: Histograms of average z-score+1 (left) and truncated final forename score (right)
Theses two histograms show the effect of the transformation introduced in step 10. In the histogram on the left, the effect of adding a value of one in the whole distribution being shifted to the right can be 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 6.3 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 6.4, 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'. The full detailed forename seed list of 41,544 forename types and their attributes is available on request to bona-fide academic researchers for further evaluation and enhancement. A sample of forename types from this list is provided in Appendix 5.

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 6.4: Summary of the contents of the final non-British forename seed list.

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.

6.4. Forename-Surname-Clustering (FSC)

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.

6.4.1. 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 6.4 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.

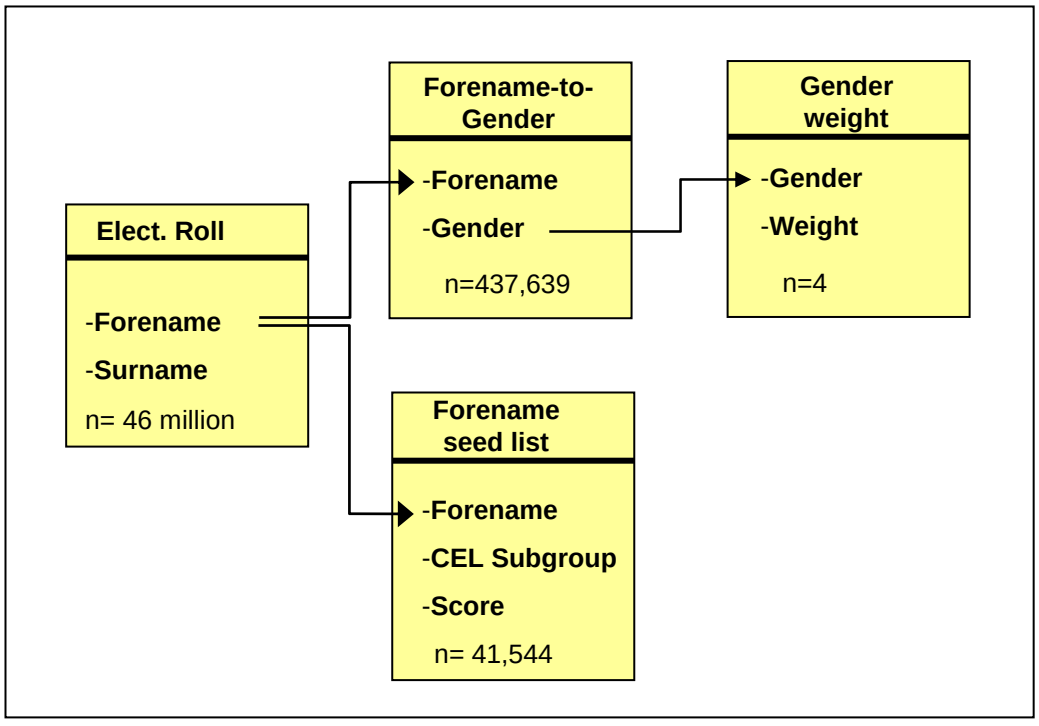

Figure 6.4: Tables relationships in cycle 1 step 1
Diagram of the relational database structure between the tables in cycle 1 step 1. The direction of the arrows represent a 'many to one' type of relationship between the tables (from arrow stem to arrow pointer). n=number of records in each table. The '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 (2007a) 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 6.4. 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 6.5

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 6.5: 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 6.4, 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 6.6.

Surname	CEL Subgroup	Forename Tokens	Forename Types	Cumulative personal score	
				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 6.6: 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 6.6), 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 6.6 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 6.6). 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 6.3.2.

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 6.6 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 6.3.2 (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.

6.4.2. 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 6.4.1. 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 6.5 shows the relationships between these three tables, which is very similar to Figure 6.4, differing only in bottom-middle table.

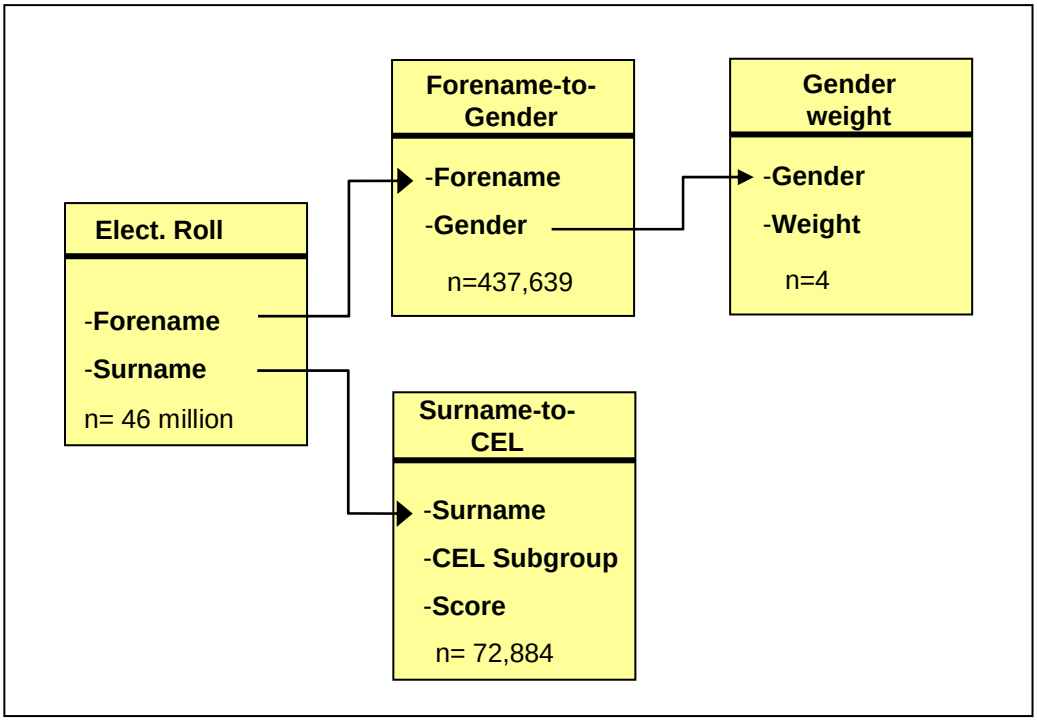

Figure 6.5: Tables relationships in cycle 2 step 1
Diagram of the relational database structure between the tables in cycle 2 step 1. The direction of the arrows represent a 'many to one' type of relationship between the tables (from arrow stem to arrow pointer). n=number of records in each table. The '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 6.5, 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.

6.4.3. 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 6.6 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.

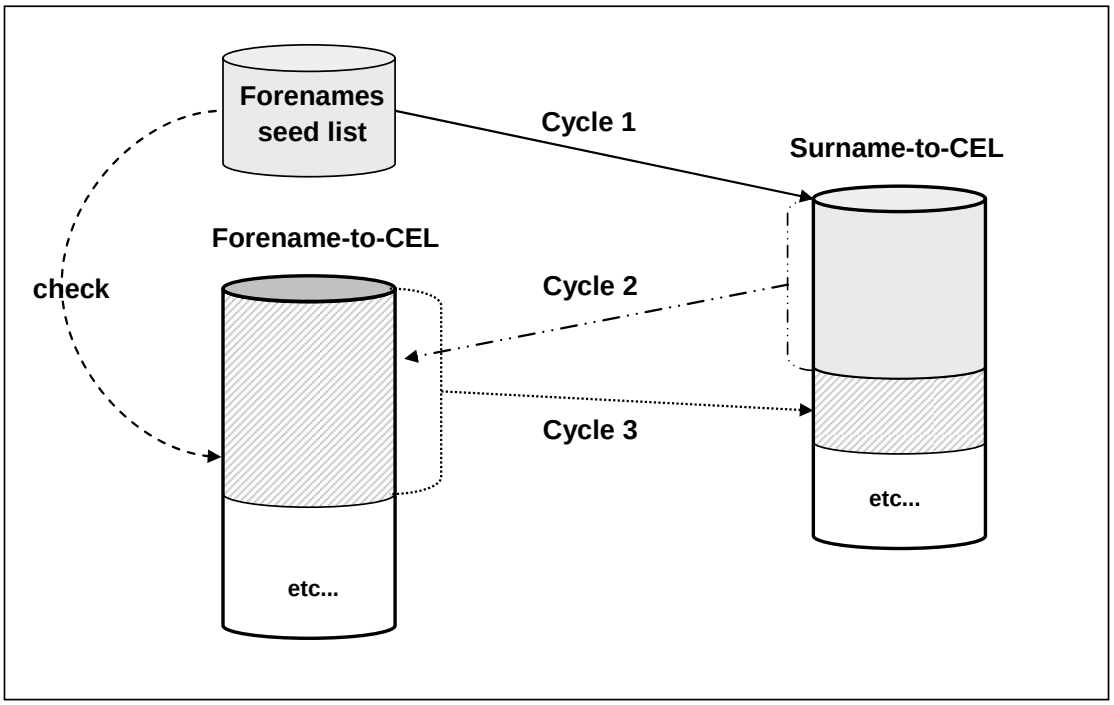

Figure 6.6: Cycles in the automated classification
This diagram shows the process flow described in Section 6.4, starting with a forenames seed list used in cycle 1 to produce a surname-to-CEL table, which is then in turn used in cycle 2 to produce a forename-to-CEL table, and so on. Only cycles 1, 2 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. These tables are available on request to bona-fide academic researchers for further evaluation and enhancement.

6.5. Enhancements to the Automated Approach

The automated approach described in this chapter presents a significant enhancement over the heuristic approach described in the previous chapter, whose limitations were listed in Section 6.1. However, the automated approach has two shortcomings. The total number of names classified is lower than in the heuristic approach, because of the high number of rarer names that the FSC technique could not classify. Furthermore, the methodology still depends on an external input; the forename seed list, pre-classified by CEL Subgroup and with accurate scores. This PhD developed such a seed list based on the previous heuristic classification, so the automated approach is still not completely independent of the previous approach. Although other researchers could also use this forenames seed list, it nevertheless remains a UK-based list and might not work as well in other contexts. Some ideas to overcome this second problem were explored and will be discussed in this section.

6.5.1. Potential enhancements that were abandoned

In order to find an alternative to the dependency on an externally and non-automatically produced forename seed list, several alternatives were explored during this PhD, but their results were deemed not sufficiently successful as to be implemented as part of the automated approach. However, there are some promising aspects that are worth mentioning here, identifying promising future research avenues in this area.

The automated approach described in this chapter relies primarily on the forename-surname clustering (FSC) technique, adapted from Tucker (2003; 2005), the assumptions of which are fully described in Chapter 4 Section 4.3.1. In this PhD and in Tucker (2003; 2005), the clustering of forenames and surnames is initially induced by a previously manually classified forename seed list that 'ignites' the FSC cycles. However, the requirement for such a seed list is an important shortcoming since it relies on the use of other less systematic techniques and subjective decisions, such as spatiotemporal analysis or text mining, or on the manual classification by experts.

In essence the FSC technique measures the correspondence between certain groups of forenames and certain groups of surnames through the identification of the highest frequencies of common bearers between the two. This problem could be represented as a matrix of *m* number of forenames as columns by *n* number of surnames as rows (*m* x *n*), and each cell would contain the counts of people that have a particular forename and surname combination. In this way the cell for column 'John' and row 'Smith' would have a count of 11,920 people in GB Electoral Register, the combination 'Pedro' and 'Garcia' 16 people, and 'Pedro' and 'Smith' none. Most of the cells in this matrix would therefore be empty, while the ones with higher frequencies would be concentrated around a few forename-surname combinations. Essentially this is a classic clustering problem in statistics, biology and social science; 'to look for systematic groups in data' (Kaufman and Rousseeuw, 2005: vii).

It is not the purpose of this thesis to describe the many different clustering techniques available, but it will suffice to mention the two main algorithms that represent the two broad types of clustering techniques: hierarchical agglomerative and iterative

relocation (Harris et al, 2005). For a comprehensive description of clustering methods see Gordon (1999). The most commonly used hierarchical agglomerative algorithm is that of Ward (1963), in which through an agglomerative or stepwise approach *n* groups each containing one object are merged together in a number of steps, using a measure of similarity or distance. At each of these steps the number of groups diminishes until all of them are finally merged into a single group containing *n* objects. The result of this classification is typically represented graphically by a dendrogram. Amongst the iterative relocation algorithms, one of the most commonly used is k-means. This is a non-parametric clustering method that creates a number of clusters (*k)* defined by the user. Its objective is to minimize the variability within the cluster by a series of iterations through which objects are moved between clusters to evaluate if the move improves the sum of squared deviations within each cluster (Aldenderfer and Blashfield, 1984).

Both hierarchical agglomerative and iterative relocation algorithms require as an input a matrix of 'distances' or (inversely) 'similarities' between the objects to be clustered, typically respectively termed the 'distance matrix' or 'similarity matrix' (Gordon, 1999). In the case discussed here, the clustering of forenames and surnames, two possible interpretations of the 'similarity matrix' could be adopted. In the first instance, a matrix can be built as the one mentioned above (an *m* x *n* matrix), measuring the frequencies of cross-occurrences between forenames (*m* columns) and surnames (*n* rows). The most 'similar' forenames and surnames will be the ones with highest cross-occurrences in the population.

A second option would entail taking just one of the two elements, for example forename types, and build a matrix of cross-occurrences between all of the different forename types (*n* forename types by *n* forename types, or *n* x *n*), their similarity being measured by the number of surname tokens that each pair of forenames types have in common. For example, the forename type pair 'Pablo – Pedro' had 204 surname tokens in common in the GB 2004 Electoral Register, while the pair 'Pablo – Maurizio' only had 7 surname tokens. The pair counts would be transformed into relative frequencies of surname tokens per forename (in percentage) for the similarity matrix. Such a matrix would be symmetrical and would have the highest measure of similarity along the principal diagonal, but this will simply measure all surname tokens associated with a single forename type - since, for example, 'Pablo' would be both the row and the column entry. Therefore, the useful part of the matrix lies on either side of the principal diagonal, indicating 'how close' is one forename to all the others in terms of their common surname tokens. Clustering just one element, such us forenames using their 'surname similarity', would produce the desired groupings of forenames as a preliminary version of the forename seed list being sought after. At that stage, the clusters will be still anonymous, that is with no CEL associated with them. The final assignment of a CEL to a cluster would have to be done by checking the possible origin of just a few names from each cluster, by using one of the non-FSC classification techniques, for example spatiotemporal analysis, text mining, or manual search.

Both of these types of matrices (*m* x *n* and *n* x *n*) and the two clustering algorithms mentioned above (Ward's hierarchical and k-means), were explored using the name datasets described in this chapter. However, the task proved to be overwhelming,

especially in technical terms because of demanding computer processing power and memory requirements and software limitations. Furthermore, it was also because of the intense efforts required to decide the number of clusters or the optimal level in the clustering hierarchy, the difficulty inherent in assigning the clusters to a specific CEL, and problems in evaluating the accuracy of such assignments.

The main problem was the technical limitations in dealing with such enormous datasets. Both types of similarity matrices contemplated here, *m* x *n* and *n* x *n*, had a dimensionality that grew exponentially with the number of name types to be clustered. For example, just using the most frequent surnames, those with a frequency of 100 or more (*n*= 28,623), and forenames, of 25 or more (*m*= 19,138), we would have a matrix of 28,623 x 19,138, that is = $5.48*10^8$ possible combinations. If all the names in the Electoral Register are put in such a matrix it would be of 983,598 x 437,639 or $4.3*10^{11}$ combinations. To put these sizes into context, the Office for National Statistics *Output Area Classification (OAC)* (Vickers and Rees, 2007) was clustered using k-means from a matrix of 223,060 output areas by 41 columns (census variables) or $9.14*10^6$, that is between 10^2 and 10^5 times simpler than the two different name matrices proposed here. The implementation of most clustering algorithms are designed to read in a similarity matrix based on just a small number of columns, since these are typically a few attributes describing the objects (rows), as in the OAC example just mentioned. These implementations cannot work with matrices of several thousands columns. Moreover, even if it were possible to process them, there are problems in preparing such matrices in the required format given that most database management software has quite stringent

restrictions on the number of columns that can be created, or the size of the total file (seeTable 6.7).

Software package	Columns limit	Rows limit
MS Excel	256 columns	65,536 rows
MS Access	256 columns	Up to 2 Gb file size
SPSS	Up to 2 Gb file size (only 8bit character names)	Up to 2 Gb file size
Oracle 10g	1,024 columns	No limit

Table 6.7: Limitations on the maximum number of columns and rows in standard software (Sources: Microsoft.com, Oracle.com, SPSS.com)

One option to overcome this problem is to build the matrices in 'row format', which means that each pair combination is listed as a row with their similarity measure. In the example above, this would be: in the *m* x *n* matrix, 'forename, surname, frequency'; or in the *n* x *n* matrix, 'forename1, forename2, surname tokens'. The latter example produces a file of three columns and n^2 rows. The problem with this format applied to these names matrices is that the majority of the name combinations would have zero frequencies. This is a case in what are known as 'sparse matrices', or matrices primarily populated with zeros (Pissanetzky, 1984). A further improvement to the row matrix format when most values are zero is the compressed sparse row format (CSR), that only stores the non-zero values but requires two supporting files to indicate the row-column combination to which each value belongs (Karypis, 2003). An example of how the amount of data stored can be made much simpler with CSR is shown in Figure 6.7.

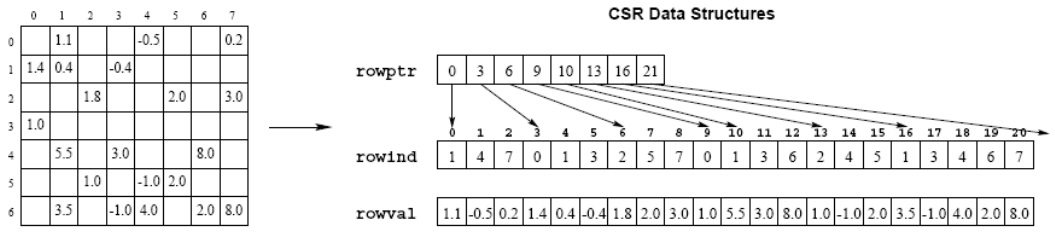

Figure 6.7: Converting a matrix to the compressed storage row format (CSR)
The matrix on the left is stored in compressed row format (CSR) by creating three files; *rowptr*, *rowind* and *rowval*. *rowval* stores the non-zero values in the matrix, starting from the top-left of the matrix and reading values within a row from left to right and moving from row to row from top to bottom. *rowind*, stores the column id corresponding to the value stored in the same position in *rowval*. *rowptr*, stores the position in *rowind* in which a change of row in the matrix is produced. Source: (Karypis, 2003: 38)

Compressed sparse row format solves the storage problem of handling the name matrices required by the clustering algorithm, but the computational requirement to populate such matrices is still very large. In order to produce an *n* x *n* matrix, a programme was built in PL/SQL (the programming language of Oracle) to compute the number of cross-occurrences between any two pairs of forenames. This required that for every pair of forename types, a full search through the whole database containing the 46 million records of the Electoral Register would be performed in order to count how many surname tokens had the two forenames in common. The number of records could be slightly reduced if the individual occurrences were aggregated to counts of forename-surname combinations, but this only brought the database down to 31,217,358 records. The programme crashed after seven days running when the temporary file created by Oracle reached a size of 52 Gb and there was no more free disk space. Several changes were done to the programme to increase its efficiency and a test was run on a sample of 100,000 people, which took 7 hours to compute. Because the time taken to process more names grows exponentially with the number of names, it was calculated that even if unlimited hard disk space were provided and the best computer available at UCL Geography was

used, this programme would still take 326 days to run on the whole Electoral Register. It was therefore decided to abandon this exploratory analysis at this point, but there are obvious room for future enhancement of computing algorithms to produce these matrices of cross-occurrences between the same type of names (i.e. forenames x forenames or surname x surnames).

Despite this problem, the *n* x *m* matrices of names are much more straightforward to produce than *n* x *n* matrices. The former just require an aggregation of the rows in the Electoral Register counting the number of forename-surname combinations, while the latter requires a lengthy computation to be carried out for every name type and all other names in the Electoral Roll. The former *n* x *m* table was produced, with 31,217,358 rows and three columns; forename, surname, and frequency. The frequency field was standardised into percentages per surname, that is, for each surname, the number of forename tokens divided by the total number of surname tokens. The resulting similarity matrix was finally stored in the CSR matrix format. Several clustering programs were tested to process this matrix, such as using k-means in SAS (SAS, 2006), partitioning methods in CLUTO (Karypis, 2003), or Self-Organising Maps (SOM) in Koh.exe (Kleiweg, 2001), but none of them was able to cluster so many rows and columns. After reducing the number of columns to just the 18,125 non-British forename types in the original seed list, the programs crashed at various stages, or provided inconsistent results when smaller samples were tested. The main causes of these inconsistencies were the high number of zeros in the matrix, since the k-means algorithm in particular is not designed to deal with a large proportion of zero values that it deems to be 'missing values'. This was also because the differences between a zero value and a very low percentage were very small, and

the k-means algorithm did not manage to detect the necessary clusters, since the technique operated more as in a binary problem (presence/absence) than as a distance measurement one (closer/further).

A new avenue for future research that would avoid the problem of sparseness in the similarity matrix is a research area known as high-dimensionality clustering. These clustering algorithms were originally designed to cluster highly dimensional gene frequency matrices in genetics studies. They differ from traditional clustering approaches in that the clustering space is divided into subspaces, and only those that are deemed 'more interesting' for clustering are used, thereby providing a better solution in terms of efficiency and clustering quality (Baumgartner et al, 2004). Degree of 'interest' is defined here as subspaces of the matrix where the greatest concentrations of similar values occur. These algorithms were not systematically tested in the context of this PhD, but exploratory work with the program *SURFING* (SUbspaces Relevant For clusterING) (Baumgartner et al, 2004) produced promising results which will be further analysed in the near future.

6.6. Conclusion

This chapter has evaluated the results of the names classification previously created through a heuristic approach. The initial heuristic classification presented a series of limitations that have been summarised in this chapter under ten major groups: the moving target problem; different data availability; lack of pre-conceived notions on optimal name classification methods; rather arbitrary rules sequencing; use of ad-hoc variables, thresholds and decisions; the effect of the ecological fallacy arising from

use of areally averaged datasets; the number of exceptions; the variability between iterations; CEL overlap and classification overriding; and the inconsistency inherent in manual checking and reclassification. As a consequence, the heuristic classification failed to pass a basic test of scientific reproducibility and presented a lack of internal consistency and simplicity.

Distilling the accumulated positive and negative experience of the exploratory phase in a robust and transparent manner, an automated and integrated approach was developed in the rest of the chapter. This automated approach has overcome most of the previous limitations, presenting a classification model that first builds a seed list of forenames that is very indicative of CEL Subgroups (using knowledge previously gained through the heuristic approach), and then uses it to classify surnames and forenames through a series of identical cycles. There are no exceptions in the automated model, the rules applied are much simpler, only two thresholds are used, and the whole model is clearly specified and built into a database programme through a series of SQL query statements that perform these processes in an automated way without manual intervention. These features mean that the methodology is much more consistent and much easier to explain and reproduce that the previous heuristic approach.

The next chapter will focus on the evaluation of the automated approach, in which it will be tested against other sources of ethnicity data. However, at this stage there is a further enhancement that could have been introduced in the automated classification; removing the requirement of a forenames seed list externally classified (in this case derived from knowledge gained through the heuristic approach). Several alternatives

were explored in this research exploiting clustering methods but because of the size and complexity of the names datasets and its relationships, these investigations did not progress to the implementation phase. These attempts were nevertheless described at the end of the chapter with the aim of informing future research in this area.

Chapter 7. Validating the CEL Name Classification

In the previous two chapters a classification of names into cultural, ethnic and linguistic groups (CEL) has been developed, first through a set of experimental rules and techniques which provided the experience upon which the final automated classification of names described in the previous chapter was based. This final automated classification constitutes a single, self-contained and robust classification of forenames and surnames and forms the main output of this PhD research. However, 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 the objective of the present chapter is to provide more than one way to validate the effectiveness of the CEL classification proposed in this thesis. 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 presented in Appendix 3, although it will usually refer to the CEL Subgroups developed in the automated phase.

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.

A selection of the best-practice examples of evaluations of name classifications found in the literature has been reviewed in Chapter 3. In that review thirteen studies were analysed, most of them from the public health literature, in which their name classifications were validated using lists of individuals where reported ethnicity, country of birth or nationality was already known, typically using patient registers. In this research two types of validation of the CEL classification have been performed, aimed at two different fields of application; classifying individuals and classifying neighbourhoods, the two aspects that have driven the justification for the research presented in this thesis. The first type, in line with the public health literature tradition, mentioned before, used a large list of individual hospital admissions in an area of London (the Boroughs of Camden and Islington) where the name and ethnicity of patients is known. The second type of validation, following a 'geography tradition', used the ethnicity data reported in the UK Census of Population at small area (Census Output Area), compared against the CEL classification of the same areas using the names in the GB Electoral Register.

The chapter is structured in four sections. Section 7.1 provides a justification and description of the algorithm used to assign a CEL at a person level, in order to move from two forename-to-CEL and surname-to-CEL tables to a classified list of full names into CELs. Section 7.2 discusses the difficulties of validating the CEL names

classification, relating to issues of differences in the ontological constructions of self-reported and names-based ethnicities. Section 7.3 presents the results of the first validation carried out using a large database of hospital admissions in the London Boroughs of Camden and Islington. Section 7.4 includes a similar validation using the ethnicity data reported in the UK Census at Output Area compared against the CEL classification of the same areas using the names in the GB 2004 Electoral Register. Finally some concluding comments on both validations are offered.

7.1. 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'.

7.2. 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 PhD 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 PhD extending from the previous onomastics literature, and discussed in Chapter 4, 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.

7.3. Validation Against Hospital Admission Ethnicity Data

This section describes the first phase in the process of validation of the CEL classification, in which, 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 PhD 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, full details of which are provided in Appendix 6.

The ethnic composition of the population of Camden and Islington is very diverse, as it can be seen in Figure 7.1, 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).

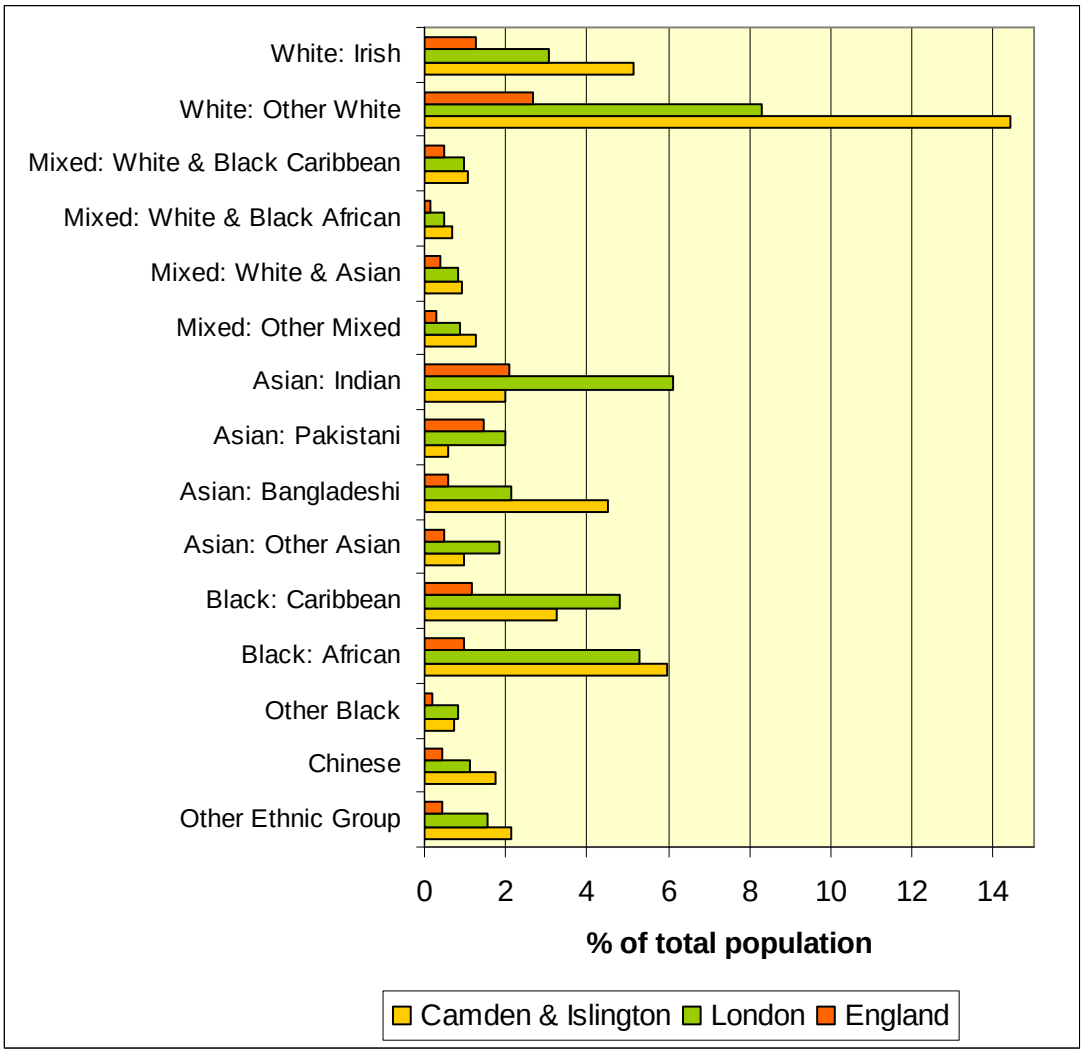

Figure 7.1: Population by ethnic minority in the London Boroughs of Camden and Islington
The chart shows the percentage of the total population that each ethnic minority represents in the London Boroughs of Camden and Islington (dark yellow), compared with the percentages for London (green) and England (dark orange). Source: ONS 2001 Census.

7.3.1. Hospital Episode Statistics data description

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, 2005a) 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.

7.3.2. 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, 2005a; 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

7.3.3. 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.

7.3.4. 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 7.1. A summary of the results obtained at CEL Group level is presented in Table 7.1, 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 7.1: Results of classifying the HES_Person table using the CEL name classification summarised at CEL Group 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 7.3.3. 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 7.2: Matrix comparing number of persons by CEL vs. HES Ethnicity using 1991 Census ethnic groups

Table 7.2 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 7.2, 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 7.2, 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 Table 3.6, in a more visual fashion.

These classification evaluation measures were calculated for the matrix shown in Table 7.2, 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 7.3 (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 7.3, 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 7.3: 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 7.2 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 7.3.1 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 7.3. The results of this table as well as of the next tables are discussed in Section 7.3.6.

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 7.4.

	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 Caribbean	n/a	n/a	n/a	n/a
E	Mixed- White and Black African	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 7.4: 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.

7.3.5. 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 7.3.4 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.

7.3.6. Discussion of results

Sections 7.3.4 and 7.3.5 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 7.3 and Table 7.4, 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 7.3 and Table 7.4 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 PhD

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 7.3 and Table 7.4, 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 7.2, 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 thesis.

7.4. Validation Against Census Small Area Ethnicity Data

The second type of validation of the CEL name classification follows a 'geography tradition' and seeks to evaluate the ability of the CEL classification to correctly identify ethnicity at the level of the small area aggregation (as opposed to unique human individuals). The validation attempted here uses the ethnicity data reported in the UK Census of Population for small areas (Census Output Area), which are compared against the CEL classification of the same areas using the names in the GB 2004 Electoral Register.

7.4.1. Data preparation

This validation requires the use of two datasets to be compared: Census 2001 Key Statistics KS06 table (ethnic group), and a new dataset to be prepared by coding the GB 2004 Electoral Register by CEL Subgroup, which is then aggregated by the Census Output Area geography, and by the Census 16 ethnic groups.

Each of the 46.3 million adults in the GB 2004 Electoral Register file, described in Chapter 4, was classified by CEL Subgroup, using the name-to-CEL tables of the automated approached described in Chapter 6, and the person allocation algorithms

explained in Section 7.1 of the current chapter. This comprises, to the author's knowledge, the first attempt ever to classify the whole population of Great Britain according to the cultural, ethnic and linguistic origin of names, and the results presented in Table 7.5 are summarized by CEL Group. The individual records were then aggregated by unit postcode, calculating the number of people per CEL Subgroup in each unit postcode, yielding a table of 1.4 million records (postcodes) and 66 columns (CEL Subgroups).

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

Table 7.5: Summary of number of people per CEL Group in the GB 2004 Electoral Register

Two further steps are required in order to make the Electoral Register dataset comparable with the Census; the aggregation of the 66 CEL Subgroups to 16 Census ethnic groups, and of the unit postcode geography to Census Output Areas. The first step is achieved through the CEL Subgroup lookup table previously used that relates

the 66 CEL Subgroups to various attributes, one of them being the sixteen 2001 Census ethnic groups, and which appears in Appendix 3. Therefore, the 66 columns are now aggregated into the 2001 Census 16 ethnic groups.

The second step requires the use of an additional dataset, the postcode directory maintained by Office of National Statistics, named the National Statistics Postcode Directory (NSPD) (formerly known as the All Fields Postcode Directory (AFPD) and previously the *Gridlink* Postcode Directory: (Office for National Statistics, 2006b). The NSPD provides a lookup table between every unit postcode in the UK to a set of higher level geographies to which it belongs. Those that are of interest to this validation exercise are; Census Output Area (OA), Lower level Super Output Area (LSOAs), Ward, and Local Authority (LA). Therefore, the 1.4 million records were separately aggregated by each of these four geographic levels, generating four tables of different spatial resolutions: 218,037 OAs, 40,883 LSOAs, 10,072 Wards, and 408 Local Authorities in Great Britain (i.e. excluding Northern Ireland). Each of these tables contains counts of persons in the Electoral Register by ethnic group based on the CEL Subgroup classification but expressed as their 2001 Census ethnic group equivalent. To avoid confusion, these four tables will be generally referred to as the CEL-GB04 datasets.

7.4.2. Data analysis: validation of CEL vs. Census ethnicity at small area

Comparison of the CEL-GB04 datasets and the Census ethnic groups by small area (2001 Census Key Statistics KS06 table) was effected by linking both datasets at each of the geographical levels for which the validation was to be performed: OA, LSOA, Ward and LA. The idea of performing the validation at four different geographical scales was to assess the sensitivity of the CEL classification to changes

in scale, as a precursor to determining the optimum geographical level of its applications.

This analysis entailed calculating correlation coefficients between the CEL-GB04 dataset and the Census ethnicity responses at the four different levels of geography. Since the two datasets do not use the same denominator (the CEL file only includes adults entitled to vote while the Census enumerates all of the resident population), the comparison was performed using the proportion of people in each ethnic group for each geographical unit. A correlation matrix was calculated for these figures using Pearson's correlation coefficient (Robinson, 1998). A summary of the results is offered in Table 7.6, which summarises the correlation coefficients at the different levels of geography for which they were calculated (OA, LSOA, Ward and LA), and also for 11 Census ethnic groups after removing the four 'Mixed' and the 'Not Stated' categories. The number of geographical units at each level and their population sizes are indicated at the bottom of the table.

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

Table 7.6: Summary of Pearson's correlation coefficients between the CEL-GB04 and 2001 Census datasets
All correlations are significant at the 0.01 level (2-tailed). Correlations ≥ 0.7 are highlighted in bold
OA = Output Area, LSOA= Lower Super Output Area, LA= Local Authority

7.4.3. Discussion of results

Most of the ethnic group categories present a high degree of correlation between the two datasets, which generally increase with area size, although correlations are not as strong at OA level. The former effect is to be expected according to the *modifiable areal unit problem* (MAUP) (Openshaw, 1984). There are however some groups for which anomalies occur: the 'White-Irish' (B), 'Any other Asian background' (L),

and ‘Any Other ethnic group’ (S) categories, and to a lesser extent the ‘Black-Caribbean’ (M) group, which each present over-all low correlations. The main reasons for this divergence are, on the one hand, the inherent vagueness of some Census categories (Mixed, Other) together with their lack of exact correspondence to the CEL Subgroups, and on the other hand, some problems detected in the distinction of Irish and Caribbean names, that are due to historic differences and a high degree of assimilation with the White-British majority. Nevertheless, the correlation coefficient of all the categories is significant at the 0.01 level (2-tailed). The stronger correlations (≥ 0.7) are highlighted in bold in Table 7.6, including 6 out of 11 categories at OA level, and especially at the LSOA level (1,500 persons in average) and coarser geographies. Some groups perform extraordinarily well across all scales; ‘White British’ (A), ‘White-Other’ (C), ‘Indian’ (H), ‘Pakistani’ (J), ‘Bangladeshi’ (K), ‘Black African’ (N) and to a lesser extent Chinese (R), probably indicating the robustness of their Census ethnic categories as well as a strong linkage between current self-perception of ethnic identity and name origins for those groups. With the exception of ‘White-Other’ and ‘Indian’, these are also the ethnic groups which performed best in the validation using the HES dataset in Camden and Islington presented in Section 7.3, an area of London with fewer residents from the ‘Indian’ ethnic group. Even more, at LSOA and higher geographies, all groups except for the ‘Other’ mentioned (L & S) perform very well when compared with the Census geographical distribution, with correlations above 0.80, many of them above 0.90 (see right columns of Table 7.6 for details).

7.5. Conclusion

The main aim set at the outset of this chapter, to provide more than one way to validate the effectiveness of the CEL classification proposed in this thesis, has been achieved through two different validation exercises, one following a public health literature tradition, using patient registers, and a separate one following a 'geography tradition' comparing Census small area statistics.

The first section in this chapter has presented a methodology that draws upon the two forename-to-CEL and surname-to-CEL tables created in Chapter 6 to create a person level CEL allocation algorithm to assign individuals with their most probable CEL (PCEL). The scores used in this process have been very useful to arbitrate in conflicts between the FCEL and SCEL of a person. This has also allowed assignment of individuals based on a final score that indicates the strength of the final allocation.

Once the person level CEL allocation method was explained, the chapter moved on to applying it to two different datasets in which the self-reported ethnicity was already known. The first of these validation exercises used a list of people admitted to hospital in London Boroughs of Camden and Islington containing forename, surname and self-reported ethnicity data. The second type of validation used a nationwide dataset, the GB 2004 Electoral Register, which was coded by CEL and aggregated by Census Output Area in order to compare the results with the Census small area statistics on ethnicity.

The results of both of the validation exercises have been very consistent, even though the nature of the two datasets was very different. The CEL names classification

achieved an overall high precision in the following ethnic groups (reported as 2001 Census categories): 'White British' (A), 'Indian' (H), 'Pakistani' (J), 'Bangladeshi' (K), 'Black African' (N) and Chinese (R). It achieved a medium precision in the 'White-Other' (C) group (high in the Census but low in the HES validation). On the other hand, the ethnic groups where the CEL classification proved less effective were: 'White-Irish' (B), 'Black-Caribbean' (M), 'Any other Asian background' (L), and 'Any Other ethnic group' (S) categories. Finally, the name approach by definition was not able to identify persons which in the 2001 Census categories identify themselves with one of the four mixed ethnicity groups (D-G), or the vague 'Any other Black background' (P) since they cannot be matched to any name group.

Excluding the problematic catch-all 'other' categories that are of dubious worth in the self-reported ethnicity data, the only two groups in which the classification is less effective are thus the 'White-Irish' (B) and 'Black-Caribbean' (M) groups. The reasons behind the high number of misallocations for these groups has been extensively justified in Sections 7.3 and 7.4, and relate to a high degree of assimilation with the White-British majority in the former group, and the British-origin of Caribbean names for the latter group. The rest of the six main ethnic groups, as reported by the census, achieve a high degree of accuracy, both measured by the measures of sensitivity, specificity, positive predictive value (PPV), and negative predicted value (NPV), using hospital admissions data, as well as by the correlation coefficient with the Census small area data. Moreover, these results are affected by the present PhD research's objective of building a name classification that maximises population coverage, as opposed to classification accuracy. That is,

when the weakest name assignments are dropped the accuracy of the classification increases and vice-versa.

Finally, these validation results should be interpreted in the light of the caveats mentioned in Section 7.2, relating to the problem of comparing constructs of different ontological nature. Hence, onomastic classifications based on cultural, ethnic and linguistic origin of names, cannot be easily compared with self-reported ethnicity. The comparisons carried out in this chapter, therefore assume that the differences between the two ontologies of ethnicity can be ignored when there is no other alternative information source. As a result, the fact that the only two ethnic groups that present major issues in the CEL classification are Irish and Caribbean ones comes as no surprise. After all, these are precisely the two groups in Britain that better reflect the ontological difference between the two constructs compared; names-based origin and self-reported identity, or Irish names and White-Irish identity, and Caribbean-British names and Black-Caribbean identity.

After the CEL classification had been evaluated a series of applications were tested in the public sector, to test its actual usefulness in the 'real world'. It can be claimed that proving that such applications work in reality is also part of the validation of the CEL methodology developed in this PhD research. One of these applications will be presented in depth in the next chapter, while others will be briefly mentioned.

Chapter 8. Applications: Residential Segregation and Ethnic Inequalities

What could be more inherently geographical than segregation?

(Brown and Chung, 2006: 125)

The literature on name-based ethnicity classifications that was reviewed in Chapter 3 is very rich in studies that have developed, validated and applied name-based methods to ascribe population ethnic origins, especially since the 1950s in the fields of public health, genetics, and demography. The search strategy used in that chapter identified 186 unique publications that either directly developed name-based methodologies or used externally available methodologies. The majority of these studies were originally conceived with a particular application in mind, using name analysis to segment a population into a few ethnic groups for further analysis of suspected differences between groups. Therefore, the primary focus of most studies in the name-based ethnicity classifications literature has been on applications, and the studies analysed in Chapter 3 have all demonstrated their value and sufficient accuracy in classifying ethnicity in the context for which they were designed. The types of applications of name-based classifications are therefore closely intertwined with the methodological developments in this field, probably because of the majority of them have been developed by strongly empirically-led health and genetics researchers.

The primary aim of this PhD research is a methodological one; to develop a new ontology of ethnicity based on personal names, leading to an alternative name-based

ethnicity classification system covering the whole population and maximising the number of ethnic groups. As a consequence, the methodology has not been developed with any one particular application in mind, nor has a specific line of examples been developed through the previous methodological chapters. However, as the second part of this thesis title suggests – 'implications for neighbourhood profiling' – the area of applications finally envisaged for the new methodology presented in the previous chapters is primarily one of geographical nature.

It is believed that one of the areas in which name-based ethnicity classifications have greatest potential is in geographical analysis of small areas, i.e. neighbourhoods, where the intersection of the majority of the factors influencing ethnic inequalities, described in Chapter 2, actually takes place and acquires an interpretable meaning in everyday practices and encounters (Amin, 2002). Moreover, there is a recognised need to differentiate the identity of neighbourhoods in the delivery of public services.

> *'We need to be better able to differentiate between locations, not just on account of their physical attributes but also by virtue of their identification with specific identities'*
>
> (Longley, 2003: 116)

Therefore, this chapter will illustrate one thread of many potential fields of applications of the CEL name to ethnicity methodology developed in this PhD. It is included here as an illustrative example of a geographical application of the methodology at the small area scale, particularly in ethnicity profiling of neighbourhoods.

The application presented here focuses upon London, which is the most ethnically diverse region of the UK, and one of the few global cities with a significant proportion of its population originally coming from all over the world. In 2001 London Non-White British population comprised 40.2% of the total population of 7.1 million people (ONS 2001 Census Key Statistics KS06 table). However, and as discussed in Chapter 1, ethnic group categories in the 2001 Census are sometimes too broad to understand the causes for residential segregation, especially in London.

Table 8.1 shows the population of each ethnic group as a share of the total population of London alongside its national average for the UK. The groups highlighted in italics are considered 'poorly studied' groups (the 'other' groups plus 'Black African') since they lump together very diverse ethnicities into meaningless 'other' 'left-overs', lost between the major ethnic groups. However, in London these poorly studied groups comprise a total population of 1.35 million people, or 18.8% of the total population and 46.7% of the ethnic minority population. It is envisaged that the CEL methodology will be specially valuable to break down these ethnic groups into finer and meaningful groups that can be further analysed. The study of the residential segregation of such groups is the main purpose of the analysis presented here.

The main application presented here intends to illustrate the potential applications of the CEL name classification to issues surrounding neighbourhood profiling and residential segregation debates. As exposed in the literature review presented in Section 2.2, these issues are of most relevance in public policy debate in Britain, and in the developed world in general.

	UK	London
White		
British	87.5%	**59.8%**
Irish	1.2%	**3.1%**
Other White	*2.6%*	***8.3%***
Mixed		
White & Black Caribbean	0.5%	**1.0%**
White & Black African	0.2%	**0.5%**
White & Asian	0.4%	**0.8%**
Other Mixed	*0.3%*	***0.9%***
Black or Black-British		
Black-Caribbean	1.1%	**4.8%**
Black-African	*0.9%*	***5.3%***
Black-Other	*0.2%*	***0.8%***
Asian or Asian-British		
Indian	2.0%	**6.1%**
Pakistani	1.4%	**2.0%**
Bangladeshi	0.5%	**2.1%**
Any other Asian background	*0.5%*	***1.9%***
Chinese or other group		
Chinese	0.4%	**1.1%**
Any other ethnic group	*0.4%*	***1.6%***
Total Non-White British	12.5%	**40.2%**
Poorly Studied Groups	*4.9%*	***18.8%***

Table 8.1: Proportion of the population by ethnic group; London vs. UK (2001 UK Census)
'Poorly studied' groups comprise the 'other' categories plus 'Black African' and are highlighted in italics. Source: Office for National Statistics 2001 Census, Key Statistics KS06 table (Crown Copyright).

Other examples of actual applications of the methodology developed in this PhD research are briefly mentioned at the end of the chapter in order to illustrate avenues for future applied research in this area. These include applications in public health, to segmenting populations by ethnic group to tackle ethnic inequalities in health at local level, and in demographic planning at local and central government, complementing current methodologies to estimate population composition by ethnic group,

especially at local level. The applications presented here do not purport to provide a comprehensive account of the specific applications for the CEL classification, but rather present a reasonably representative range of examples of potential implementations of the methodology to common problems identified in the geographical ethnic inequalities literature.

Section 8.1 presents the justification of the analysis of residential segregation in London and introduces the methods used. Section 8.2 analyses in detail four of the five traditional dimensions of residential segregation, drawn from the sociological literature, while Section 8.3 expands on additional dimensions and approaches from a geographical perspective. Section 8.4 summarises the overall results found and discusses the issues identified. Finally, Section 8.5 briefly mentions other applications of the CEL methodology in the area of ethnic inequalities in health and population studies in local government, with the purpose to illustrate some of the potential future applications of the methodology developed in this PhD research.

8.1. Residential Segregation in London. Introduction and Methods

8.1.1. Introduction

The main application presented in this chapter seeks to illustrate the relevance of the CEL methodology to the issues identified in the literature review and highlighted in Section 2.2. In particular, it intends to show how name analysis can be a feasible alternative to self-reported ethnicity information, when analysing apparent segregation of neighbourhoods. This pertains to the criticised persistence of a skin colour criterion when defining segregation, around a White / Non-White divide, which usually ascribes Non-White residential concentrations with negative

connotations (Simpson, 2004). However, as justified in Chapter 2, the reality of neighbourhood segregation is more likely to be based upon a complex spectrum of 'skin tones' or culturally diverse neighbourhoods, and it is believed that name analysis can be useful to reveal its complex geography.

This section will not go deeper into the issue of the meaning of a 'segregated' or an 'integrated' neighbourhood or city. However, it intends to show how the spatial distribution of an alternative ontology of ethnicity based on name origins, can change established perceptions of the nature of the most segregated ethnic groups and the level of segregation of particular neighbourhoods. Therefore, the focus of this example will be on ethnic group categorisations at much finer levels than the ethnic minority aggregations typically studied in Britain;– viz. South Asian (Indian, Pakistani and Bangladeshi) (Peach, 1998), Black (Phillips, 1998), or Muslim (Peach, 2006; Peach and Owen, 2004). As such, this contribution seeks to provide new evidence about the ethnic groups categorised as 'Other' in official statistics (Connolly and Gardener, 2005). More contributions of this kind, which might stem from future applications of the CEL name classification, should help to advance the debate about the ontology of ethnicity and segregation, and how it may affect the results of geographical analysis at the neighbourhood level.

The example presented here entails classification of the names of London's population, as per the 2004 Electoral Register, into 66 CEL Subgroups, in order to analyse the level of segregation of ethnic groups and neighbourhoods at very fine scales (CEL Subgroup and Census Output Area). Segregation is measured using

traditional indices of segregation, taken from the sociological and geographical literatures, as well as using spatial autocorrelation measures.

8.1.2. Data preparation and methods

The dataset used in this analysis is the 'CEL-classified' 2004 Electoral Register for Greater London, which contained 5 million electors, individually classified into 66 CEL Subgroups as per the process described in Section 7.1. As a result, 99.79% of the individuals could be allocated with a CEL Subgroup, what constitutes a remarkable achievement in terms of population coverage. A summary table of the sizes of each of these CEL Subgroups is listed in Table 8.2. Individuals were then aggregated into the 131,721 unit postcodes of the Capital, computing counts of people per CEL Subgroup and postcode unit. Finally this table was further aggregated into Output Areas (OAs), a geographical unit that is apt for London-wide analysis since its average size is 285 people in London and there is a total of 24,100 OAs. The linkage between postcode units and OAs was made using the National Statistics Postcode Directory (NSPD) (Office for National Statistics, 2006b) as previously described in Section 7.4.1. The NSPD directory was also used to aggregate both postcode units and OAs up to higher level geographies (ordered in increasing size; Lower Super Output Areas -LSOA-, Wards, and London Borough). Each of these geographies was mapped through a GIS using OS CodePoint boundaries for the postcode units and the Census administrative geographies for the OAs and their higher level administrative aggregations.

The analysis involved the calculation of a set of well-established residential segregation indices at each of the different levels of geography described above. A software application called *Segregation Analyser*, developed by Apparicio et al

(2005), was used to compute the residential segregation indices for all of the CEL Subgroups at a range of different geographical scales. This tool significantly simplified this task, since it computes over 40 different segregation indices using as an input a geographical boundary file of the area with the population headcounts per areal unit and ethnic group. This software application is available from the *Centre Urbanisation, Culture et Societé* in Quebec City part of the *Institut National de la Recherche Scientifique* (INRS), available as follows:

http://www.inrs-ucs.uquebec.ca/inc/Groupes/LASER/Segregation.zip (last accessed 07/09/2006)

However, because of computer memory limitations the segregation indices at postcode unit level for London (n=131,721) could not be calculated using the Segregation Analysis tool, because of the intensive process of dealing with very small geographical units. Therefore, the calculations were applied to Output Areas (n=24,100) and higher order aggregations.

CEL Subgroup	Total Pop.	%
ENGLISH	2,876,980	57.47%
IRISH	414,038	8.27%
SCOTTISH	323,847	6.47%
WELSH	222,429	4.44%
HINDI INDIAN	156,269	3.12%
PAKISTANI	140,548	2.81%
SIKH	83,968	1.68%
BANGLADESHI	72,829	1.45%
ITALIAN	71,967	1.44%
NIGERIAN	68,596	1.37%
GREEK	61,296	1.22%
MUSLIM MIDDLE EAST	48,114	0.96%
PORTUGUESE	44,780	0.89%
SPANISH	44,679	0.89%
FRENCH	40,264	0.80%
SRI LANKAN	39,269	0.78%
JEWISH	35,984	0.72%
HONG KONGESE	35,609	0.71%
GHANAIAN	35,255	0.70%
TURKISH	34,359	0.69%
POLISH	33,270	0.66%
GERMAN	33,264	0.66%
PAKISTANI KASHMIR	32,061	0.64%
INDIA NORTH	31,888	0.64%
NORWEGIAN	24,927	0.50%
SOMALIAN	20,376	0.41%
HINDI NOT INDIAN	12,643	0.25%
BLACK CARIBBEAN	11,554	0.23%
MUSLIM SOUTH ASIAN	11,380	0.23%
EUROPEAN OTHER	9,091	0.18%
BALKAN	9,035	0.18%
CHINESE	8,874	0.18%
SOUTH ASIAN OTHER	8,484	0.17%
VIETNAM	8,415	0.17%
INTERNATIONAL	6,214	0.12%
RUSSIAN	5,539	0.11%
DUTCH	5,477	0.11%
SWEDISH	5,155	0.10%
AFRICAN	4,879	0.10%
IRANIAN	4,761	0.10%
DANISH	4,592	0.09%
SIERRA LEONIAN	3,854	0.08%
JAPANESE	3,469	0.07%
AFRIKAANS	3,036	0.06%
EAST ASIAN	2,645	0.05%
HUNGARIAN	2,603	0.05%
ARMENIAN	2,436	0.05%
MUSLIM	2,335	0.05%
BLACK SOUTH AFRICA	2,161	0.04%
FINNISH	2,099	0.04%
MUSLIM NORTHAFRICAN	2,044	0.04%
ALBANIA	1,908	0.04%
CZECH & SLOVAKIAN	1,660	0.03%
UKRANIAN	1,629	0.03%
LEBANESE	1,404	0.03%
NORDIC	1,174	0.02%
MUSLIM STANS	1,155	0.02%
KOREAN	1,139	0.02%
ROMANIAN	1,085	0.02%
BALTIC	1,061	0.02%
ERITREAN	1,053	0.02%
ETHIOPIAN	918	0.02%
MALAYSIA	891	0.02%
UGANDAN	812	0.02%
CONGOLESE	598	0.01%
UNKNOWN NAME	10,546	0.21%
VOID NAME	90,715	1.81%
GRAND TOTAL	**5,006,490**	**100.00%**

Table 8.2: List of the 66 CEL Subgroups and their total and relative population sizes in London (2004)
The table is ordered by decreasing population size. The category ‘unknown name’ has been added, although it does not constitute a CEL Subgroup per se.

8.2. The Traditional Dimensions of Residential Segregation

8.2.1. Selection of segregation indices

Drawing upon Massey and Denton's (1988) famous 'five dimensions of residential segregation', a selection of indices was made, one for each dimension of evenness, exposure, concentration, and clustering. No index of centralisation was used because of the multiplicity of historic town centres in London. The dimension of centralisation was devised for American cities where ethnic minorities typically occupy the inner city area, which comprises a well defined core, and gradually move out to the suburbs as they become more integrated (Peach et al, 1981). This process does not follow a similar pattern in Europe, and in London the multiplicity of historic town centres complicates the role of the functional city centre as an area of immigration settlement. Since centralisation indices are based on a single centre, and calculate a distance to the centre function, it was deemed irrelevant to the London case.

An exploratory analysis of residential segregation indices was carried out, including all of the indices reviewed by Massey and Denton (1988), spatial indices proposed in the subsequent literature (Wong, 2003; 2004), segregation classifications based on thresholds (Brimicombe, 2007; Johnston, Voas et al, 2003), and recent reviews of the adequacy of each of the most common residential indices (Simpson, 2004; 2007). The indices proposed by Massey and Denton (1988) as the best representative for each of the five dimensions, were those with higher loadings in the factor analysis carried out by these authors. Using these indices and comparing them with more complex indices such us those including spatial features by Wong (2003; 2004)

produced very similar results, and therefore simpler indices were preferred. As a result of this selection process four indices were finally adopted for further analysis, as per the following list:

- **Evenness**

  ID; Index of Dissimilarity $ID = \frac{1}{2}\sum_{i=1}^{n}\left|\frac{x_i}{X} - \frac{t_i - x_i}{T - X}\right|$ (Duncan and Duncan, 1955)

- **Exposure**

  ${}_xP_x^*$; Isolation Index ${}_xP_x^* = \sum_{i=1}^{n}\left[\frac{x_i}{X}\right]\cdot\left[\frac{x_i}{t_i}\right]$ (Bell, 1954) (Lieberson, 1981)

- **Concentration**

  ACO; Relative Concentration Index

$$ACO = 1 - \left\{\frac{\left[\sum_{i=1}^{n}\left(\frac{x_i A_i}{X}\right) - \sum_{i=1}^{n}\left(\frac{t_i A_i}{T_1}\right)\right]}{\left[\sum_{i=n_2}^{n}\left(\frac{t_i A_i}{T_2}\right) - \sum_{i=1}^{n_1}\left(\frac{t_i A_i}{T_1}\right)\right]}\right\}$$ (Massey and Denton, 1988)

  (spatial units are sorted by area size in ascending order)

- **Clustering**

  ACL; Absolute Clustering Index

$$ACL = \left[\frac{\sum_{i=1}^{n}\left(\frac{x_i}{X}\right)}{\sum_{j=1}^{n} c_{ij}x_j} - \frac{X}{n^2\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}}\right] \Bigg/ \left[\frac{\sum_{i=1}^{n}\left(\frac{x_i}{X}\right)}{\sum_{j=1}^{n} c_{ij}t_j} - \frac{X}{n^2\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}}\right]$$

  (Massey and Denton, 1988)

Key to the formulas:

X = Total population of group X in the whole area/city

x_i = Total population of group *X* in spatial unit *i*

x_j = Total population of group *X* in spatial unit *j*

T= Total population in the whole area/city

t_i= Total population in spatial unit *i*

t_j= Total population in spatial unit *j*

T_1 = The sum of all t_i in areal unit 1 to areal unit n_1

T_2 = The sum of all t_i in areal unit n_2 to areal unit n

c_{ij} = cell value of the binary connectivity matrix (1 where *i* and *j* are contiguous and 0 otherwise)

A_i = Area of spatial unit $_i$

$_xP_x^*$ = Probability of a member of ethnic group X entering into contact with a member of the same group within an area of residence

For a review of these indices, equations and their theoretical justification, see Massey and Denton (1988) and the original sources (Bell, 1954; Duncan and Duncan, 1955; Lieberson, 1981); for their implementation in Segregation Analyser, which correspond to the formulas presented here, see Apparicio et al (2005).

These four indices represent four of the five dimensions of residential segregation, and their meaning will be described in subsequent sections devoted to each dimension. Additional dimensions are dealt with in the next section. These indices were calculated for every CEL Subgroup at the Output Area level. The results of all of the calculations described here are presented in the next subsections. As a result of these calculations, a series of measures of residential segregation were produced for each of the 66 CEL Subgroups in Greater London. Those individuals who could not be classified by CEL Subgroup in the personal allocation algorithm (only 0.21%), were assigned with an additional code ‘Unknown Name’, bringing the total number of categories to 67. This ‘Unknown Name’ category has been treated as a separate

CEL Subgroup and indices were calculated for it to double check that it did not present any particular pattern and hence that their distribution is completely random. Two other CEL Subgroups that are included in the list of 67 are termed 'International' and 'Void' names. International names are those names, primarily forenames, that are widely adopted across CELs and are deemed to be of an 'international' nature, as opposed to any particular CEL. 'Void' names are those that have been identified as names but in a different category, for example surnames recorded as forenames, or those common mistakes in data quality assurance, such us honorifics (i.e. Mr., Ms, Dr. etc).

Unless otherwise specified, most of the figures that follow only take into account the most frequent 46 CEL Subgroups, for reasons of ease of representation and discussion. These correspond to the CEL Subgroups with a total population size in London greater than 3,000 people, which in the list shown in Table 8.2 corresponds to the 46 subgroups that are more numerous than the 'East Asian' category.

8.2.2. Evenness

Evenness was measured through the classic index of dissimilarity (ID) (Duncan and Duncan, 1955), which is portrayed by many as *the* segregation index (Simpson, 2007). The index of dissimilarity represents the proportion of the group's population that would have to move between areas in order for the group to become distributed in the same way as the rest of the population (evenly distributed, hence the name of this dimension).

$$ID = \frac{1}{2}\sum_{i=1}^{n}\left|\frac{x_i}{X} - \frac{t_i - x_i}{T - X}\right| \text{ (Duncan and Duncan, 1955)}$$

(See Section 8.2.1 for explanation of variables)

The ID index was calculated for the CEL Subgroups in London, and the results for the most frequent 46 CEL Subgroups are listed in Table 8.3. In this table the CEL Subgroups are ordered by descending index of dissimilarity (ID), noted by the rank, alongside their absolute population size in London. The most segregated CEL Subgroups in London according to the ID index are; Afrikaans, Sierra Leonean, Japanese, Iranian and African (a category encompassing other Black African names not included in the rest of CEL Subgroups). This is an interesting result, since these are not precisely the groups that come up at the top on the segregation literature on London (Johnston et al, 2002a; Peach, 1996a; Peach, 1999a). This demonstrates the value of the CEL methodology in uncovering the residential patterns of carefully defined disaggregate ethnic groups. The least segregated CEL Subgroups (out of the most frequent 46 CEL Subgroups) are Irish, Scottish, Welsh, English, and 'Void Names' (a category including invalid entries in the Electoral Register). This is likely to arise because of the ubiquity of these groups across the Capital, as a result of the long-established nature of these groups in London.

However, Table 8.3 suggests that there is a relationship between the size of the CEL Subgroup and the level of the segregation index. In order to corroborate this, Figure 8.1 shows the scatterplot of both items; the index of dissimilarity (ID) on the vertical axis and the total population size on the horizontal axis for the 46 CEL Subgroups, both represented in logarithmic scale. The plot shows a clear negative relationship

between the ID index and population size, which is confirmed by a regression line plotted between the points using a linear fit, whose R^2 is 0.805.

Rank	CEL Subgroup	Total Pop.	ID	Rank	CEL Subgroup	Total Pop.	ID
1	AFRIKAANS	3,036	0.909	24	JEWISH	35,984	0.620
2	SIERRA LEONEAN	3,854	0.908	25	GHANAIAN	35,255	0.611
3	JAPANESE	3,469	0.905	26	SOMALIAN	20,376	0.585
4	IRANIAN	4,761	0.875	27	NIGERIAN	68,596	0.580
5	AFRICAN	4,879	0.868	28	INDIA NORTH	31,888	0.574
6	DANISH	4,592	0.864	29	HINDI INDIAN	156,269	0.573
7	VIETNAM	8,415	0.862	30	GREEK	61,296	0.557
8	SWEDISH	5,155	0.854	31	HONG KONGESE	35,609	0.549
9	RUSSIAN	5,539	0.843	32	PAKISTANI KASHMIR	32,061	0.537
10	DUTCH	5,477	0.840	33	NORWEGIAN	24,927	0.516
11	INTERNATIONAL	6,214	0.789	34	GERMAN	33,264	0.499
12	SOUTH ASIAN OTHER	8,484	0.788	35	PAKISTANI	140,548	0.495
13	CHINESE	8,874	0.787	36	POLISH	33,270	0.478
14	BALKAN	9,035	0.774	37	MUSLIM MIDDLE EAST	48,114	0.469
15	EUROPEAN OTHER	9,091	0.761	38	PORTUGUESE	44,780	0.464
16	HINDI NOT INDIAN	12,643	0.754	39	SPANISH	44,679	0.459
17	BLACK CARIBBEAN	11,554	0.739	40	FRENCH	40,264	0.445
18	MUSLIM SOUTH ASIAN	11,380	0.734	41	ITALIAN	71,967	0.386
19	UNKNOWN NAME	10,546	0.695	42	VOID	90,715	0.341
20	SIKH	83,968	0.670	43	ENGLISH	2,876,980	0.249
21	SRI LANKAN	39,269	0.665	44	WELSH	222,429	0.206
22	BANGLADESHI	72,829	0.644	45	SCOTTISH	323,847	0.188
23	TURKISH	34,359	0.620	46	IRISH	414,038	0.180

Table 8.3: Index of Dissimilarity (ID) by CEL Subgroups in London at Output Area level
ID; Index of Dissimilarity (Duncan and Duncan, 1955), Rank; Rank ordered by ID in descending order. The table only lists the most frequent 46 CEL Subgroups (those with a total population size in London greater than 3,000 people) ranked by the index of dissimilarity (ID). The total population size of each CEL Subgroup in London is also listed. See text for explanation of 'International', 'Void' and 'Unknown Names' categories.

Nevertheless, this finding is at odds with the consensus in the literature stating that ID index is independent of the groups' size (Massey and Denton, 1988) (Simpson, 2004). However, it is also known that the index of dissimilarity is dependent on the number of areas in which a city is divided (Voas and Williamson, 2000), especially 'where the group numbers are small or the areal grid is very finely drawn' (Peach, 1996a: 218). This seems to be the factor most affecting the relationship shown in this analysis, since there are 24,100 OAs in London and the total size of most of the

groups in London are either below this figure or just above it, and hence very difficult that a group would be evenly spread across all of them.

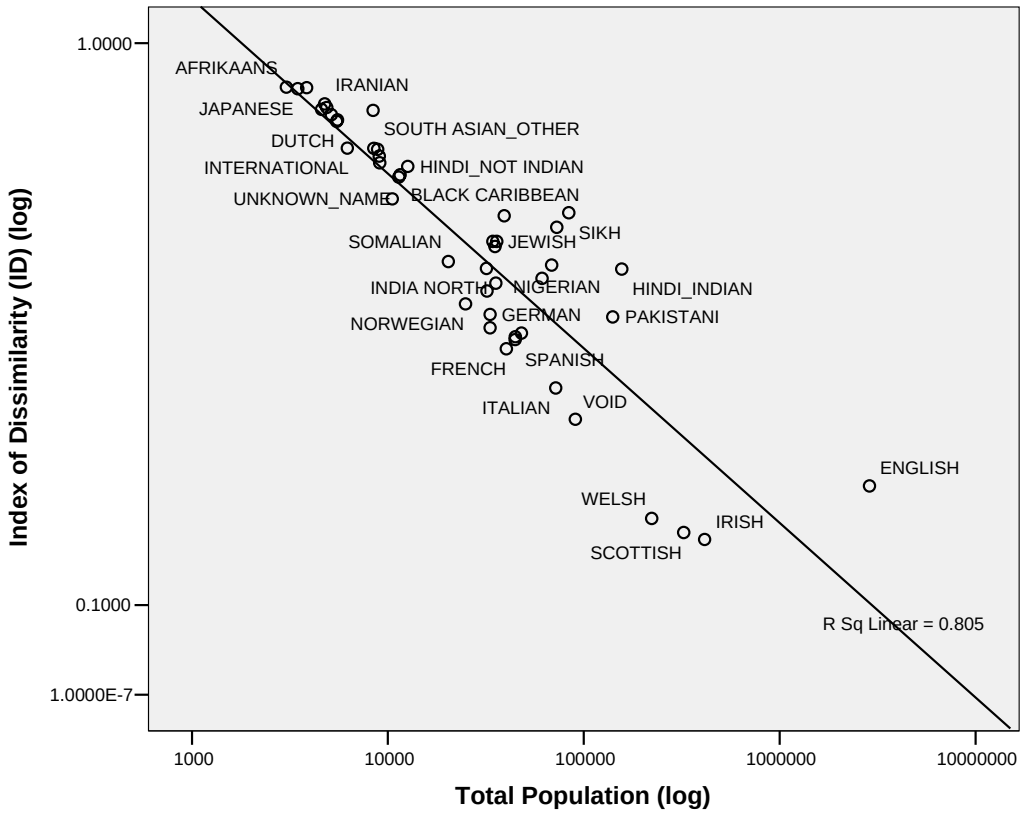

Figure 8.1: Scatterplot of CEL Subgroups Index of Dissimilarity (ID) at Output Area level vs. their total population size in London
This scatterplot only includes the most frequent 46 CEL Subgroups with a total population size in London greater than 3,000 people. The Index of Dissimilarity (Duncan and Duncan, 1955) calculated at Output Area level is represented on the vertical axis and the total population size of the CEL Subgroup in London on the horizontal axis. A trend line between the points is plotted using a linear fit, the R^2 of which is 0.805, demonstrating the relationship between segregation index and population size.

In any case, it is interesting to look at deviations from this relationship between group's size and the dissimilarity index in Figure 8.1, which are also readily apparent in Table 8.3. It is striking to notice the position of the English CEL Subgroup, which according to its disproportionate size would be expected to be the least segregated group of all, while the other three co-British Isles CELs (Irish, Scottish and Welsh) are less segregated than the English group, as expected by their population sizes.

Other groups which are more segregated than expected by their population size are Hindi Indian, Pakistani, Sikh, Jewish, Iranian and Greek.

Besides population size, another factor that ought to account for difference in the index of dissimilarity is the length of time since migration, since some ethnic groups have been longer established in Britain are likely to have lower residential segregation. To test this point, Figure 8.2 shows a scatterplot of the index of dissimilarity (ID) of 26 CEL Subgroups in London, against the average year of arrival in Britain of people born in countries associated with those CEL Subgroups. The year of arrival information corresponds only to current residents in the London Borough of Camden who have been born abroad, sourced from the General Practice register of Camden Primary Care Trust, a funding partner of this PhD research. A caveat to take into consideration is that both the ID index and the average year of arrival are drawn from different populations (respectively London and Camden) and from different ontologies of ethnicity (respectively name-based and country of birth). Despite this difference, Figure 8.2 shows that there is a positive relation between year of arrival and the index of dissimilarity, and although the linear regression R^2 is 0.336, it initially validates the hypothesis of length of residence as an additional factor, together with population size, explaining differences in the level of segregation between CEL Subgroups measured by the ID index.

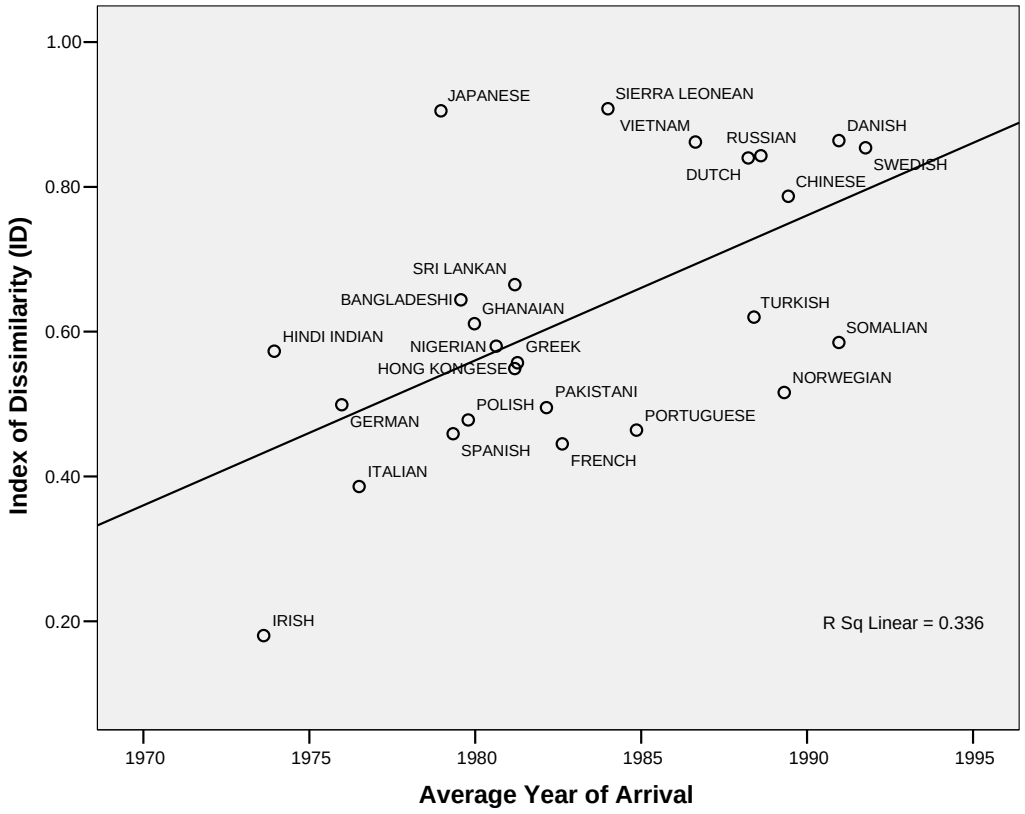

Figure 8.2: Index of dissimilarity vs. average year of arrival in Britain
This scatterplot shows on the vertical axis the index of dissimilarity of 26 CEL Subgroups in London, against the average year of arrival in Britain on the horizontal axis. The year of arrival information corresponds only to current residents in the London Borough of Camden who have been born abroad. Country of birth has been matched to their associated CEL Subgroup. Source: General Practice register, Camden Primary Care Trust.

8.2.3. Exposure

Exposure measures the degree of potential contact, or physical interaction, between two groups *within* geographic areas of a city, by virtue of sharing a common area of residence (Massey and Denton, 1988). The index of exposure most widely used is the index of isolation P* initially proposed by Skevy and Williams (1949), modified by Bell (1954) and popularised by Lieberson (1981). The version of the isolation index calculated here is ${}_xP_x^*$, which measures the probability of a member of ethnic group X entering into contact with a member of the same group within an area of residence, in this case an Output Area in London.

$$ {}_{x}P_{x}^{*} = \sum_{i=1}^{n} \left[\frac{x_i}{X} \right] . \left[\frac{x_i}{t_i} \right] \quad \text{(Bell, 1954) (Lieberson, 1981)} $$

(See Section 8.2.1 for explanation of variables)

The name 'index of isolation' is rather unfortunate, since a high value of this index means a high probability of finding a member of the same ethnic group living in the same area, that is being 'highly exposed', but not necessarily that this group is isolated from itself or other groups in surrounding areas. The results of the calculation of this index of isolation are shown in Table 8.4 following the same layout as described in Table 8.3.

Rank	CEL Subgroup	Total Pop.	P*	Rank	CEL Subgroup	Total Pop.	P*
1	ENGLISH	2,876,980	0.587	24	PAKISTANI KASHMIR	32,061	0.019
2	SIKH	83,968	0.168	25	FRENCH	40,264	0.019
3	BANGLADESHI	72,829	0.150	26	POLISH	33,270	0.018
4	HINDI INDIAN	156,269	0.137	27	HINDI NOT INDIAN	12,643	0.016
5	IRISH	414,038	0.093	28	SOMALIAN	20,376	0.015
6	PAKISTANI	140,548	0.085	29	NORWEGIAN	24,927	0.014
7	JEWISH	35,984	0.074	30	JAPANESE	3,469	0.013
8	SCOTTISH	323,847	0.073	31	MUSLIM SOUTH ASIAN	11,380	0.013
9	GREEK	61,296	0.062	32	CHINESE	8,874	0.013
10	NIGERIAN	68,596	0.058	33	BLACK CARIBBEAN	11,554	0.013
11	WELSH	222,429	0.052	34	SIERRA LEONIAN	3,854	0.012
12	SRI LANKAN	39,269	0.046	35	SOUTH ASIAN OTHER	8,484	0.012
13	TURKISH	34,359	0.033	36	IRANIAN	4,761	0.012
14	VOID	90,715	0.030	37	BALKAN	9,035	0.012
15	GHANAIAN	35,255	0.029	38	EUROPEAN OTHER	9,091	0.010
16	ITALIAN	71,967	0.029	39	RUSSIAN	5,539	0.010
17	VIETNAM	8,415	0.026	40	AFRICAN	4,879	0.010
18	MUSLIM MIDDLE EAST	48,114	0.025	41	SWEDISH	5,155	0.010
19	HONG KONGESE	35,609	0.025	42	UNKNOWN NAME	10,546	0.009
20	INDIA NORTH	31,888	0.024	43	DANISH	4,592	0.009
21	PORTUGUESE	44,780	0.023	44	DUTCH	5,477	0.009
22	SPANISH	44,679	0.022	45	AFRIKAANS	3,036	0.008
23	GERMAN	33,264	0.019	46	INTERNATIONAL	6,214	0.007

Table 8.4: Index of Isolation (P*) by CEL Subgroups in London at Output Area level
P*; Index of Isolation (Lieberson, 1981), Rank; Rank ordered by the P* index in descending order. The table only lists the most frequent 46 CEL Subgroups (those with a total population size in London greater than 3,000 people) ranked by the index of isolation (P*). The total population size of each CEL Subgroup in London is also listed. See text for explanation of 'International', 'Void' and 'Unknown Names' categories.

As expected, the most exposed group by far is the English group, since it is the majority population and its members have the highest probability of meeting each other in the same Output Area of residence. The next three more exposed groups are; Sikh, Bangladeshi and Hindi Indian, which are the groups usually picked up by the segregation literature about London (see for example Brimicombe, 2007). This means that members of these three ethnic groups are more likely to find someone from their own ethnic group in the Output Areas where they live than of any other ethnic minority. Furthermore, the fact that the analysis performed here, that uses name-based ethnicity from electoral registration records, gives such a similar result to the findings of other researchers using Census data, is in a sense another way of validating the methodology presented in this thesis.

The index of isolation is by definition correlated with the size of the group, in this case positively correlated, however not as much as the index of dissimilarity. The scatterplot in Figure 8.3 shows this relationship between P* and population size, but the linear regression R^2 is 0.517 – suggesting a much weaker over-all fit than that of the index of dissimilarity (R^2 is 0.805). Apart from the three CEL Subgroups already mentioned (Sikh, Bangladeshi and Hindi Indian), there are some others that have strikingly high values of P* relative to what might be expected given their population size. These include Jewish, Vietnamese, Japanese and Swedish CEL Subgroups. On the other hand, CEL Subgroups which are less exposed than might be expected given their population sizes are the Irish, Scottish, Welsh, Ghanaian, Muslim (Middle East), Spanish, and Portuguese.

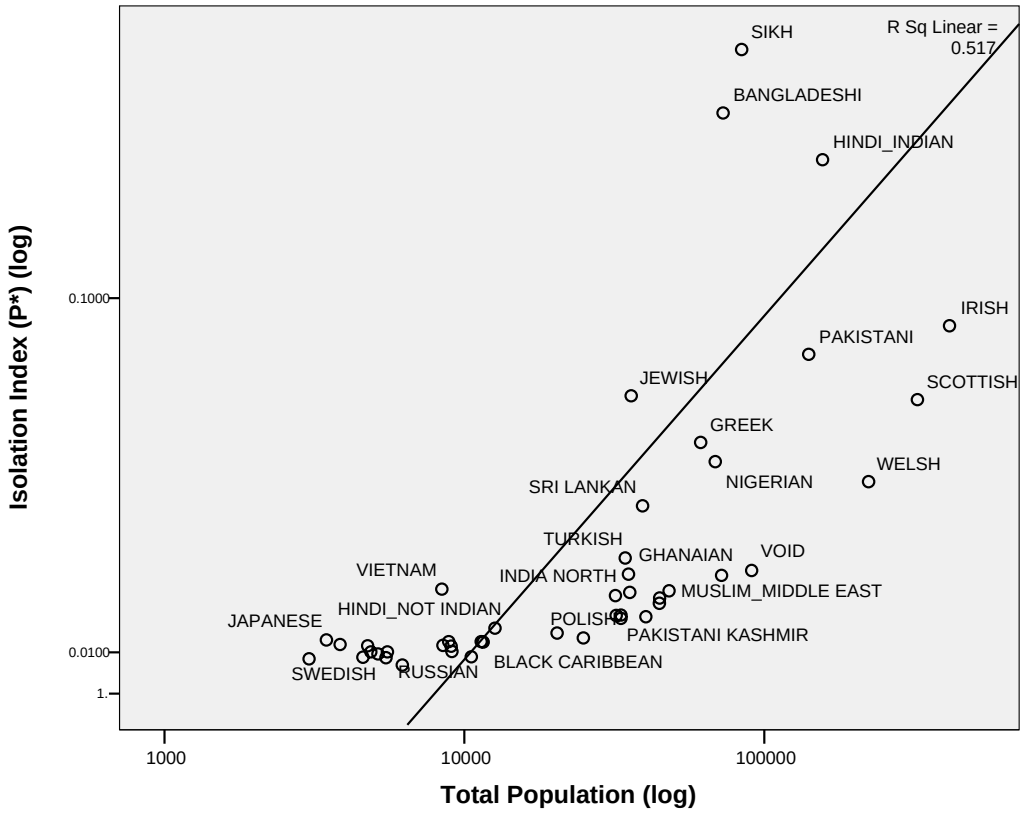

Figure 8.3: Scatterplot of CEL Subgroups Index of Isolation (P*) at Output Area level vs. their total population size in London
This scatterplot only includes the most frequent 46 CEL Subgroups, each with a total population size in London greater than 3,000 people. The 'English' CEL Subgroup is an outlier and falls outside the plotting area: it has been omitted from the plot for ease of visual interpretation. The Index of Isolation P* (Lieberson, 1981) calculated at Output Area level is represented on the vertical axis and the total population size of the CEL Subgroup in London on the horizontal axis. A trend line between the points is plotted using a linear fit, the R^2 of which is 0.517, showing a quite a strong relationship between P* and population size.

8.2.4. Concentration

Concentration refers to the relative amount of physical space occupied by a group in a city. The index of absolute concentration ACO was proposed by Massey and Denton (1988) (see formula in Section 8.2), and computes the total area inhabited by a group, and compares this figure with the minimum and maximum spatial concentration that could be inhabited by the group in a given city or area.

$$ACO = 1 - \left\{ \frac{\left[\sum_{i=1}^{n} \left(\frac{x_i A_i}{X} \right) - \sum_{i=1}^{n} \left(\frac{t_i A_i}{T_1} \right) \right]}{\left[\sum_{i=n_2}^{n} \left(\frac{t_i A_i}{T_2} \right) - \sum_{i=1}^{n_1} \left(\frac{t_i A_i}{T_1} \right) \right]} \right\}$$ (Massey and Denton, 1988)

(Spatial units are sorted by area size in ascending order. See Section 8.2.1 for explanation of variables.)

The maximum spatial concentration is reached when all members of the group live in the smallest space possible (i.e. in just one or very few of the smallest spatial units), while the minimum spatial concentration correspond to a situation where the members of the group live in the largest spatial units in the city. The ACO index varies from 0 to 1, where a score of 1 indicates that the group experiences the maximum spatial concentration possible (all members live in the smallest spatial units), and a score of 0 the minimum spatial concentration possible, in other words, the maximum deconcentration possible.

The results of this index are rather deceptive, since all CEL Subgroups obtain very similar and high values of ACO, except for the British Isles CELs (English, Welsh, Scottish, Irish). If these four CEL Subgroups are excluded (which respectively have ACO values of 0.396, 0.900, 0.882, and 0.867), the mean ACO for the remaining 42 Subgroups is 0.977 with a standard deviation of 0.015. This result might suggest that they all present a highly concentrated spatial pattern, but in reality it is an artefact of applying the ACO index to a large number of fine ethnic groups that are spread over a large number of small areas. The ACO index was designed to measure binary situations in US cities between a white majority and a Non-White minority, at census tract level (average size 4,000 people), where in this example there are 66 groups and

spatial units which average 285 people (OAs). Furthermore, OAs are by definition homogeneous in population size, and hence large differences between the densities of the areas studied, in an urban area like London, are highly unlikely. No other alternative spatial concentration index is available in the literature that is designed for such situations.

8.2.5. Clustering (I): the sociological approach

The clustering dimension measures the degree to which members of a group inhabit areas which are contiguous and closely packed, that is, if their geographical distribution presents a clustered pattern. There are several measures of clustering in the geographical literature, which are extensions to the 'checkerboard problem' (Geary, 1954), but in the first instance an index from the sociological literature will be computed here, namely the absolute clustering index (ACL) (Massey and Denton, 1988).

$$ACL = \left[\frac{\sum_{i=1}^{n} \left(\frac{x_i}{X} \right)}{\sum_{j=1}^{n} c_{ij} x_j} - \frac{X}{n^2 \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}} \right] \Bigg/ \left[\frac{\sum_{i=1}^{n} \left(\frac{x_i}{X} \right)}{\sum_{j=1}^{n} c_{ij} t_j} - \frac{X}{n^2 \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}} \right] \quad \text{(Massey and Denton, 1988)}$$

(See Section 8.2.1 for explanation of variables)

The absolute clustering index ACL (Massey and Denton, 1988), expresses the average number of members of a group in neighbouring spatial units as a proportion of the total population in those neighbouring units (see formula in Section 8.2). It varies from a minimum of 0 (low clustering) to a maximum that approaches but never equals 1 (high clustering).

Rnk	CEL Subgroup	Total Pop.	ACL	Rnk	CEL Subgroup	Total Pop.	ACL
1	ENGLISH	2,876,980	0.235	24	POLISH	33,270	0.006
2	SIKH	83,968	0.149	25	PAKISTANI KASHMIR	32,061	0.006
3	HINDI INDIAN	156,269	0.106	26	FRENCH	40,264	0.006
4	BANGLADESHI	72,829	0.106	27	HINDI NOT INDIAN	12,643	0.005
5	PAKISTANI	140,548	0.055	28	SOMALIAN	20,376	0.005
6	JEWISH	35,984	0.052	29	JAPANESE	3,469	0.004
7	GREEK	61,296	0.042	30	SOUTH ASIAN OTHER	8,484	0.004
8	NIGERIAN	68,596	0.028	31	BLACK CARIBBEAN	11,554	0.003
9	SRI LANKAN	39,269	0.028	32	SIERRA LEONIAN	3,854	0.003
10	IRISH	414,038	0.021	33	MUSLIM SOUTH ASIAN	11,380	0.003
11	SCOTTISH	323,847	0.016	34	BALKAN	9,035	0.003
12	TURKISH	34,359	0.016	35	CHINESE	8,874	0.003
13	INDIA NORTH	31,888	0.012	36	IRANIAN	4,761	0.003
14	GHANAIAN	35,255	0.011	37	NORWEGIAN	24,927	0.003
15	WELSH	222,429	0.010	38	SWEDISH	5,155	0.003
16	VIETNAM	8,415	0.010	39	AFRICAN	4,879	0.002
17	VOID	90,715	0.010	40	RUSSIAN	5,539	0.002
18	ITALIAN	71,967	0.009	41	EUROPEAN OTHER	9,091	0.002
19	MUSLIM MIDDLE EAST	48,114	0.008	42	DANISH	4,592	0.002
20	PORTUGUESE	44,780	0.008	43	AFRIKAANS	3,036	0.002
21	HONG KONGESE	35,609	0.007	44	UNKNOWN NAME	10,546	0.002
22	GERMAN	33,264	0.007	45	DUTCH	5,477	0.002
23	SPANISH	44,679	0.007	46	INTERNATIONAL	6,214	0.001

Table 8.5: Absolute Clustering Index (ACL) by CEL Subgroups in London at Output Area level
ACL; Absolute Clustering Index (ACL) (Massey and Denton, 1988), Rnk; Rank ordered by ACL in descending order. The table only lists the most frequent 46 CEL Subgroups (those with a total population size in London greater than 3,000 people) ranked by ACL. The total population size of each CEL Subgroup in London is also listed. See text for explanation of 'International', 'Void' and 'Unknown Names' categories.

The results for the calculation of the ACL index in London are shown in Table 8.5, which only lists the most frequent 46 CEL Subgroups. The most clustered group is again the English CEL, since it has most neighbours of its own group, followed by the Sikh, Hindi Indian, and Bangladeshi groups. The spatial clustering of these three groups has been persistently identified in the recent segregation literature on London (Brimicombe, 2007; Peach, 2006). Following these groups in the clustering ranking are the Jewish and Greek groups. Again the Jewish case has been repeatedly reported in the literature (Brimicombe, 2007; Peach, 2006), but the Greek group has not been studied before since it is not measured separately from the 'White Other' ethnic

group or the Christian religion in the UK Census. The Greek group has already been highlighted as having a segregated pattern in the indices previously described, and presents an example of the advantages of using a name-based classification in segregation studies, which will be further discussed later in this section. Amongst the less clustered groups, there are several Nordic CELs (Norwegian, Swedish, and Danish), some European small groups (Dutch, Russian, 'European Other'), African and Afrikaans, and finally the Unknown Name and International Names groups, which is reassuring to find at the bottom of the clustering table since they might be expected to share no common characteristics.

The relationship between ACL and group population size is still positive but very weak, as can be seen in the scatterplot between the two variables presented in Figure 8.4. The R^2 of the linear regression is 0.418, a consequence of the wide range of outliers in this linear relationship. However, this relationship seems to hold true for CEL Subgroups above a threshold level of a total population size of approximately 60,000 people, while below it, the ACL index barely grows with population size (bottom left part of Figure 8.4). This is a consequence of the population size effect discussed above, since below the 60,000 threshold there are fewer than 2.5 people per Output Area on average, and the mechanics of the indices applied here were not designed for such small concentrations of people per unit area.

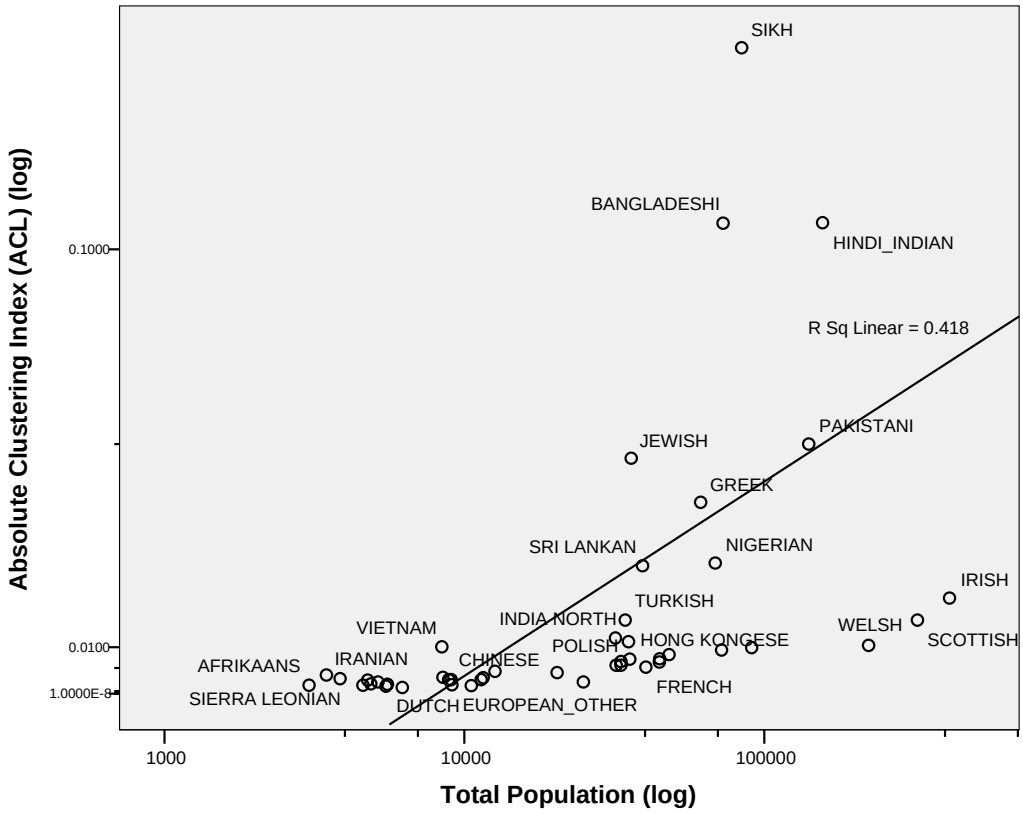

Figure 8.4: Scatterplot of CEL Subgroups' Absolute Clustering Index (ACL) at Output Area level vs. their total population size in London
This scatterplot only includes the most frequent 46 CEL Subgroups with a total population size in London greater than 3,000 people. The 'English' CEL Subgroup is an outlier and falls outside the plotting area. It has been omitted from the plot for ease of visual interpretation. The Absolute Clustering Index (ACL) (Massey and Denton, 1988), calculated at Output Area level, is represented on the vertical axis and the total population size of the CEL Subgroup in London on the horizontal axis. A trend line between the points is plotted using a linear fit, the R^2 of which is 0.418.

8.3. Additional Dimensions and Approaches to Measuring Residential Segregation

In the previous section the most commonly used indices to measure four of the five traditional dimensions of residential segregation (Massey and Denton, 1988) were reviewed and applied to the CEL-classified Electoral Register for London. In this section two additional aspects of residential segregation will be separately measured: spatial clustering of ethnic groups using a geographical approach, and the degree of

diversity of areas, using an index of entropy. These two measures complement the four indices already presented since they represent aspects not adequately reflected by the previous measures, more precisely; measures are not global/spatially invariant across the study area, and that focus on the over-all ethnicity composition of each neighbourhood rather than separately on each particular ethnic group.

8.3.1. Clustering (II): the geographical approach

An alternative view of segregation can be achieved by using spatial autocorrelation statistics, which measure the tendency of similar values to cluster together in space (Goodchild, 1986). Therefore, it seems pertinent to apply such measures to study residential segregation from a geographical analysis perspective, as has been proposed by some authors (Owen, 2006). The most widely accepted measures of spatial autocorrelation are Moran's I and Geary's C, which at their simplest are global measures providing a value for the whole study area (Fotheringham et al, 2000). A spatially variable measure of autocorrelation is preferred here to measure differences between areas. One particular instance are local indicators of spatial autocorrelation (LISA) (Anselin, 1995) such as the Local Moran statistic:

$$I_i = z_i \sum_j w_{ij} z_j \quad \text{(Anselin, 1995)}$$

where the observations z_i and z_j are given in standard deviations from the mean $\left[z_i = (X_i - \overline{X}); z_j = (X_j - \overline{X})\right]$, and the summation over j is such that only neighbouring values are included. Neighbourhood is defined by a weight matrix w_{ij} representing contiguity, which in this application represents binary adjacency (1 adjacent and 0 non-adjacent) between the i^{th} and j^{th} points (0 or 1) – other definitions of neighbourhood may also be accommodated.

The Local Moran statistic was calculated for the OAs in London and the 66 CEL Subgroups, using *GeoDa,* an exploratory spatial data analysis (ESDA) software tool (Anselin and Regents of the University of Illinois, 2004). The weights matrix was defined using a *Rook* adjacency criterion taking into account both first and second order neighbourhoods (a window of area's immediate areal units comprising immediately adjacent neighbours plus zones adjacent to these neighbours).

The purpose of using the Local Moran's statistic is to investigate and identify local clusters of spatial autocorrelation. In the analysis performed here the purpose is to identify the areas within London of highest and lowest clustering of each CEL Subgroup. While the value of Moran's *I* varies between -1 and 1, indicating the range from strong negative autocorrelation to strong positive autocorrelation (in a similar fashion to the correlation coefficient), the value range of the Local Moran has no particular bounds. Values range from a negative figure to a positive figure for each spatial unit, indicating strong negative autocorrelation to strong positive autocorrelation. However, the amount of correlation is given in relative terms denoting variation in spatial autocorrelation at local level, and its final value depends on the immediate neighbouring values whose weighted average difference from the mean is built into the final value. Therefore the most appropriate scale to interpret the final LISA results is to create a relative classification of each areas' local autocorrelation. In the analysis reported here, the results of the Local Moran *I* statistic were represented in a choropleth map for the most significant CEL Subgroups (p values < 0.05) classifying all output areas into five types of spatial correlation, following Anselin (2004):

- *High-high*; output areas with high proportions of people from the CEL Subgroup next to areas with similar values.
- *Low-low*; output areas with low proportions of people from the CEL Subgroup next to areas also with similar values.
- *High-low*; output areas with high proportions of people from the CEL Subgroup next to areas with low values.
- *Low-high*; output areas with low proportions of people from the CEL Subgroup next to areas with high values.
- *No clustering*; output areas with no significant LISA, and thus whose p-values > 0.05

In this scale, 'high' values are statistically significant ($p < 0.05$) and positive LISAs while low values are negative and significant. The high-high and low-low adjacency types suggest clustering of similar values, whereas the high-low and low-high locations indicate spatial outliers (i.e. they represent departures from uniformity in spatial distribution, hence areal differentiation at the scale of the mapped areal units). 22 out of the total 66 CEL Subgroups were selected representing the third with a larger number of high clustering Output Areas in London. The 22 maps of the five types of local clustering of LISA are shown in Figure 8.5, Figure 8.6, Figure 8.7 and Figure 8.8. These maps use the following colour scheme: bright red for the high-high association, bright blue for low-low, light blue-purple for low-high, light red-pink for high-low, and white for areas with no clustering.

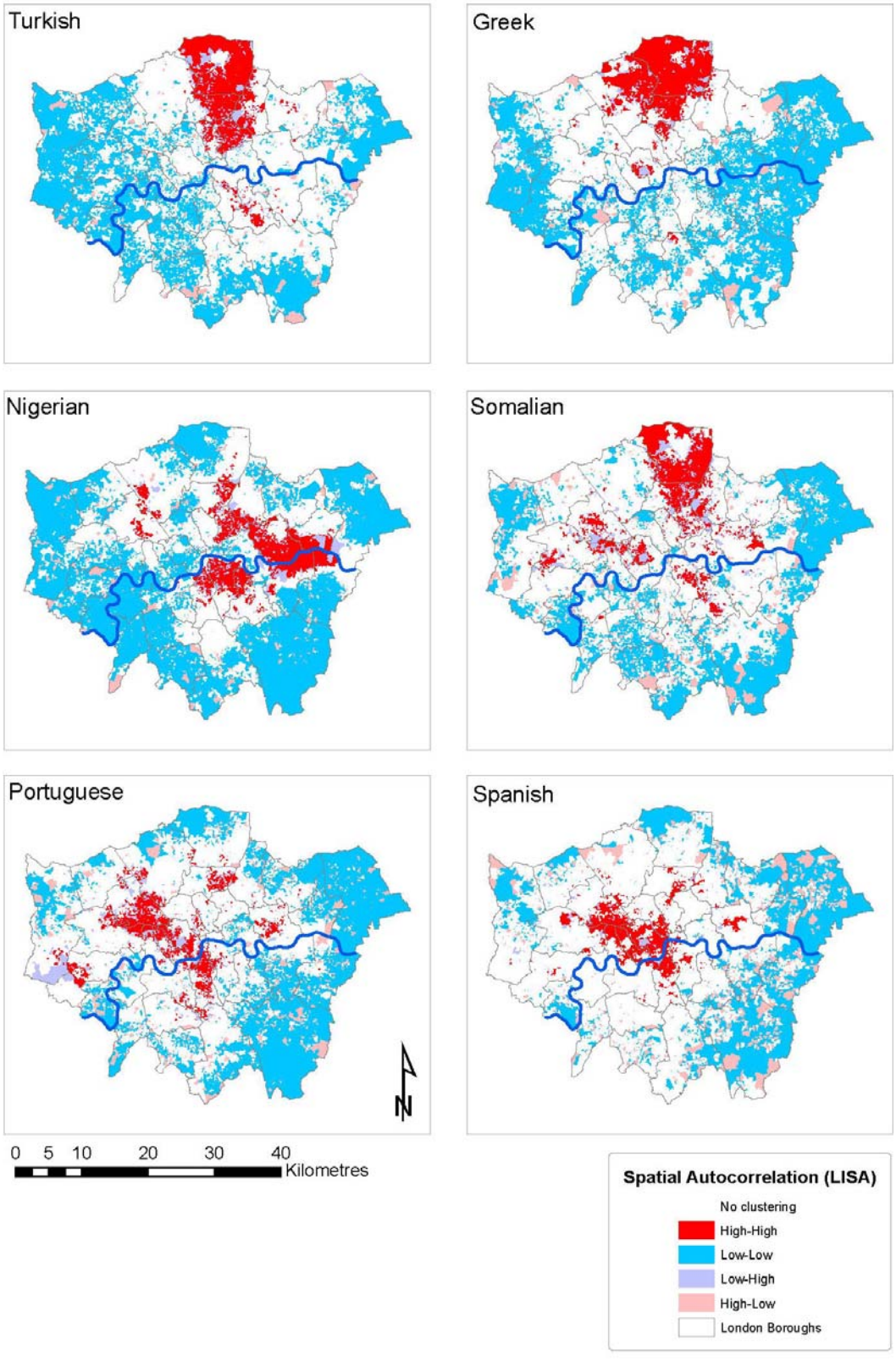

Figure 8.5: Maps of local indicators of spatial autocorrelation (LISA): Turkish, Greek, Nigerian, Somali, Portuguese and Spanish CELs

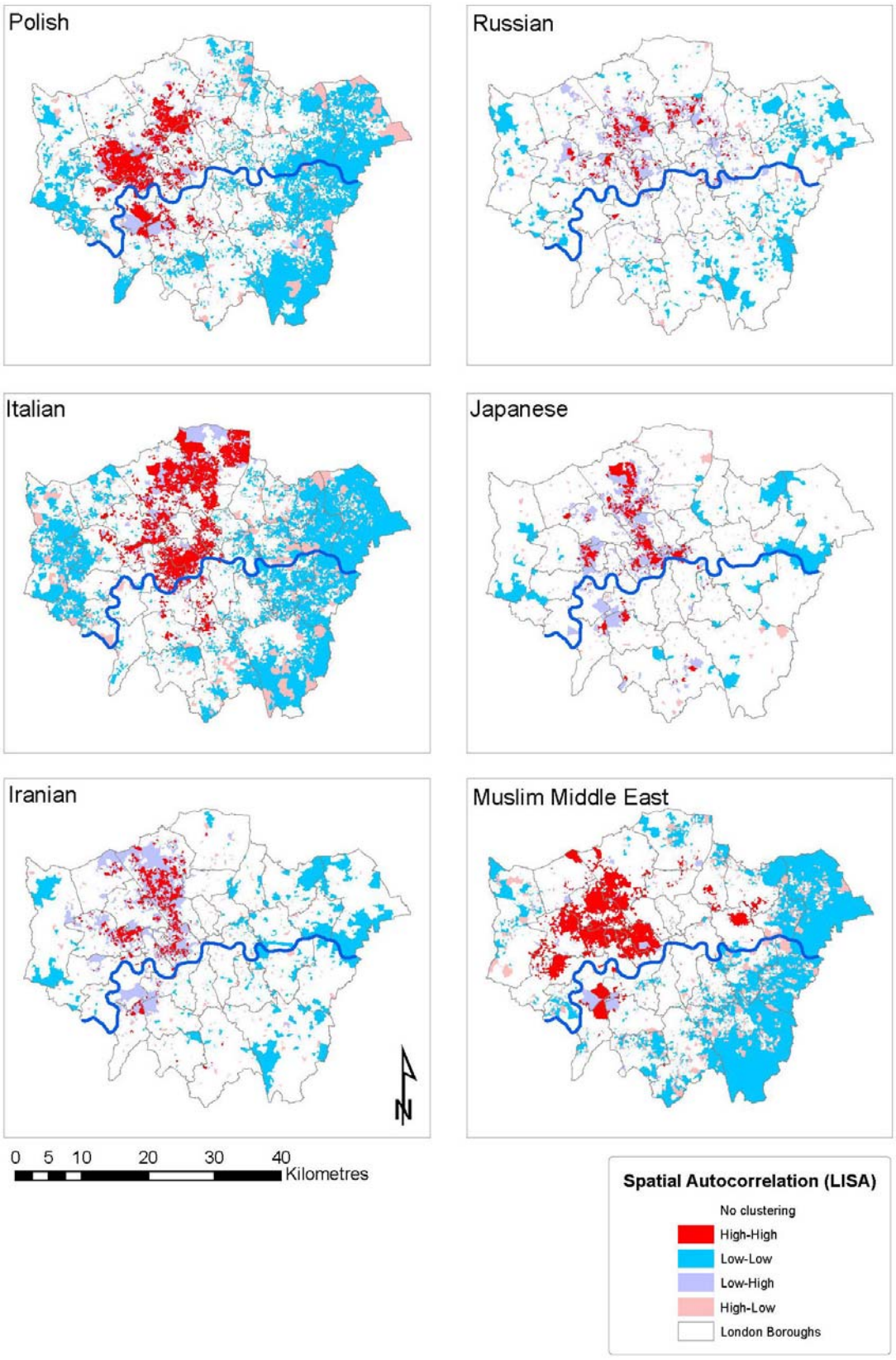

Figure 8.6: Maps of local indicators of spatial autocorrelation (LISA): Polish, Russian, Italian, Japanese, Iranian and Muslim Middle East CELs

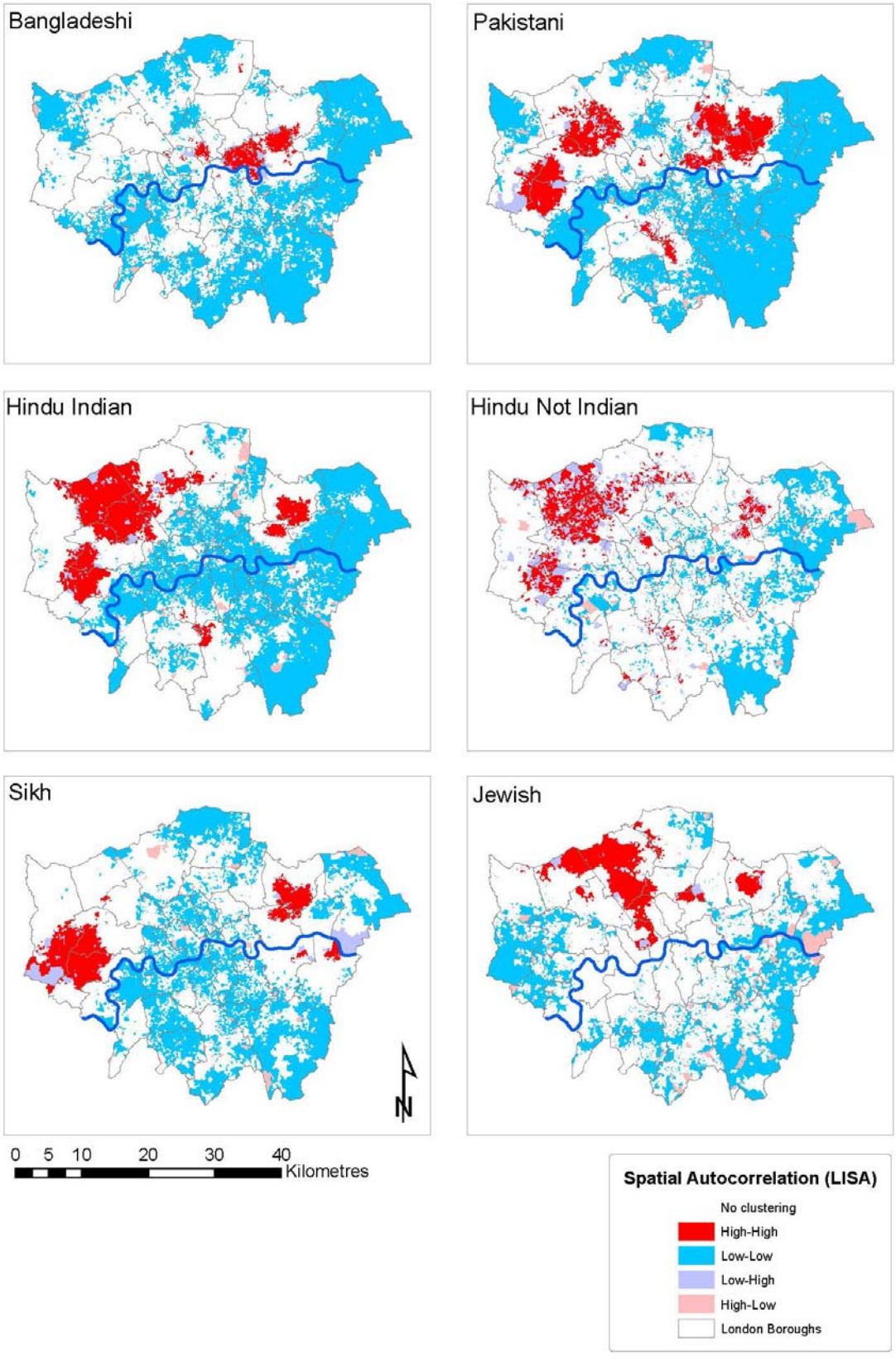

Figure 8.7: Maps of local indicators of spatial autocorrelation (LISA): Bangladeshi, Pakistani, Hindu Indian, Hindu Not Indian, Sikh and Jewish CELs

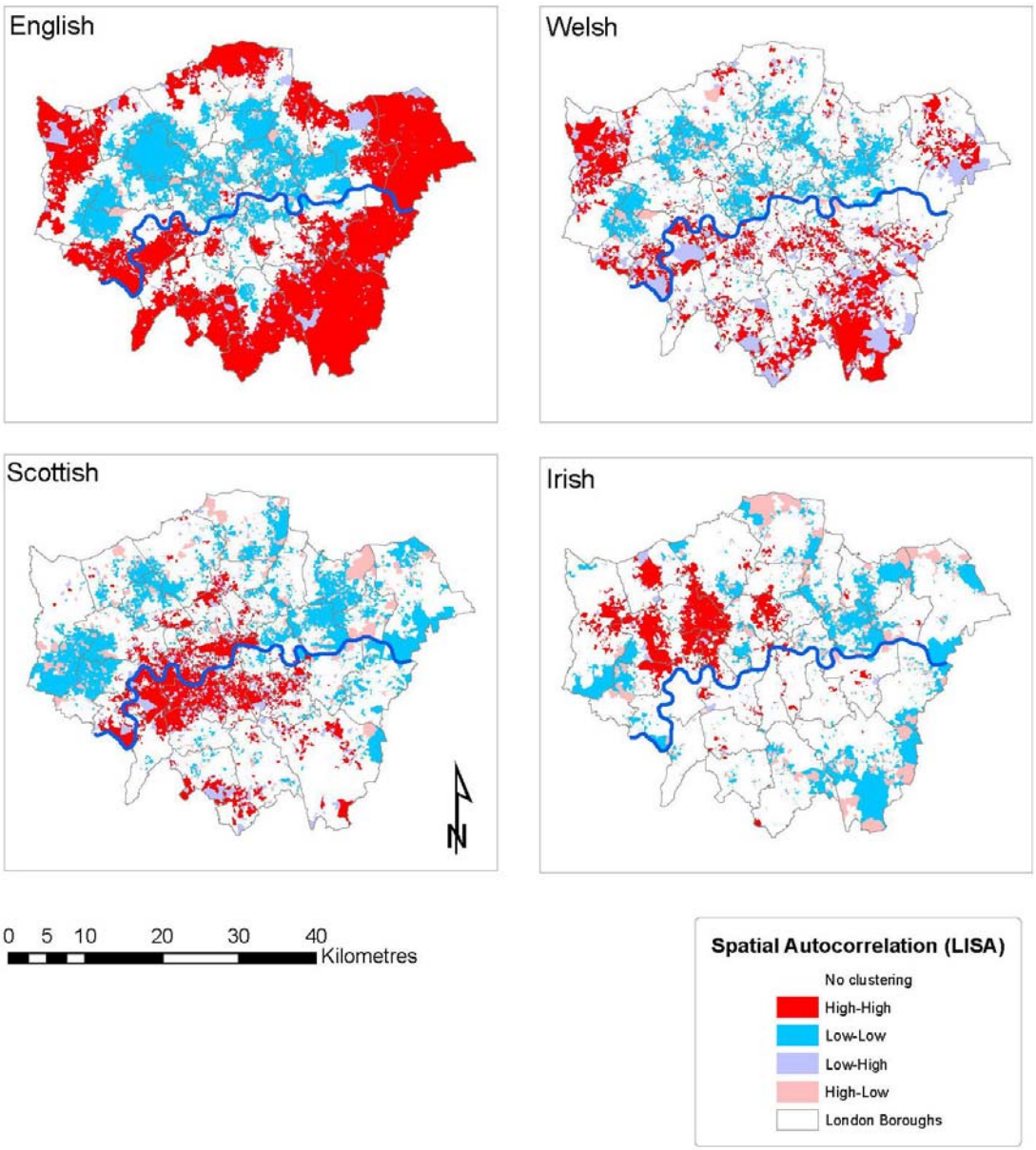

Figure 8.8: Maps of local indicators of spatial autocorrelation (LISA): English, Welsh, Scottish and Irish CELs

These 22 maps show the unique patterns of geographical distribution of these CEL Subgroups, summarised by the areas where each of them is most or least clustered. A summary of some of the most evident features of the clustering patterns will be commented here, stressing the value of the name-based technique adopted here as opposed to the results that would have been obtained using just Census ethnicity data.

Figure 8.5 and Figure 8.6 show twelve clustering maps for ethnic groups that are not separately reported in the UK 2001 Census ethnicity classification; Turkish, Greek, Nigerian, Somali, Portuguese, Spanish, Polish, Russian, Italian, Japanese, Iranian, and Muslim Middle East. To the author's knowledge this is the first time that these fine groups have been mapped in London using a universal register, such as the Electoral Register, and a broad definition of ethnic origin, as opposed to country of birth data which is common in the literature (Peach, 1999a). These maps show the unique spatial clustering patterns of each CEL, in which each group seems to occupy a distinct set of areas within the city. However, of these twelve groups, eleven appear to cluster in an area comprising approximately a third of the Capital's total area, in what constitutes the Northwest third of the whole city, from the North-Central to the Southwest bounds of the city (approximately postal areas N, NW, WC, W, WC, EC and the west of SW). The exception is the Nigerian CEL which is predominantly clustered in the East of London on both sides of the river, following historic settlement areas of Black Africans in London.

It is surprising to notice the degree of overlap between areas of high clustering of Turkish and Greek names in North London, perhaps indicative of the cultural closeness of these groups when they live abroad despite their historical grievances at home. However, Greeks are more distributed towards the northern periphery of London, especially in and around the Boroughs of Enfield and Barnet, while Turks are more concentrated in Inner London, especially in Hackney and Haringey, sharing Enfield with Greeks.

Output Areas where Somali names are most clustered are found in several parts of the city, probably because of the sparse availability of public housing into which this community was originally accommodated following the refugee arrivals from the Horn of Africa in the early 1990s. A bigger cluster in Haringey and Enfield can also be discerned.

Portuguese and Spanish names clusters share a clear common pattern of settlement in West London that spreads throughout the Boroughs of Brent, Ealing, Chelsea and Kensington, Westminster, and Lewisham. This reveals the commonalities in cultures and preferences between Spanish speaking and Portuguese speaking communities in London, which comprise people originating in over 25 countries in Latin America, the Iberian Peninsula and some African countries. The spread over very affluent and less affluent areas of inner west London suggests a diverse range of socio-economic backgrounds of members of these CEL groups. Further analysis of these differences using postcode unit level names data in combination with geodemographic classifications would shed light upon these local differences.

The Polish CEL (Figure 8.6) is highly clustered in the Boroughs of Ealing and Barnet with some other smaller clusters in West and Southwest London. The version of the Electoral Register used for this analysis is from 2004, and hence the pattern revealed here is three years old at the time of finishing this thesis. However, it is known that the Polish ethnic group has been one of the fastest growing in Britain since Poland joined the EU in May 2004, with immigration from Poland from May 2004 to March 2007 estimated in 294,000 workers plus their families (Border and Immigration Agency et al, 2007). Therefore, it would be very interesting to repeat this clustering

exercise with a current version of the Electoral Register or even better with a patient register, in order to see how these geographical patterns have changed in London. As regards the clustering of Russian names, this group is much smaller than the Polish group and is concentrated in a number of hotspots scattered in inner Northwest London.

Italian names are clustered in several Boroughs in Central and North London, following a pattern of historic settlement of Italian communities in Central London and in Enfield. Clustering of Japanese and Iranian names follows a surprisingly similar pattern; concentrated in Westminster, Chelsea and Kensington, west of Camden, Barnet and east of Ealing. This similarity could be explained by the relative wealth of the areas where members of these two communities live. Finally, Muslim names associated with the Middle East, that is, generally with Arab language patterns, are highly concentrated across several Boroughs in the West of London.

The six maps in Figure 8.7 represent the CEL Subgroups associated with the most commonly reported ethnic minorities in the literature; Bangladeshi, Pakistani, Hindu Indian and Hindu Not Indian, Sikh, and Jewish CEL Subgroups. The local clusters of each of these groups correspond to the areas repeatedly identified in the literature using Census derived data (Johnston et al, 2002a; Owen, 2006; Peach, 2006). It is interesting to notice the way in which Pakistanis share common neighbourhoods with the Hindu Indian and Bangladeshi neighbourhoods that are themselves very segregated from each other. Given that these are ethnic group categories that are reported in the Census, and that were compared with the CEL classification through

the validation exercise described in Chapter 7, the results of both methodologies tell a very similar story in London.

Figure 8.8 includes the LISA maps of the British Isles CELs; English, Welsh, Scottish and Irish, whose degree of segregation is rarely analysed by the literature. The areas of high clustering of the English CEL are the reverse of the combined maps shown so far for non-British groups, and are mainly concentrated in the southeast and outer rim of London. This map clearly shows the result of a sort of 'centrifugal force' that hollows out Inner London of English names and clusters them in the outer suburbs, especially in the southeast. It is also interesting to notice specific clusters of Welsh and Scottish names in the west and southwest of London. The analysis of these last three ethnic groups constitutes an innovative type of analysis and findings since this information is not collected in official statistics. It could be argued perhaps that the Scots immediately north of the river Thames could be recent north-south migrants in rental housing areas, and some of those south of the river could just be Black Caribbeans with Scottish surnames (which are known to be very common in the Caribbean). The Welsh pattern seems to mirror the English clusters, and could be more dispersed because of small numbers.

Finally, the map of Irish clustering indicates areas of settlement of Irish migrants that might mirror Old Commonwealth immigration patterns in London, suggesting that there are still some less well established migrants from Ireland in London. However, as it was found in the validation exercise described in Chapter 7, Irish names were one of the two ethnic groups, together with Black Caribbeans, where name derived ethnicity vs. Census ethnicity presented a larger degree of mistmatch.This was

explained then by a difference in the perceptions of Irish identity between generations of people with Irish names. One aspect that is worth investigating in the future is comparing the areas of these two types of Irish identity self-identification to study the differentials between their demographic and migration profiles. Furthermore, anticipating a question about national identity in the 2011 Census, a similar type of future analysis for the rest of the British Isles CELs will be very illustrative of collective identity formation processes at local level.

8.3.2. Diversity

Beyond the five dimensions of segregation analysed in the previous section, and as was discussed in Section 2.2, it has been recognised that there are two other aspects related to the measurement of segregation; *movement* (Simpson, 2007), which analyses changes in segregation over time taking into account migration and demographic structure, and *diversity* (based on Edward Simpson, 1949), which measures how close a set of groups are to equal numbers within an area.

Since no temporal change data on names were available, the measurement of movement could not be calculated in this exercise (although this is an interesting avenue for future research in this direction). However, the measurement of diversity was added to the four indices previously described. An index of entropy or diversity, derived from the ecological literature (Simpson, 1949), was calculated to measure the level diversity of each Output Area, H, expressed by the number and size of ethnic groups as per the following formula (Thiel and Finezza, 1971):

$$H = -\sum_{i=1}^{n} \frac{\left(\frac{P_{ij}}{P_j}\right)\ln\left(\frac{P_{ij}}{P_j}\right)}{\ln n}$$

where:

n= number of groups

P_{ij}= Population of group *i* in spatial unit j

P_j= Sum of population of all groups 1 to *n* in spatial unit j

This index is sometimes known as the Multigroup Entropy Index, the Information Theory Index, or Theil's *H*. The values of *H* vary from 0 (no diversity) to 1 (maximum diversity), and there is single value for each of the areas, in this case each OA in London. The frequency distribution of the *H* index across all OAs, calculated in this analysis, is summarised in a histogram shown in Figure 8.9 that shows a near-to-normal shape of the frequency distribution, which is slightly negatively skewed (its skewness is -0.208). However, when the same results are mapped, as shown in Figure 8.10, systematic differences in OA diversity become very apparent. The map in Figure 8.10 confirms the aggregated results of the clustering processes unveiled by the previous maps for each individual CEL Subgroup, although here the number of groups rather than group size is driving the values of the diversity index. The areas of higher diversity are predominantly found in the northern half of London, with the Boroughs of Brent, Newham and Westminster leading the diversity league measured at OA level.

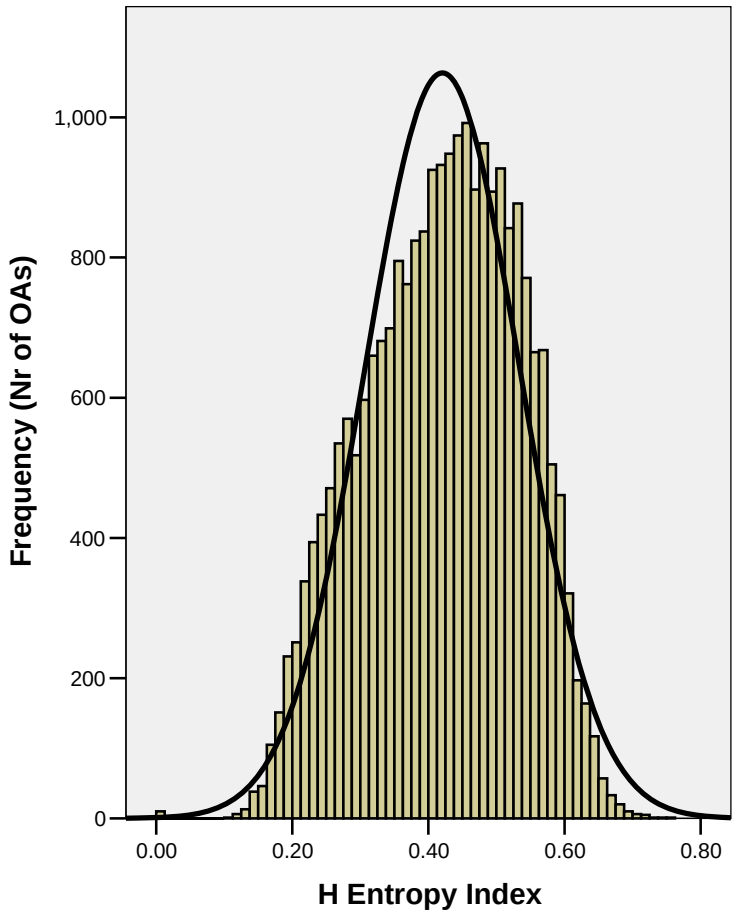

Figure 8.9: Frequency distribution of the *H* entropy index by OA in London
The histogram shows the frequency distribution of the H entropy index of diversity (Thiel and Finezza, 1971) by output area level in London, with each count representing one OA. A Normal distribution with mean $H = 0.4$ is included for reference purposes.

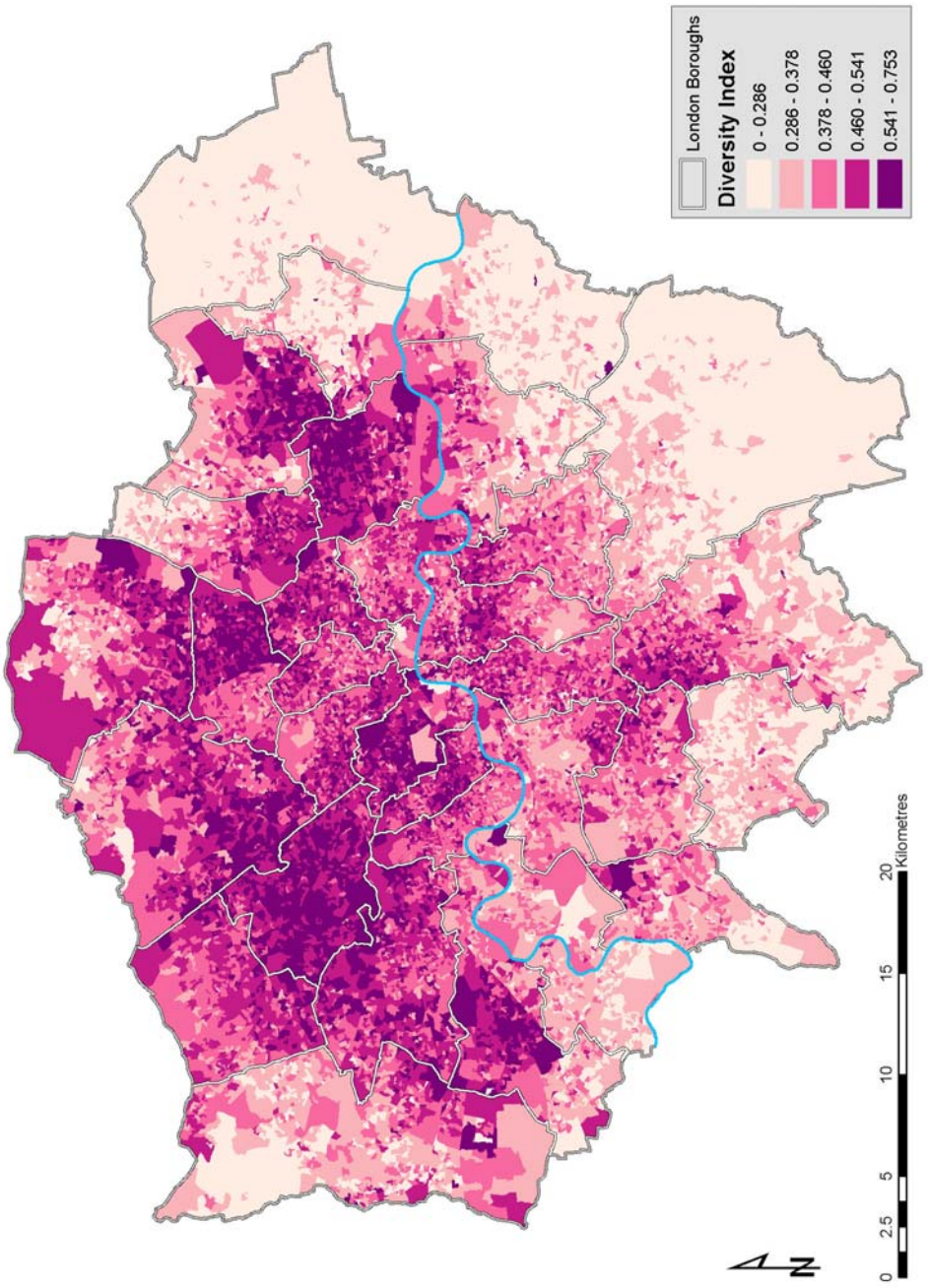

Figure 8.10: Map of ethnic diversity in London at Output Area level, measured by the Multigroup Entropy Index (*H*)

8.4. Discussion of Residential Segregation Results

8.4.1. Scale effect

As a result of the analysis carried out in the previous sections, the issue of the scale dependency of the indices has emerged in the calculation of most of them. The purpose of this section is to investigate the sensitivity of the main measures of segregation to changes in the geographical scale of measurement as well as to changes in the level of aggregation of the ethnic groups analysed. The index of dissimilarity (ID) is used here since it is deemed to be independent of the relative size of the ethnic group (Massey and Denton, 1988; Peach, 1996a), although it is influenced by the number of areas and the fineness of the grid used (Voas and Williamson, 2000). The objective is to compare the effect that changes in geographical scale and ethnic group unit definition have on the resulting ID index, using both the CEL dataset and the 2001 UK Census ethnicity data. The different geographical scales calculated were Output Area (OA), Lower Super Output Area (LSOA), Ward and London Borough levels. A summary of the number and sizes of geographical units at each of these scales is shown in Table 8.6

	OA	LSOA	Ward	Borough
Average Persons / Geographical Unit	285	1,443	10,931	208,011
Number of Geographical Units	24,100	4,758	628	33

Table 8.6: Summary of geographic units' characteristics

Firstly, the 66 CEL Subgroups were aggregated into a set of 17 aggregations of CEL groups in order to analyse the effect of a phenomena that could be termed the 'Modifiable Ethnic Unit Problem' (MEUP), drawing a parallelism with the

'Modifiable Areal Unit Problem' (MAUP) (Openshaw, 1984). Furthermore, this scale of analysis makes the CEL results more comparable with the Census dataset. These CEL groups were defined as follows; British (including English, Scottish and Welsh), Irish, Eastern European (including ex-communist countries), Spanish-Portuguese, Western Europe (the rest of Europe not included in the previous groups), Black Caribbean, Somali, African (including all other Black African CELs), Greek or Greek Cypriot, Jewish, Chinese, Japanese, Bangladeshi, Pakistani, Hindu (all Hindu CELs), Sri Lankan, Sikh, and Other Muslim (Muslim CELs not included in the rest). Calculations of the index of dissimilarity (ID) were made for each of these 17 CEL groups at each of the four geographical levels Output Area (OA), Lower Super Output Area (LSOA), Ward and London Borough.

The ID index was also calculated for the Census 2001 ethnic groups (Key Statistics KS06 table) for the 33 London Boroughs at Output Area (OA) level (comprising 7,158,904 Census respondents and 24,100 OAs), and higher geographies (LSOA, Ward and Borough). The detailed characteristics of this Census dataset were described when explaining the evaluation of the CEL methodology in Section 7.4 and repetition is avoided here. The Census ethnicity dataset is the main source of ethnicity information used in the literature to calculate indices of segregation, so here the intention is to compare it with the results using the CEL classification in order to highlight the advantages of the methodology presented in this thesis.

The 'radar' charts shown in Figure 8.11 and Figure 8.12 represent a graphical comparison of the ID index for both the Census and the CEL datasets at each of the four geographical scales; OA, LSOA, Ward and Borough. As expected, the level of

segregation increases as the size of the geographical unit is reduced (Wong, 2004), although the strength of this scale effect shows substantial variations by ethnic group. If segregation were to increase with decrease in the size of geographical units in the same way for each of the groups, all of the lines in Figure 8.11 and Figure 8.12 would look like parallel concentric rings. However, in the Census-based Figure 8.11, all the 'Mixed' ethnic groups are much more segregated at OA than at LSOA level. A similar difference is noticeable in the CEL-based Figure 8.12, for the Eastern European, Pakistani, Black Caribbean, and Irish CEL groups. Therefore, these groups show processes of more pronounced segregation at smaller geographical units.

Another aspect worth mentioning is the relatively homogeneity of values in the index of dissimilarity measured at the coarser scales, i.e. the Ward and the Borough levels. These present very smooth profiles of segregation across ethnic groups. This finding is surprising since these are the geographical scales at which most of the segregation studies in Britain are based (Johnston et al, 2002a; Peach, 2006; Simpson, 2005a).

Moreover, the advantage of the much finer CEL categories is apparent in Figure 8.12, which reveals the differential patterns of residential segregation between finely defined ethnic groups. For example, the Greek group's index of dissimilarity at OA level is nearly double (0.55) that of Western Europeans (0.3). In general the CEL dataset produces a more segregated pattern than the Census for the same areal units, because of its much finer ethnic group categories and the consequent more intricate representation of underlying segregation patterns.

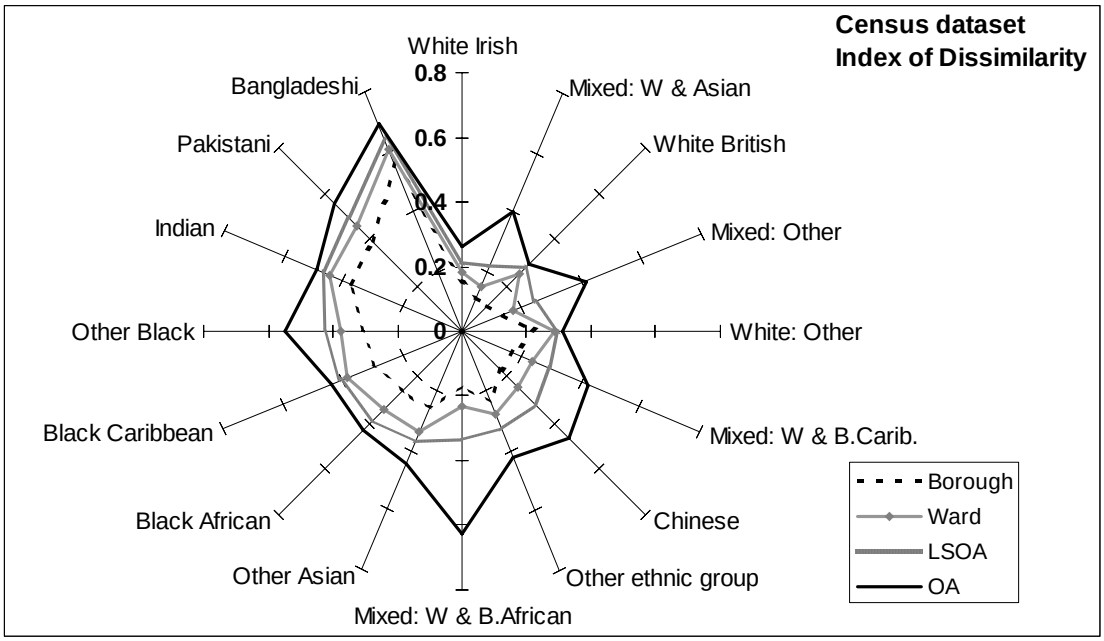

Figure 8.11: Index of dissimilarity of the Census dataset at four different geographical scales

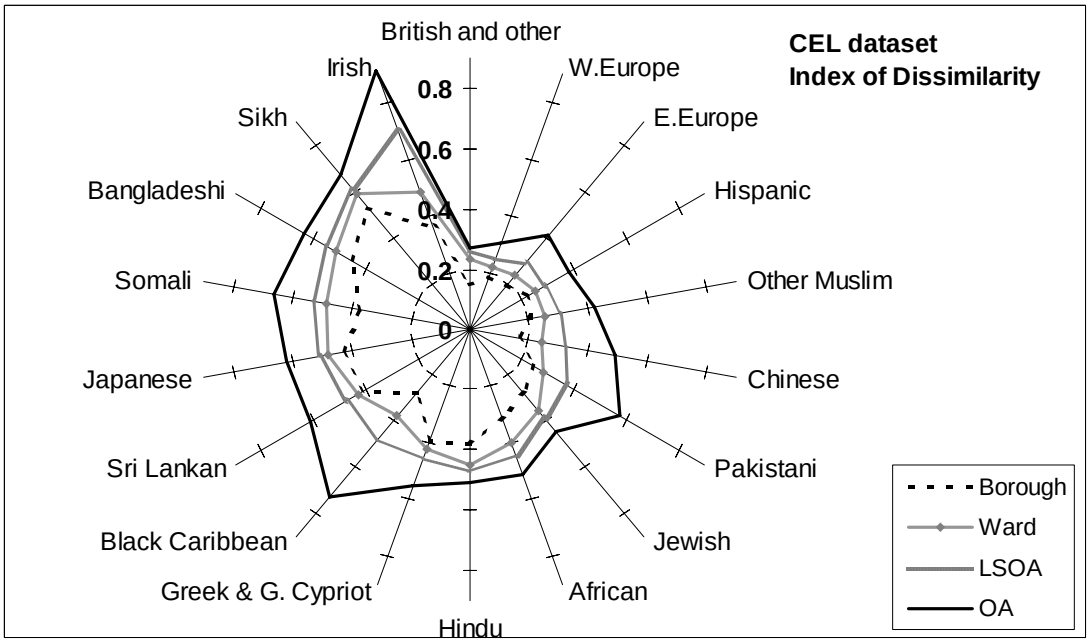

Figure 8.12: Index of dissimilarity of the CEL dataset at four different geographical scales
Both figures represent the index of dissimilarity ID (Duncan and Duncan, 1955)calculated for the Census (Figure 8.11) and CEL (Figure 8.12) datasets at four different geographical scales. The ethnic categories are ordered by their average ID value, showing increasing segregation in a clockwise direction from '12 o'clock'.

Furthermore, changes in the ontology of ethnicity can have a significant effect in segregation levels. In Figure 8.12 the newly created CEL aggregations of Western and Eastern Europe show a distinct segregation pattern at OA level, with Eastern

European CELs slightly more segregated (ID = 0.40) than Western European ones (ID= 0.30). This presents a distinct pattern that might be explicable by the differential history of these groups in terms of settlement and socioeconomic profile. In another example, while the Census-based 'Black African' in Figure 8.11 presents an ID index of 0.43 at OA level, the CEL-based Somali group in Figure 8.12 shows a higher ID index of 0.66, denoting an increase in segregation that arises from use of a more detailed ontology of ethnicity.

However, when the effects of the two last aspects of changes of scale are compared; aggregations of geographical units (MAUP) and aggregation of ethnic groups (MEUP), it seems that having more information about CEL group is as much or even more important than having greater spatial detail. This is illustrated with an example in Table 8.7, that shows the effect on the Index of Dissimilarity (ID) of changing between ontologies of ethnicity (MEUP); Census based 'Black African' and CEL based 'Somali', vs. changing the areal aggregation of the calculation (MAUP), at Borough (district), Ward, Lower Super Output Area (LSOA), and Output Area (OA). The MAUP index compares within each ontology of ethnicity the ID value at each geographical scale with the one at OA level (=100). The MEUP index compares the ID of the Somali group with the ID of the Black African (=100) at each of the geographical scales. The conclusion is that while the MAUP effect introduces loss of information (MAUP index) at each scale of aggregation, the relative difference between the two ontologies of ethnicity remains practically constant (MEUP index), therefore corroborating the existence of the MEUP effect.

		Borough	Ward	LSOA	OA
Black African (Census)	ID	0.26	0.35	0.39	0.43
	MAUP index OA=100	60	80	90	100
Somali (CEL)	ID	0.37	0.48	0.53	0.66
	MAUP index OA=100	56	73	80	100
MEUP index (ID Somali / ID Black African * 100)		141	139	136	153

Table 8.7: Effect of MAUP and MEUP on Black African and Somali Index of Dissimilarity in London

ID; Index of Dissimilarity, MAUP; Modifiable Areal Unit Problem, MEUP; Modifiable Ethnic Unit Problem.

The table shows the effect on the Index of Dissimilarity (ID) of changing between ontologies of ethnicity (MEUP); Census based 'Black African' and CEL based 'Somali', vs. changing the areal aggregation of the calculation (MAUP), at Borough (district), Ward, Lower Super Output Area (LSOA), and Output Area (OA). The MAUP index compares within each ontology of ethnicity the ID value at each geographical scale with the one at OA level (=100). The MEUP index compares the ID of the Somali group with the ID of the Black African (=100) at each of the geographical scales. The conclusion is that while the MAUP effect introduces loss of information (MAUP index) at each scale of aggregation, the relative difference between the two ontologies of ethnicity remains practically constant (MEUP index), therefore corroborating the existence of the MEUP effect.

Taken together, there are three inter-related aspects to these observations: the size and number of areal units, the fineness of the ethnic group units, and the ontology of ethnicity (self-reported vs. name-based). All have an impact in the level of segregation that is reported for a particular group. In other words, the granularity and ontology of the units upon which segregation indices are calculated have an important effect on the results, as it has been demonstrated through the comparison presented in Figure 8.11, Figure 8.12 and Table 8.7. It is envisaged that the name-based methodology developed in this thesis will allow future analysts to re-aggregate ethnic groups and geographical units in various flexible ways in order to perform scale-sensitivity analysis of MAUP and MEUP of these indices.

8.4.2. Summary and discussion of overall residential segregation results

The analysis of residential segregation in London presented in the previous three sections has produced a series of interesting results that will be summarised here. The results of the indices calculated here for each CEL Subgroup and the four dimensions of evenness, exposure, concentration, and clustering, are summarised in Table 8.8. In order to rank all of the 46 Subgroups evaluated here from high to low overall segregation an average composite index has been created as follows:

$$Average\ Composite\ Index = (ID + P^* + ACO + ACL)/4$$

where ID= Index of Dissimilarity, P*= Index of Isolation, ACO= Absolute Concentration Index and ACL= Absolute Clustering Index. Standardisation of the four indices was not performed since they are bounded by a 0 to 1 scale, for ease of overall interpretation. However, as it will be seen the 'English' groups is an outlier in most indices and this could have an impact in the final result.

Table 8.8 summarises the value and rank of each index of segregation for each of the four dimensions, alongside the composite index summarising them. The table is ordered by this composite index from high to low overall segregation. It is interesting to note at first sight that the final rank of this composite index is not solely determined by population size. It could be argued that the averaging of the indices is smoothing the population size effect in some of the individual indices discussed in the previous sections.

According to this composite index, the ten most segregated groups are: Sikh, Sierra Leonean, Japanese, Afrikaans, Vietnamese, Iranian, Bangladeshi, African, Danish, and Swedish. Amongst them, only the Sikh and Bangladeshi have previously been

identified as being amongst the most segregated groups in the Capital (Brimicombe, 2007; Peach, 2006), in practice because they are easily identifiable ethno-religious groups in the Census. Amongst the others, two types of segregation might be taking place at the Output Area level: more affluent or highly educated groups seeking exclusive areas of residence (Japanese, Danish, Swedish, Afrikaans and Iranian); and more socio-economically constrained groups (Vietnamese, Sierra Leonean, and African) being constrained to a restricted range of neighbourhoods.

At the opposite end of the segregation scale the following groups present lower overall segregation at Output Area level; Muslim Middle East, Portuguese, English, Spanish, French, Italian, Void, Welsh, Irish, Scottish. Amongst these groups, and as has been reported throughout the chapter, the British Isles CELs comprise the largest and least segregated groups in London (English, Welsh, Scottish, and Irish). The other major group that could be identified seems to be a set of southwest European CELs whose names are well established in the Capital and are more evenly distributed according to the four dimensions of segregation (Portuguese, Spanish, French and Italian). It is comforting to see the 'Void' category presenting low segregation, indicating that there is no direction or pattern in the errors found in the input data.

			Evenness		Isolation		Concentrtn		Clustering		Avg.
Rank	CEL Subgroup	Total Pop.	ID	Rnk	P*	Rnk	ACO	Rnk	ACL	Rnk	Composite Index
1	SIKH	83,968	0.670	20	0.168	2	0.941	41	0.14	2	0.482
2	SIERRA LEONEAN	3,854	0.908	2	0.012	34	0.994	1	0.00	32	0.479
3	JAPANESE	3,469	0.905	3	0.013	30	0.990	8	0.00	29	0.478
4	AFRIKAANS	3,036	0.909	1	0.008	45	0.992	2	0.00	43	0.478
5	VIETNAMESE	8,415	0.862	7	0.026	17	0.990	5	0.01	16	0.472
6	IRANIAN	4,761	0.875	4	0.012	36	0.990	7	0.00	36	0.470
7	BANGLADESHI	72,829	0.644	22	0.150	3	0.973	27	0.10	4	0.469
8	AFRICAN	4,879	0.868	5	0.010	40	0.991	3	0.00	39	0.468
9	DANISH	4,592	0.864	6	0.009	43	0.989	10	0.00	42	0.466
10	SWEDISH	5,155	0.854	8	0.010	41	0.990	6	0.00	38	0.464
11	RUSSIAN	5,539	0.843	9	0.010	39	0.989	11	0.00	40	0.461
12	DUTCH	5,477	0.840	10	0.009	44	0.988	12	0.00	45	0.460
13	SOUTH ASIAN OTHER	8,484	0.788	12	0.012	35	0.987	15	0.00	30	0.448
14	CHINESE	8,874	0.787	13	0.013	32	0.987	16	0.00	35	0.447
15	INTERNATIONAL	6,214	0.789	11	0.007	46	0.991	4	0.00	46	0.447
16	BALKAN	9,035	0.774	14	0.012	37	0.988	13	0.00	34	0.444
17	EUROPEAN OTHER	9,091	0.761	15	0.010	38	0.985	17	0.00	41	0.440
18	HINDI NOT INDIAN	12,643	0.754	16	0.016	27	0.982	20	0.00	27	0.439
19	BLACK CARIBBEAN	11,554	0.739	17	0.013	33	0.989	9	0.00	31	0.436
20	HINDI INDIAN	156,269	0.573	29	0.137	4	0.926	42	0.10	3	0.436
21	MUSLIM SOUTH	11,380	0.734	18	0.013	31	0.985	18	0.00	33	0.434
22	JEWISH	35,984	0.620	24	0.074	7	0.967	36	0.05	6	0.428
23	SRI LANKAN	39,269	0.665	21	0.046	12	0.973	30	0.02	9	0.428
24	UNKNOWN NAME	10,546	0.695	19	0.009	42	0.987	14	0.00	44	0.423
25	TURKISH	34,359	0.620	23	0.033	13	0.975	23	0.01	12	0.411
26	NIGERIAN	68,596	0.580	27	0.058	10	0.969	35	0.02	8	0.409
27	GHANAIAN	35,255	0.611	25	0.029	15	0.980	21	0.01	14	0.408
28	GREEK	61,296	0.557	30	0.062	9	0.954	39	0.04	7	0.404
29	SOMALIAN	20,376	0.585	26	0.015	28	0.982	19	0.00	28	0.397
30	PAKISTANI	140,548	0.495	35	0.085	6	0.947	40	0.05	5	0.395
31	INDIA NORTH	31,888	0.574	28	0.024	20	0.970	32	0.01	13	0.395
32	HONG KONGESE	35,609	0.549	31	0.025	19	0.972	31	0.00	21	0.388
33	PAKISTANI KASHMIR	32,061	0.537	32	0.019	24	0.975	22	0.00	25	0.384
34	NORWEGIAN	24,927	0.516	33	0.014	29	0.974	26	0.00	37	0.376
35	GERMAN	33,264	0.499	34	0.019	23	0.969	34	0.00	22	0.373
36	POLISH	33,270	0.478	36	0.018	26	0.973	29	0.00	24	0.369
37	MUSLIM MIDDLE E.	48,114	0.469	37	0.025	18	0.970	33	0.00	19	0.368
38	PORTUGUESE	44,780	0.464	38	0.023	21	0.974	24	0.00	20	0.367
39	ENGLISH	2,876,980	0.249	43	0.587	1	0.396	46	0.23	1	0.367
40	SPANISH	44,679	0.459	39	0.022	22	0.974	25	0.00	23	0.366
41	FRENCH	40,264	0.445	40	0.019	25	0.973	28	0.00	26	0.361
42	ITALIAN	71,967	0.386	41	0.029	16	0.960	37	0.00	18	0.346
43	VOID	90,715	0.341	42	0.030	14	0.956	38	0.01	17	0.334
44	WELSH	222,429	0.206	44	0.052	11	0.900	43	0.01	15	0.292
45	IRISH	414,038	0.180	46	0.093	5	0.867	45	0.02	10	0.290
46	SCOTTISH	323,847	0.188	45	0.073	8	0.882	44	0.01	11	0.290

Table 8.8: Summary of the four dimensions of segregation and composite index
ID= Index of Dissimilarity, P*= Index of Isolation, ACO= Absolute Concentration Index, ACL= Absolute Clustering Index. Average Composite Index = (ID + P*+ ACO + ACL) /4
This table summarises the value and rank of each index of segregation for each of the four dimensions, alongside a composite index that summarises them all.

The scatterplot in Figure 8.13 presents, in a similar fashion as the ones shown before, a comparison of the average composite index described here with the total population. It demonstrates the negative correlation of the composite index with the group's size, whose linear regression has an R^2 of 0.541. However, as can be seen, there are very stark outliers in this relationship, with several high leverage points. The Sikh, Bangladeshi, Hindi-Indian, and English present very high segregation relative to their population sizes, while the Welsh, Scottish and Irish have lower than expected levels of segregation, followed by Italian, Portuguese, French and Spanish.

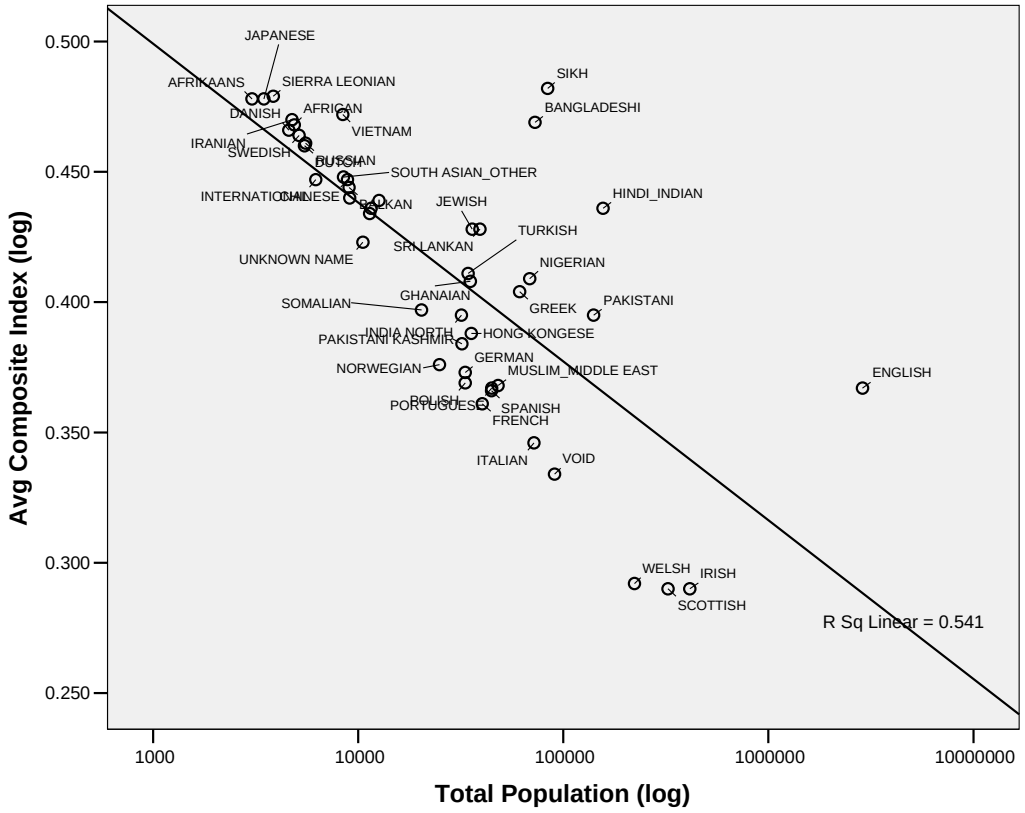

Figure 8.13: Scatterplot of average composite index vs. total population size

Underlying the relationship exposed by Figure 8.13 is the problem repeatedly mentioned in this chapter, namely of the dependency of the segregation indices on the size and number of ethnic groups. This problem is of course linked to the scale

dependency analysed in the previous section, and the three aspects (scale, size and number of ethnic groups) are closely intertwined. However, most of these issues are usually ignored by much of the segregation debate outside the specialised literature. One reason for this is that these issues are difficult to unveil, and it is only when data are available to sufficient level of geographical and nominal disaggregation, as in the examples presented here, that the issues of scale, size and number of ethnic groups become so apparent.

Furthermore, the analysis presented here has made evident that segregation indices were designed with a preconceived idea of residential segregation as being formed solely by a white / Non-White dichotomy. For example, the English CEL Subgroup ranks first in the isolation index, with a P* index of 0.587, only followed in the distance by the Sikh group with P* of 0.168. Is the English group the most isolated of all ethnic groups? The reason behind this bizarre finding is because this index is not designed to be used on the 'majority' ethnic group, but only with one or a few minorities. A similar situation applies to the concentration indices, since the formula is just designed to have a majority group and one or just very few ethnic groups all with substantial population size.

8.5. Other Applications of the CEL Methodology

The bulk of this chapter on applications of the CEL methodology has been dedicated to elaborate a single example in depth; the analysis of residential segregation at very fine geographical and nominal scales. However, many other possible applications of the CEL methodology developed in this PhD research have been envisaged and some

already explored. Because of space limitations in this thesis and the number of external collaborators that have been involved in some of them, they will only be briefly mentioned in this last section.

8.5.1. Ethnic inequalities in health

One of the dimensions of ethnic inequalities that was exposed by the literature review in Section 2.1.2. is ethnic inequalities in health. The organisations charged with tackling and reducing these inequalities at local level frequently lack the appropriate information sources to understand the differential distribution of health outcomes by ethnic groups, and so are limited in their abilities to identify the wider determinants of health that underpin poor outcomes (London Health Observatory, 2005). One rare opportunity to rectify this situation lies in the classification of people's names that are part of the vast datasets about the health of the population. Names analysis has indeed proved very useful in segmenting populations by their most probable ethnic group of origin ever since the 1950s, as per the literature review carried out in Chapter 3.

The CEL methodology presented here has already been applied within the context of a Knowledge Transfer Partnership between University College London (UCL) and Camden Primary Care Trust (PCT) in London, in which the author has worked during his PhD research. Moreover, through links developed during this KTP project with various NHS organisations, the CEL methodology has been tested at various steps of its developmental stages using different health datasets and applications. A few of them will be mentioned here, in order to illustrate other potential areas of application of this methodology.

One of the early examples of applications of the CEL name methodology has been the profiling of the population of General Practices in Camden by ethnic group, based on the names on patients as well as on the countries of birth when they have been reported in the patient register. An example of this type of population profiling for one General Practice in Camden PCT is given in Figure 8.14, showing the rich variety of ethnic groups present, that goes well beyond the coarse categories available in the Census ethnicity classification.

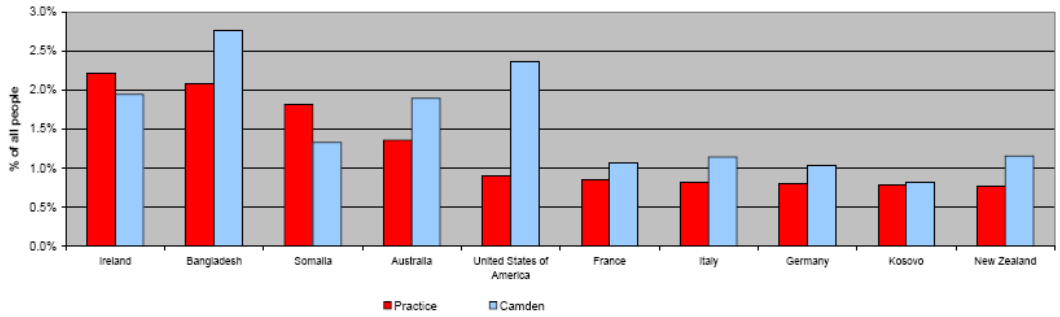

Figure 8.14: Main ethnic groups in the population registered in a general practice in Camden PCT
The chart shows the main ethnic groups that comprise the population of a general practice in Camden PCT's area of responsibility. Ethnicity was estimated using the name-based CEL methodology in addition to country of birth information. Population sizes are relative, given as a percentage of the registered population in the practice (red) and compared to the overall Camden PCT average (blue)

In another application of the CEL methodology, a report was produced for Camden PCT analysing the most likely ethnicity of women who did not respond to breast screening calls, when no ethnicity information had been recorded. Amongst the major findings of the report were that 26% of the non-respondents were women drawn from ethnic minorities, of which 38% had Bangladeshi names, 29% Jewish, 18% Irish and 15% Greek names. This allowed Camden PCT to re-design their breast screening strategy, setting up a survey of ethnic minorities and communication campaign in community centres of the Borough where these ethnic groups are most

heavily represented. Similar types of analysis have been performed at Camden PCT on the birth register, the death register, the hospital episodes statistics dataset, and the general practice patient register, in order to analyse differential patterns of health outcome or to identify population groups at greatest risk of certain public health conditions. Without the name-based methodology, this segmentation of populations would have not been possible.

The CEL methodology has also been used by other PCTs in London to perform similar analysis. Islington PCT used the CEL methodology to identify the ethnicity of patients diagnosed with diabetes. This was done by using the information in the patient register, where the names and the diabetes condition of each individual patient are known but not their individual ethnicity. Ethnicity was then assigned using the CEL methodology developed in this PhD. For the purpose of conducting health equity audits, the CEL-based diabetes prevalence information was then compared with the estimated percentages, aggregated by ethnic group, reported by each general practice (GP). Table 8.9 shows a summary of the percentages of patients diagnosed with diabetes by ethnic group, comparing the figures reported by GPs with those from the CEL-based analysis. This exercise allowed Islington PCT to go back to the general practices where major differences were found and work with them towards improved reporting mechanisms, as well as improved overall access to diabetes diagnosis services.

The words of Islington PCT Public Health Information Officer, are an illustrative testimony in describing the usefulness of this type of analysis:

'The CEL classification of names has helped me a lot. As a public health information officer, when I am asked to find the ethnicity of a certain group of patients and produce reports, I always come across the difficulty of finding out the ethnicity information, because of missing data, or because it is very difficult to formulate the right queries through NHS systems'

(Marina Kukanova, Public Health Information Officer, Islington Primary Care Trust, in an e-mail sent to the author in April 2007)

Ethnic Group	GP assigned	CEL method
White British	57%	41%
Irish	4%	8%
White Other	10%	17%
Indian	2%	2%
Pakistani	0%	1%
Bangladeshi	4%	7%
Other Asian	2%	2%
Black Caribbean	6%	6%
Black African	9%	12%
Chinese	1%	1%
Other	5%	2%
Total	100%	100%

Table 8.9: Patients diagnosed with diabetes in Islington PCT. GP-assigned ethnicity vs. CEL name-based ethnicity
GP= General Practice (third-party reported ethnicity), CEL method= Cultural, Ethnic and Linguistic method (name based)

This is just a very small sample of the number of applications of this type in the health sector. In the last two years of this PhD research the author has received numerous requests of health analysts struggling to segment their population by ethnicity, as equal opportunity practices have been rolled out and interest beyond the typical segmentation by gender, age and geography have been exhausting the basic factors explaining some of the inequalities in health. Another two examples of this

type of demand is depicted by the following e-mail messages received from different health organisations:

> *'(…)I work for a PCT in Birmingham and we are trying to find out which ethnic groups our breastfeeding peer workers deliver a service to. We have the names of all the mums but ethnicity has never been recorded. Can you help me in anyway? Many thanks'*
>
> (Ian Mather, Specialist Trainee in Public Health, Heart of Birmingham PCT, e-mail sent to the author in October 2005)

> *' (…) I saw your webpage on CASA UCL and wonder if you know of software that analyses data by name to pick out different ethnic groups e.g. Chinese? I am secretariat to the Advisory Group on Hepatitis who currently has a working group looking at case-finding options in minority ethnic groups in the UK who originate from countries of intermediate or high hepatitis prevalence. (…) However data on hepatitis infection by ethnicity in the UK is lacking and prevalence here could be different to prevalence in country of origin.'*
>
> (Claire Swales, Advisory Group on Hepatitis secretariat, Expert Advice Support Officer, Health Protection Agency. E-mail sent to the author in June 2007)

It is envisaged that the CEL name classification system developed in this PhD should be made available to these and other types of users through an easy-to-access software platform. Currently the CEL classification runs as a series of queries in an Oracle database and the names have to be sent to the author for coding. As part of the

future plans for developments beyond this PhD, such software should be developed as a stand-alone tool with which users can directly interact. This is also the form of access to the existing South Asian name-to-ethnicity algorithms, *Nam Pechan* and *SANGRA*.

8.5.2. Population studies

Another area of potential applications of the CEL methodology is in population studies, beyond the case of measuring residential segregation which has already been discussed. The CEL classification presents a high potential to be applied in broader population studies about ethnicity, such as: in ethnic group population forecasting by small area (Large and Ghosh, 2006); monitoring migration (Stillwell and Duke-Williams, 2005); diagnosing possible Census undercount (Graham and Waterman, 2005); analysing the geography of ethnic inequalities (Dorling and Rees, 2003) or of mortality and morbidity (Boyle, 2004); evaluating equal opportunity policies (Johnston et al, 2004) and political empowerment processes (Clark and Morrison, 1995); and improving public and private services to ethnic minorities (Van Ryn and Fu, 2003). A central contention of this thesis has been that each of these research and public policy areas presents a lack of appropriate, timely and detailed data on ethnicity. Moreover, this problem is increasing as the last round of Census data age and new migration flows are changing the composition and demands for public services. Improved methods in these areas are thus of key policy importance in today's multi-cultural society.

One example of a real application of the CEL classification in this area has been the attempt to improve population projections by ethnic groups and to better understand the geographical distribution of ethnic groups in the London Borough of

Hammersmith and Fulham (hereinafter H&F). Using the full version of the Electoral Register (as opposed to just the publicly available edited version), H&F compared the nationality of the 119,551 electors in December 2006 with the CEL-derived ethnicity of their names at the individual level, establishing a lookup table between the most likely CEL for each nationality. The conceptions of identity behind these two concepts are obviously quite different: nationality is a legal aspect associated with one's place of birth, or of residence, or of those of one's parents, while the CEL class reflects the culture or language of origin of one's forename and surname. However, the results of this comparison are very encouraging.

Nationality was matched to CEL ethnicity using the lookup table between the CEL Types and the major languages, geographies and Census ethnicity information that is reported in Appendix 3. A total of 101,387 people of a number of 'Anglo-Saxon nationalities' were excluded from this analysis, since British is the default nationality of many second generation ethnic minorities, and other nationalities are primarily comprised of people with British names (English, Welsh and Scottish names). The following nationalities were excluded; British, British Commonwealth, Canadian, Australian, New Zealander, Malta, and other small territories of the Commonwealth.

Table 8.10 shows a summary of this validation exercise in Hammersmith and Fulham. It includes 'CEL Subgroup' and the 'total number of people', which are derived from applying the CEL methodology to the names in the full version of the H&F Electoral Register, and 'Nationality match' and 'Perc. match' which show the counts and percentage of those whose reported nationality in the Electoral Roll matched the name-CEL Subgroup, as per the lookup table described above

(columns). The over-all match rate is 73%, which bearing in mind the different ontological nature of nationality and name-based cultural, ethnic and linguistic origin, is indeed a promising result.

CEL Subgroup (name)	Total people	Nationality match	Perc. match
UGANDAN	3	3	100%
AFRIKAANS	70	63	90%
IRISH	1,967	1,705	87%
POLISH	1,365	1,166	85%
FRENCH	1,490	1,272	85%
NIGERIAN	141	115	82%
SWEDISH	195	159	82%
PORTUGUESE	760	612	81%
GREEK	436	328	75%
ITALIAN	1,998	1,452	73%
FINNISH	64	46	72%
BALTIC	43	29	67%
CZECH & SLOVAKIAN	65	42	65%
HINDI INDIAN	146	94	64%
GERMAN	527	334	63%
GHANAIAN	97	61	63%
HUNGARIAN	93	58	62%
SIERRA LEONEAN	28	17	61%
DANISH	129	78	60%
BLACK CARIBBEAN	15	9	60%
SPANISH	1,245	712	57%
DUTCH	429	224	52%
PAKISTANI	201	80	40%
BANGLADESHI	98	38	39%
SRI LANKAN	57	19	33%
AFRICAN	25	7	28%
MALAYSIA	6	1	17%
HONG KONGESE	157	14	9%
BALKAN	55	4	7%
TOTAL	**11,905**	**8,742**	**73%**

Table 8.10: Validation of the CEL methodology against nationality in Hammersmith and Fulham (London)
The CEL Subgroup is derived from name analysis of the full version of the Electoral Register in the London Borough of Hammersmith and Fulham, showing the total number of people per CEL Subgroup. The 'Nationality match' column is the number of people whose nationality matched the one assigned for that CEL Subgroup. A total of 101,387 people with the following 'Anglo-Saxon nationalities' were excluded from this analysis; British, British Commonwealth, Canadian, Australian, New Zealander, Malta, and other small territories of the Commonwealth. (Source: Data courtesy of Martin Robson, Research Officer, Planning Division, London Borough of Hammersmith & Fulham)

Figure 8.15 presents a comparison between two maps with the distribution of Polish people by output area in Hammersmith and Fulham. Both maps are based on the full

version of the Electoral Register as of December 2006. The map on left represents Polish ethnicity based on the CEL methodology and the one on the right on the self-reported nationality. Both maps show a very similar pattern, facilitating a visual corroboration of the agreement between these two datasets and methodologies to assign population ancestry/national identity.

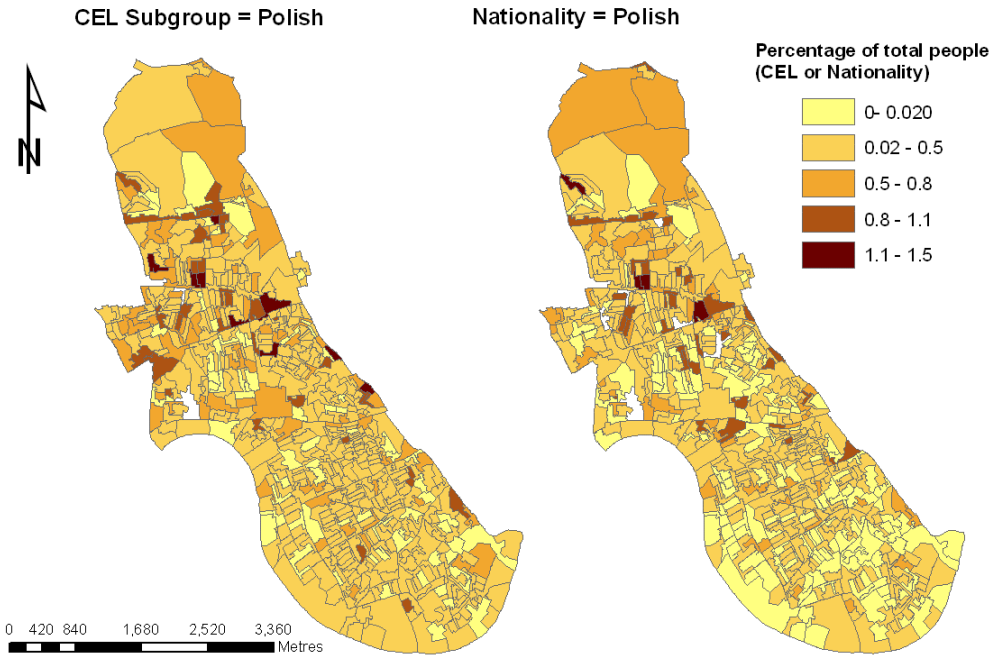

Figure 8.15: Maps of Polish CEL Subgroup vs. Polish nationality in Hammersmith and Fulham (London)
These maps present a comparison of the distribution of Polish people by output area in the London Borough of Hammersmith and Fulham. Both maps are based on the full version of the Electoral Register as of December 2006. The map on left represents Polish ethnicity based on the CEL methodology and the one on the right on the self-reported nationality. Both show a very similar pattern. (Source: Map by the author, using data courtesy of Martin Robson, Research Officer, Planning Division, London Borough of Hammersmith & Fulham)

H&F have subsequently used the name-coded Electoral Register to enrich the knowledge about the 96,306 electors with British or British Commonwealth nationalities inferring their most likely ethnicity based on their names. This information was fed into several planning areas of local government in order to supplement the population projections by ethnic group and small area.

8.6. Conclusion

While the rest of this thesis has focused on different developmental aspects of the CEL methodology, this chapter has presented potential applications of the methodology to ascribe ethnicity using people's names. Most of the chapter has been devoted to one of the many potential applications, given its high relevance to current debates in contemporary society: namely, the study of ethnic residential segregation, and in particular in London.

The application of the CEL methodology to this purpose has opened up new opportunities for much finer analysis of several dimensions of residential segregation in terms of the size of the geographical units and ethnic group boundaries, and the frequency of update. This example has also raised several key questions about the relevance of widely adopted segregation indices which were developed with a very simplistic conception of society based on a 'racial duality' of neighbourhoods, which does not resemble the complexity of contemporary cities, especially outside the US. The large number of ethnic groups and quantity of small neighbourhoods, accommodated by the analysis introduced in this chapter, has brought new challenges to traditional segregation indices that were designed to deal with two or a very few ethnic groups, and zoning schemes that comprise only tens of coarse geographical units.

Despite these challenges, the analysis presented in this chapter has confirmed the conclusions reached by previous studies of segregation in London: namely, the higher degree of residential segregation of some of the South Asian ethno-religious groups, especially Sikh, Indian and Bangladeshi, as well as the Jewish religious

minority. Moreover, the use of name-based ethnicity classifications has suggested a much more complex reality of highly segregated small groups across the socio-economic spectrum: Japanese, Iranian, Danish, Swedish, Sierra Leonean, Afrikaans, Other African, and Vietnamese. In some dimensions, such as evenness and clustering, other groups such us Greek and Turkish CELs show a higher level of segregation than expected by their total population sizes. On the other hand, the three 'Celtic CELs', Welsh, Scottish and Irish show a very low level of segregation across all dimensions, even less so than the English majority.

The number of geographical units considered, the group's population size and the average length of residence of each CEL Subgroup seems to be the three key factors in explaining the major variations observed in the segregation indices in London. In the scatterplots that relate each of these indices and the CEL Subgroups population sizes, there are some CELs that fall outside the main regression trend lines. These should be the ones that receive future attention to investigate the other factors that might explain their atypical behaviour. Most of these groups have been highlighted under each of the dimensions of segregation analysed here.

A commonly used tool in geographical analysis, local spatial autocorrelation, has been also applied here to the study of segregation through the computation of local indicators of spatial association. This tool has proven its ability to delineate local clusters of concentration of the main ethnic groups in London neighbourhoods. Moreover, the use of a diversity index has also allowed the classification of the Capital's output areas according to the number and size of ethnic groups present in each of them, pinpointing the areas that are more diverse. Most of these are found

north of the River Thames and within Inner London. The development of more examples that use different innovative tools from different disciplines, such as the two mentioned here, will make more significant contributions through cross-fertilization between disciplines concerned with residential segregation and socio-spatial differentiation processes.

Finally, the last section in this chapter has briefly mentioned a few examples of other potential applications of the CEL methodology in tackling ethnic inequalities in health and in population forecasting and planning at local level. These constitute a small gallery of applications, in order to illustrate the very wide potential applicability of the CEL classification. From a methodological standpoint, it will only be through the persistent application of the CEL classification to different settings and contexts that its wider validation can take place.

Chapter 9. Conclusions – The Cultural, Ethnic and Linguistic Classification of Names

Identity, though complex, can be encoded in a name

(Seeman, 1980: 129)

9.1. Reflections on Names, Identity, Populations and Neighbourhoods

This closing chapter marks a point of arrival in a journey that started justifying the need for new and innovative methods of analysing ethnicity in the study of inequalities and segregation, and ends with a developed and tested new ontology and classification of ethnicity based on name origins. During this journey, the PhD has gathered enough evidence to substantiate the association between collective identities, languages, genes, places and personal names. It has concluded that if current multicultural cities are composed of a rich variety of culturally diverse neighbourhoods and populations, their inhabitant's names leave in them an 'identity trail' through processes of migration, settlement, mobility and intergenerational transmission of culture. Drawing a physical geography parallel, such processes of 'transportation, sedimentation and erosion' of collective identities may be revealed by studying the 'names geomorphology' left by that identity trail in contemporary populations and neighbourhoods. Therefore, names can serve as very valuable markers of collective identity. This PhD has then set the objective to develop a

methodology to decipher such markers in population registers, based on an ontology of ethnicity designed for this purpose.

As a result, the cultural, ethnic and linguistic classification of names has been developed through this PhD, expanding previous and partial efforts in the literature, and taking a multidisciplinary approach that has borrowed ideas, methods and data sources from the fields of public health, population genetics, historical demography, linguistics, computer science and marketing. The core chapters of this thesis have been dedicated to carefully describing the techniques to develop such taxonomy and classification of names, initially through an exploratory phase and later in an automated and reproducible approach. Outstanding amongst all the techniques used, because of its extraordinary classificatory power, is Forename-Surname Clustering (FSC), which allows the effective clustering of surname types through the cross-occurrences of common forename bearers, and vice-versa for clustering forename types. As a result, two tables have been produced, a surname-to-CEL table including 225,576 surnames, and a forename-to-CEL table including 98,624 forenames, each alongside their most likely CEL and a measure of the degree of association between the name and the CEL. A personal allocation algorithm has been developed to finally assign the most probable CEL to each individual person, using both forename and surname. The intention is to make these tables available on request to bona-fide academic researchers for further evaluation and enhancement.

Moreover, the thesis has also reported the validation of the CEL classification, in multiple applications; one following a public health literature tradition, using patient registers, and a separate one following a ‘geography tradition’ comparing Census

small area statistics. The results of both of the validation exercises have been very consistent, even though the nature of the two datasets was very different, and showed high accuracy level of the CEL methodology in identifying the main ethnic groups in the independent datasets. However, these results should be interpreted in the light of the caveats mentioned relating to the problems inherent in comparing constructs of different ontological nature. Hence, onomastic classifications based on cultural, ethnic and linguistic origin of names cannot be easily compared with self-reported ethnicity.

Finally, the last part of the journey was spent developing an example of the application of the CEL classification to the study of residential segregation in London. This application was chosen given its high relevance to current debates of ethnic residential segregation, in particular in highly diverse cities such as London, and because it constitutes a research area in which geography is the key to understanding social processes. This application has opened up new opportunities for much more sophisticated and detailed analysis of several dimensions of residential segregation in terms of both the size of the geographical units and the definition of ethnic group boundaries. Based on the latter, a new concept of the Modifiable Ethnic Unit Problem (MEUP) has been proposed, differentiating it from the effect of the former, better known as the Modifiable Areal Unit Problem (MAUP) (Openshaw, 1984). Furthermore, the frequency of update of name-based datasets have a great potential compared to traditional sources in the study of residential segregation, since resident characteristics may change rapidly during decennial inter-censal periods. This type of detailed analysis would not be possible without the contribution of the ontology of ethnicity based of names proposed by this thesis.

Many other areas of applications can be envisaged at this stage, some of which have been explicitly mentioned in Chapter 8 in the fields of ethnic inequalities in health and population studies. The number of other possible applications are too numerous to mention, but future work by the author will focus on extending the usage of the CEL classification. These lie beyond the scope of this PhD thesis and form part of the 'next stage in the journey'.

9.2. Advantages and Limitations of the CEL Classification

The name-based ethnicity classification methodology developed in this PhD offers a series of advantages over traditional information sources such as censuses of population. Amongst them, it can be used to develop a more detailed and meaningful classification of people's origins (finer categories based on a very large number of languages, vs. just 10 to 20 ethnic groups in the Census), it offers improved updating (annually through registers with substantial population coverage, such as electoral or patient registers), it better accommodates changing perceptions of identity than ethnicity self-classification (through independent assignment of ethnicity and or cultural origins according to name), and it is made available, subject to confidentiality safeguards, at the individual or household level (rather than an aggregated Census area). Moreover, the research literature suggests that an important advantage of name-based classifications lies in their capability to provide an ethnicity classification when self-reported ethnicity is not available. This is the case in most population registers and data-sets pertaining to individuals, and automated classification provides solutions at a fraction of the cost of alternative methods. However, this advantage may dissipate over time, as the recording of self-reported

ethnicity becomes routine, and data linkage methods make possible wide linkage through population registers (Bhopal et al, 2004; Blakely et al, 2000).

Compared to other name classifications in the literature, the CEL classification offers a number of advantages that overcome some of the issues with previous name analysis identified in Chapter 3. These include that other methods: only classify a few ethnic groups; ignore spatio-temporal differences in the frequency distribution of names and the selective nature of migration; fail to exploit differences in the strength of association between a name and an ethnic group. The advantages of the CEL classification are that: it classifies the entire population of a country (in this case optimised for the UK), rather than just a specific subset; it classifies the names into all of the likely cultural, ethnic and linguistic groups (CELs) found in a society, rather than just one or a very few ethnic groups; it allows the end user to aggregate highly disaggregate CELs into new groups according to different criteria that are appropriate and tailored to each individual application, consistent with the core CEL taxonomy; and finally, it offers a measure of strength through name scores and person-allocation scores, that allows the user to adapt the classification to the sensitivity of the specific research purpose. It can be argued that these advantages constitute important contributions of this PhD to existing knowledge on name-based ethnicity classifications.

In spite of all of these advantages, the CEL classification suffers from some of the same limitations as its predecessors in the literature: autocorrelation of names amongst family members; name spelling errors and name normalisation issues; inconsistent transcriptions or transliterations of name into different alphabets or

errors induced by pronunciation; names usually only reflecting patrilineal heritage; and different histories of name adoption, naming conventions and surname change. Moreover, if exogamy outside CEL groups increases, as is anticipated in the near future, the method's discriminatory ability may decline. These are all problems that the final user of the CEL classification will have to solve as they specifically affect their data configuration and particular research objectives.

Finally, there is a series of ethnical considerations that should be taken into account when handling names data about individuals and inferring the CEL or ethnicity assignments based upon them. These have not been specifically addressed by this thesis, but future researchers should be exercise caution in considering the sensitivity of the context in which their research is conducted, so as to prevent any potential risk of incorrect inference. Nevertheless, it should be restated that the classification of populations into groups of common culture, ethnicity and language based on the origin of names cannot, by definition, replace self-assigned ethnicity, in the sense that the information based upon names is not intended to replace personal perception of identity, but rather provide the most plausible externally assigned one. As has been discussed in this thesis, this is not necessarily a bad thing, but this important distinction should be taken into consideration at every step of the research when using this type of method.

9.3. Future Research

Two avenues for future research are envisaged at this stage. One is inherently methodological, and focused upon development of the core classification proposed in

this thesis; and the other lies in the development of new types of applications using this methodology.

9.3.1. Methodological improvements

Amongst the methodological aspects, there are three major needs for future enhancements that arise out of this research: fully automating the classification of names; expanding the classification to cover several countries; and using contextual information to improve the CEL allocations.

The first type of enhancement, full automation of the classification of names, arises from the need to find an alternative to dependency on an externally and non-automatically produced 'forename seed list', that is currently used to 'ignite' the automated classification presented in this thesis. Some alternative methods that have already been explored in this PhD are the account of the investigations reported in Section 6.5. These entail the use of subspace clustering techniques of very large two-way sparse matrices of names, which although offering some promising aspects, as yet produce results deemed insufficiently successful because of the size and complexity of the names datasets and its relationships. As a consequence, these investigations did not progress to the implementation phase. These attempts are nevertheless very promising and should inform future research in this area.

The second type of enhancement lies in expanding the classification to cover several countries, since the current system is only optimised to classify the population resident in the UK. Each country has a unique set of 'host' and 'foreign' names, that is closely linked to very specific geo-historical contexts, and that is currently

changing very rapidly because of international migration flows. The first stage of this would involve expanding the number of names in the database and the number of CELs in the taxonomy in order to represent the majority of names in a number of other countries where the CEL classification could be used to classify ethnic groups, specifically in Western Europe, North America, and Australia, but in many other regions as well. The second stage would imply the development of country-specific name scores, to avoid cases of name overlap between CELs (for example 'Martin' is an English surname in Britain but a Spanish surname in Spain), and to develop an algorithm to decide between competing CELs depending on the country for which the population is being classified.

The third type of enhancement is to develop a person CEL allocation algorithm that takes into account contextual information in order to improve the allocations. Such contextual assignment is somehow related to the last point made about the international context, and relates to situations where other multiple sources of information could be used in combination with a name's ethnicity, such as for example the address of a person's residence or his or her place of birth. For example, a Jewish component of a name could be given more weight in the allocation if the person lives in Golders Green in London (an area known to have high concentrations of Jewish people). Moreover, most population registers (birth, death and patient registers) consist of a name, address and place of birth, and the last two components have great potential to enhance the accuracy of the final person CEL allocation. This enhancement would entail building lookup tables of contextual information between geographies and CELs, using Census information and other sources.

9.3.2. Future types of applications

The second avenue for future research concerns the type of applications that this methodology permits, that were somehow unthinkable before. There are endless possibilities of exploiting the CEL methodology through the 'traditional applications' that have been mentioned in Chapter 8, contributing to some of the research gaps identified by the literature in respect to facilitating better understanding of ethnicity through advanced analysis of population composition and change in contemporary society and cities.

However, it is envisaged that the methodology presented in this thesis will make it possible to ask totally different types of questions to the traditional research enquiries in social science, heavily determined by the availability of up-to-date and accurate Census data (particularly of recent immigrant or mobile populations). These of course are difficult to predict at this stage, but are likely to emerge from new ways of analysing population collective identity at the individual level, or at very small area level, i.e. the household, building or street block/postcode unit. They will also be related to the developing opportunities to: link datasets about the same individuals longitudinally through space and time; or to analyse relationships within and between different aggregations of CEL groups and geographical units (assessing the MEUP-MAUP effect). Such types of linkages and clustering through CELs will make it possible to address new questions about population change over time and space, sometimes using very fine temporal resolution datasets, such as frequently updated electoral or patient registers. An example of an area where these questions look promising is in the analysis of the *movement* dimension of residential segregation recently proposed by Simpson (2007).

Moreover, the linkages between the different aspects of identity revealed by names can feed enormous amounts of complex information into current debates concerning social networks (Wasserman and Galaskiewicz, 1995). An example of these possibilities in shown in Figure 9.1, which represents the relationships between surnames in a sample of 5,000 people by showing a network of the 'forename distances' between them, that is, whether surnames are related through common forename bearers or not. It can be appreciated how some surnames are only related to some other surname types, constituting neatly defined clusters, which in essence is the characteristic exploited by the FSC technique. However, the possibilities for social network research of these relationships are endless, especially if the CEL classification and geography are added to the network.

These are just some examples of how it is anticipated that researchers will be able to develop interesting new types of applications of the CEL classification using individual person level data and finely spatially disaggregated data.

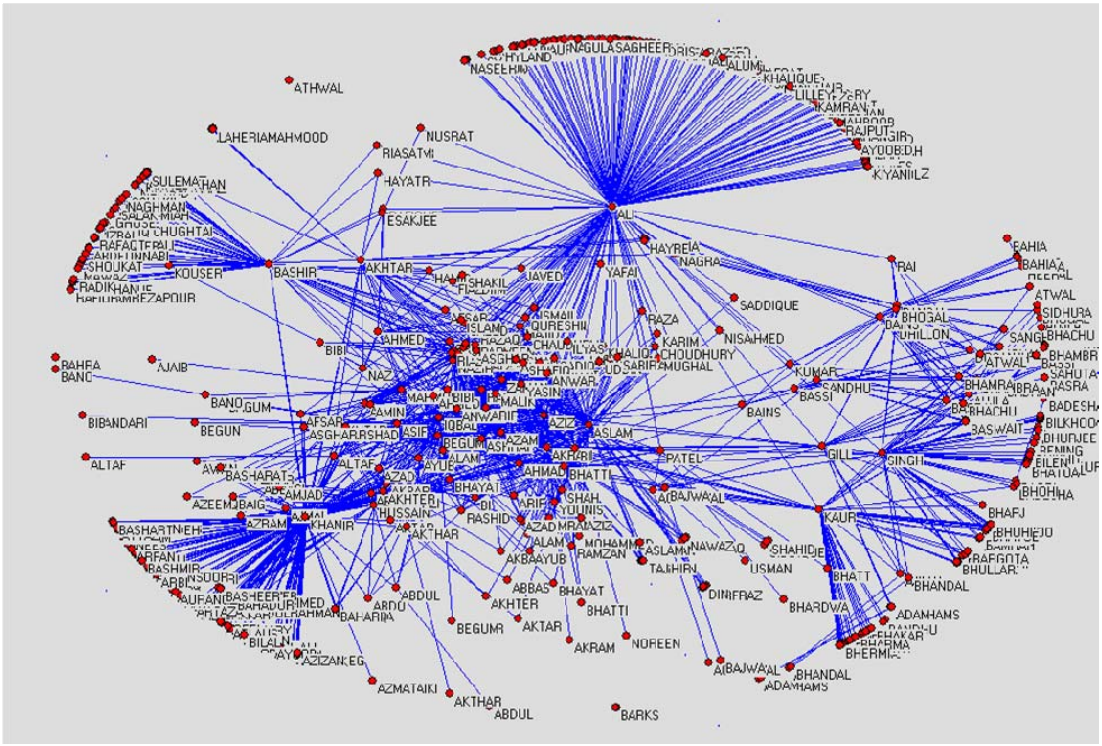

Figure 9.1: Network of 'forename distance' between the surnames of a sample of 5,000 people
The network represents the relationships between surnames in a sample of 5,000 people drawn from non-British surnames in the GB 2004 Electoral Register. It shows a network of those surnames related by the 'forename distances' between them, that is, whether surnames are related through common forename bearers or not. Distances and locations are randomly selected by the software *Pajek*. Some surnames are only related to small cluster of similar ones, while a few act as 'surname hubs' in between different clusters. Source: Prepared by the author using GB 2004 Electoral Register and social network analysis software *Pajek* (*http://vlado.fmf.uni-lj.si/pub/networks/pajek/*).

9.4. Concluding Statement

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 PhD 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 thesis 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 thesis will assist them in this difficult task.

References

Abbotts J, Williams R, Smith GD. 1999. Association of medical, physiological, behavioural and socio-economic factors with elevated mortality in men of Irish heritage in West Scotland. *Journal of Public Health Medicine* **21**(1): 46-54

Abrahamse AF, Morrison PA, Bolton NM. 1994. Surname analysis for estimating local concentration of Hispanics and Asians. *Population Research and Policy Review* **13**: 383-398

Adebayo C and Mitchell P. 2005. Patient Profiling. Presented at *GEONom*, London, 25 May. Available at: www.casa.ucl.ac.uk/geonom/Initial_meeting Accessed: 12/05/2006.

Afshari R and Bhopal R. 2002. Changing pattern of use of 'ethnicity' and 'race' in scientific literature. *Int J Epidemiol* **31**: 1074-1076

Agyemang C, Bhopal R, Bruijnzeels M. 2005. Negro, Black, Black African, African Caribbean, African American or what? Labelling African origin populations in the health arena in the 21st century. *Journal of Epidemiology and Community Health* **59**(12): 1014-1018

Akenson DH. 1984. Why the accepted estimates of ethnicity of the American people, 1790, are unacceptable. *The William and Mary Quarterly* **41**(1): 102-119

Aldenderfer MS and Blashfield RK. 1984. *Cluster Analysis*. London: Sage.

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.

American Sociological Association. 2002. *Statement of the American Sociological Association on the Importance of Collecting Data and Doing Social Scientific Research on Race. Adopted by the elected Council of the American Sociological Association.* Available at: http://www.asanet.org/galleries/default-file/asa_race_statement.pdf. Accessed: 20/09/2006.

Amin A. 2002. Ethnicity and the multicultural city: living with diversity. *Environment and Planning A* **34**(6): 959 - 980

Anselin L. 1995. Local indicators of spatial association - LISA. *Geographical Analysis* **27**: 93-115

Anselin L and Regents of the University of Illinois. 2004. *GeoDa Analysis Software (version 0.9.5-i).* Available at: https://www.geoda.uiuc.edu/. Accessed: 31/03/2007.

Apparicio P, Petkevitch V, Charron M. 2005. *Une application C#.Net pour le calcul des indices de ségrégation résidentielle.* Document de recherche (2005-02). INRS. Urbanisation Culture et Societé. Available at: http://www.inrs-ucs.uquebec.ca/pdf/inedit2005_02.pdf Accessed: 10/12/2005.

Aranda Aznar J. 1998. La mezcla del pueblo Vasco. *Empiria* **1**: 121-177

Arday SL, Arday DR, Monroe S, Zhang J. 2000. HCFA's Racial and Ethnic Data: Current Accuracy and Recent Improvements. *Health Care Financing Review* **21**(4): 107-116

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. 2002. Collective terminology to describe the minority ethnic population: the persistence of confusion and ambiguity in usage. *Sociology* **36**(4): 803-816

Aspinall PJ. 2003. Who is Asian? A category that remains contested in population and health research. *Journal of Public Health Medicine* **25**(2): 91-97

Aspinall PJ. 2005. The operationalization of race and ethnicity concepts in medical classification systems: issues of validity and utility. *Health Informatics Journal* **11**(4): 259-274

Aspinall PJ. 2007. Approaches to Developing an Improved Cross-National Understanding of Concepts and Terms Relating to Ethnicity and Race. *International Sociology* **22**(1): 41-70

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.

Barker HF. 1926. Our Leading Surnames. *American Speech* **1**(9): 470

Barker HF. 1928. How We Got Our Surnames. *American Speech* **4**(1): 48

Barrai I, Rodriguez-Larralde A, Mamolini E, Manni F, Scapoli C. 2000. Elements of the surname structure of Austria. *Annals of Human Biology* **27**(6): 607-622

Barrai I, Rodriguez-Larralde A, Mamolini E, Manni F, Scapoli C. 2001. Isonymy structure of USA population. *American Journal of Physical Anthropology* **114**: 109-128

Barrai I, Rodriguez-Larralde A, Manni F, Ruggiero V, Tartari D, et al. 2003. Isolation by Language and Distance in Belgium. *Annals of Human Genetics* **68**(1): 1-16

Barrai I, Scapoli C, Beretta M, Nesti C, Mamolini E, et al. 1996. Isonymy and the genetic structure of Switzerland I. The distributions of surnames. *Annals of Human Biology* **23**: 431-455

Barrier NG. 1981. *The Census in British India*. New Delhi: Manohav.

Barry H, 3rd and Harper AS. 2000. Three last letters identify most female first names. *Psychological Reports* **87**(1): 48-54

Bateson G. 1979. *Mind and Nature: a Necessary Unity*. Glasgow: Fontana.

Baumgartner C, Plant C, Railing K, Kriegel HP, Kroger P. 2004. Subspace selection for clustering high-dimensional data. Presented at *ICDM '04. Fourth IEEE International Conference on Data Mining*, Brighton, UK, 1-4 Nov. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1410261 Accessed: 23/11/2006.

BBC Radio 4. 2005. *The Colour-coded Prescription.* Transcription of radio program broadcasted on 17 Nov 2005. Available at: http://news.bbc.co.uk/nol/shared/spl/hi/programmes/analysis/transcripts/17_11_05.txt. Accessed: 20/11/2005.

Bell W. 1954. A Probability Model for the Measurement of Ecological Segregation. *Social Forces* **32**(4): 357-364

Bhopal R. 1997. Is research into ethnicity and health racist, unsound, or important science? *British Medical Journal* **314**(7096): 1751-

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. 2007. *Ethnicity, Race, and Health in Multicultural Societies*. Oxford: Oxford University Press.

Bhopal R and Donaldson L. 1998. White, European, Western, Caucasian, or what? Inappropriate labeling in research on race, ethnicity, and health. *American Journal of Public Health* **88**(9): 1303-1307

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

Birkin M and Clarke G. 1998. GIS, Geodemographics, and spatial modeling in the UK financial service industry. *Journal of Housing Research* **9**(1): 87-111

Blakely T, Woodward A, Salmond C. 2000. Anonymous linkage of New Zealand mortality and Census data. *Australian and New Zealand Journal of Public Health* **24**(1): 92-95

Boal FW. 2000. *Ethnicity Housing: Accommodating Differences*. Aldershot, UK: Avebury.

Bonaventura P, Gori M, Maggini M, Scarselli F, Sheng J. 2003. A hybrid model for the prediction of the linguistic origin of surnames. *IEEE Transactions on Knowledge and Data Engineering* **15**(3): 760-763

Border and Immigration Agency, Department for Work and Pensions, HM Revenue & Customs, Department for Communities and Local Government. 2007. *Accession Monitoring Report, A8 Countries. May 2004 - March 2007.* Interdepartamental Online Report. London. Available at: http://www.ind.homeoffice.gov.uk/6353/aboutus/AccessionMonitoringReport11.pdf. Accessed: 22/05/2007.

Bouwhuis CB and Moll HA. 2003. Determination of ethnicity in children in the Netherlands: two methods compared. *European Journal of Epidemiology* **18**(5): 385-388

Boyle P. 2004. Population geography: migration and inequalities in mortality and morbidity. *Progress in Human Geography* **28**(6): 767-776

Boyle PJ, Sarah Curtis, Elspeth Graham, Eric Moore. 2004. *The Geography of Health Inequalities in the Developed World: Views from Britain and North America*. Aldershot, UK: Ashgate.

Brimicombe A. 2007. Ethnicity, religion and residential segregation in London: evidence from a computational typology of minority communities. *Environment and Planning B*: (in press)

Brown LA and Chung S-Y. 2006. Spatial segregation, segregation indices and the geographical perspective. *Population, Space and Place* **12**(2): 125-143

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

Buechley RW. 1961. A Reproducible Method of Counting Persons of Spanish Surname. *Journal of the American Statistical Association* **56**(293): 88-97

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

Buechley RW. 1976. *Generally useful ethnic search system: GUESS* (mimeo) Cancer Research and Treatment Center. University of New Mexico. Albuquerque.

Buechley RW, Dunn J, Linden G, Breslow L. 1957. Excess lung cancer mortality rates among Mexican women in California. *Cancer* **10** 63-43

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

Bunting M. 2006. Living separate lives. *The Guardian* 29/11/2006. Available at: http://commentisfree.guardian.co.uk/madeleine_bunting/2006/11/post_727.html. Accessed: 23/01/2007.

Burgess S, Wilson D, Lupton R. 2005. Parallel lives? Ethnic segregation in schools and neighbourhoods. *Urban Studies* **42**: 1027

Cantle T. 2001. *Community Cohesion. A Report of the Independent Review Team.* Home Office,. London. Available at: http://image.guardian.co.uk/sys-files/Guardian/documents/2001/12/11/communitycohesionreport.pdf. Accessed: 19/02/2007.

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

Carter J, Fenton S, Modood T. 1999. *Ethnicity and employment in higher education*. Policy Studies Institute report, 865. London: Policy Studies Institute.

Castells M. 1997. *The power of identity. Information age: economy, society and culture*. Vol. 2. Oxford: Blackwell.

Castles S and Miller MJ. 2003. *The age of migration*. 3rd ed. Basingstoke, UK: Palgrave Macmillan.

Cavalli-Sforza LL. 1997. Genes, peoples, and languages. *Proceedings of the National Academy of Sciences* **94**(15): 7719-7724

Cavalli-Sforza LL. 2001. *Genes, Peoples, and Languages*. London: Penguin Books.

Cavalli-Sforza LL and Cavalli-Sforza F. 1995. *The Great Human Diasporas*. Reading, Massachusetts: Addison-Wesley.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. *The History and Geography of Human Genes*. New Jersey: Princeton University Press.

Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J. 1988. Reconstruction of Human Evolution: Bringing Together Genetic, Archeological and Linguistic Data. *Proceedings of the National Academy of Sciences* **85**: 6002-6006

Chapman RR and Berggren JR. 2005. Radical contextualization: contributions to an anthropology of racial/ethnic health disparities. *Health: An Interdisciplinary Journal for the Social Study of Health, Illness and Medicine* **9**(2): 145-167

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

Chiapella AP and Feldman HI. 1995. Renal failure among male Hispanics in the United States. *American Journal of Public Health* **85**(7): 1001-1004

Chiarelli B. 1992. The use of family names in the study of human migration during the last two centuries. *Mankind Quarterly* **33**(1): 69-77

Choi BCK, Hanley AJ, Holowaty EJ, Dale D. 1993. Use of surnames to identify individuals of Chinese ancestry. *American Journal of Epidemiology* **138**: 723-734

Choi KH and Sakamoto A. 2005. *Who is Hispanic? Hispanic ethnic identity among African Americans, Asian Americans, and Whites.* PRC Working Paper Series. Rep. No. 04-05-07, Population Research Centre. University of Texas at Austin. Available at: http://www.prc.utexas.edu/working_papers/wp_pdf/04-05-07.pdf Accessed: 22/02/2005.

Christopher AJ. 2002. "To define the indefinable": population classification and the census in South Africa. *Area* **34**(4): 401-408

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

Clarke C, Ley D, Peach C. 1984. *Geography and Ethnic Pluralism*. London: Allen & Unwin.

Colantonio SE, Lasker GW, Kaplan BA, Fuster V. 2003. Use of surname models in human population biology: a review of recent developments. *Human Biology* **75**(6): 785-807

Coldman AJ, Braun T, Gallagher RP. 1988. The classification of ethnic status using name information. *Journal of Epidemiology and Community Health* **42**(4): 390-395

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

Commission for Racial Equality. 2002. *Community Cohesion. Our Responsibility.* Commission for Racial Equality. London. Available at: http://www.cre.gov.uk/downloads/cohesion.pdf. Accessed: 03/10/2006.

Compact Oxford English Dictionary. 2005. Oxford: Oxford University Press.

Comstock RD, Castillo EM, Lindsay SP. 2004. Four-Year Review of the Use of Race and Ethnicity in Epidemiologic and Public Health Research. *Am. J. Epidemiol.* **159**(6): 611-619

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.

Cook D, Hewitt D, Milner J. 1972. Uses of the surname in epidemiologic research. *American Journal of Epidemiology* **95**: 38-45

Cope I. 2005. The 2011 Census. In *Census: Present and Future*. Leicester: ESRC/JISC Census Population Program

Coronado GD, Koepsell TD, Thompson B, Schwartz SM, Wharton RS, et al. 2002. Assessing cervical cancer risk in Hispanics. *Cancer Epidemiology Biomarkers and Prevention* **11**(10 Pt 1): 979-984

Couzin J. 2005. To what extent are genetic variation and personal health linked? . *Science* **309**: 81

Crow JF and Mange A. 1965. Measurements of inbreeding from the frequency of marriages between persons of the same surnames. *Eugenics Q.* **12**: 199-203

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

Curtis S and Jones IR. 1998. Is there a place for geography in the analysis of health inequality? *Sociology of Health & Illness* **20**(5): 645-672

Dance P. 2007. Personal communication.

Danmarks Statistik. 2006. *Mest populære for- og efternavne for alle danskere.* Available at: http://www.dst.dk/Statistik/Navne/pop.aspx. Accessed: 23/07/2006.

Darwin C. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.

Darwin GH. 1875. Marriages between first cousins in England and their effects. *Journal of the Statistical Society of London* **38**: 153-184

Degioanni A and Darlu P. 2001. A Bayesian approach to infer geographical origins of migrants through surnames *Annals of Human Biology* **28**(5): 537-545

Department of Communities and Local Government. 2006. *Commission on Integration and Cohesion.* London. Available at: http://www.communities.gov.uk/index.asp?id=1501520. Accessed: 20/03/2007.

Department of Health. 2003. *Tackling Health Inequalities: A Programme for Action.* Health Inequalities Unit (HID). London. Available at: http://www.dh.gov.uk/assetRoot/04/01/93/62/04019362.pdf. Accessed: 15/08/2005.

Department of Health. 2005a. *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.

Department of Health. 2005b. *Tackling Health Inequalities: Status Report on the Programme for Action.* Health Inequalities Unit. Available at: http://www.dh.gov.uk/assetRoot/04/11/76/98/04117698.pdf Accessed: 15/08/2005.

Dipierri JE, Alfaro EL, Scapoli C, Mamolini E, Rodriguez-Larralde A, et al. 2005. Surnames in Argentina: A population study through isonymy. *American Journal of Physical Anthropology* **128**: 199-209

Dorling D. 2005a. *Human Geography of the UK*. London: Sage.

Dorling D. 2005b. Why Trevor is wrong about race ghettos *The Observer* Sunday September 25. Available at: http://www.guardian.co.uk/race/story/0,11374,1577790,00.html

Dorling D and Rees P. 2003. A nation still dividing: the British census and social polarisation 1971 - 2001. *Environment and Planning A* **35**(7): 1287-1313

Duncan OD and Duncan B. 1955. A Methodological Analysis of Segregation Indexes. *American Sociological Review* **20**(2): 210-217

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

Electoral Commission. 2002. *Changes to the electoral register.* Electoral Commission. London. Available at: http://www.electoralcommission.org.uk/files/dms/optout_6781-6314__E__S__W__.pdf. Accessed: 24/07/2006.

Engels F. 1987. *The Condition of the Working Class in England*. Harmondsworth, England: Penguin, Originally published in 1845.

Equifax. 2007. ER voter opt-out hits record levels. *Data Strategy* (07 February): 6

Eriksen TH. 2002. *Ethnicity and Nationalism*. Second Edition. London: Pluto Press.

Experian Ltd. 2004. *Mosaic United Kingdom. The consumer classification for the UK.* Nottingham. Available at: http://www.business-strategies.co.uk/upload/downloads/mosaic%20uk%20brochure.pdf. Accessed: 02/11/2006.

Fernandez EW. 1975. *Comparison of Persons of Spanish Surname and Persons of Spanish Origin in the United States.* Technical Paper No. 38. U.S. Bureau of the Census,. Washington.

Fiscella K and Fremont AM. 2006. Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity. *Health Services Research* **41**(4p1): 1482-1500

Fortuijn JD, Musterd S, Ostendorf W. 1998. International Migration and Ethnic Segregation: Impacts on Urban Areas. *Urban Studies* **35**(3): 367 - 370

Fotheringham SA, Brunsdon C, Charlton M. 2000. *Quantitative Geography*. London: Sage.

Frazier JW, Margai FM, Tettey-Fio E. 2003. *Race and place: equity issues in urban America*. Boulder, CO: Westview Press.

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

Fucilla JG. 1943. The Anglicization of Italian Surnames in the United States. *American Speech* **18**(1): 26-32

Gatrell A, Thomas C, Bennett S, Bostock L, Popay J, et al. 2000. Understanding health inequalities: locating people in geographical and social spaces. In *Understanding Health Inequalities*, Graham H (eds.), Open University Press: Brighton 156-169

Geary RC. 1954. The Contiguity Ratio and Statistical Mapping. *Incorporated Statistician* **5**: 115-141

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

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

Gill P, Bhopal R, Wild S, Kai J. 2005. Limitations and potential of country of birth as proxy for ethnic group. *British Medical Journal* **330**(7484): 196

Goodchild MF. 1986. *Spatial Autocorrelation*. Catmog 47. Norwich: Geo Books.

Gordon AD. 1999. *Classification*. 2nd Ed. London: Chapman and Hall.

Gordon RG, Jr. (ed.). 2005. *Ethnologue: Languages of the World.* SIL International. Dallas, Tex. Available at: http://www.ethnologue.com/. Accessed: 21/11/2006.

Gould S. 1984. *The mismeasure of man*. London: Pelican.

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

Graves JLJ. 2002. *The Emperor's New Clothes. Biological theories of Race at the Millennium*. New Brunswick, NJ: Rutgers University Press.

Guibernau M, Rex J, (eds.). 1997. *The Ethnicity Reader. Nationalism, Multiculturalism and Migration*. Cambridge: Polity Press.

Gutgesell M, Terrell G, Labarthe D. 1981. Pediatric blood pressure: ethnic comparisons in a primary care center. *Hypertension* **3**(1): 39-47

Hage BH, Oliver RG, Powles JW, Wahlqvist ML. 1990. Telephone directory listings of presumptive Chinese surnames: an appropriate sampling frame for a dispersed population with characteristic surnames. *Epidemiology* **1**(5): 405-408

Haggett P. 2001. *Geography. A Global Synthesis*. Harlow, England: Prentice Hall.

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

Hanks P, Hardcastle K, Hodges F. 2003. *Oxford Dictionary of First Names* Oxford: 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

Harland JO, White M, Bhopal RS. 1997. Identifying Chinese populations in the UK for epidemiological research experience of a name analysis of the FHSA register. Family Health Services Authority. *Public Health* **111**: 331-337

Harris R, Sleight P, Webber R. 2005. *Geodemographics: neighbourhood targeting and GIS*. Chichester, UK: John Wiley and Sons.

Hartshorne C and Weiss P. 1931. *Collected Papers of Charles Sanders Peirce*. Cambridge, MA. cited in Burks AW. 1949. Icon, Index, and Symbol. Philosophy and Phenomenological Research 9(4): 673-689: Harvard University Press.

Harvey D. 2005. *A Brief History of Neoliberalism*. New York: Oxford University Press.

Haskey J. 2002. *Population projections by ethnic group. A feasibility study.* Studies on Medical and Population Subjects No. 67. Statistics OfN. Available at: http://www.statistics.gov.uk/downloads/theme_population/SMPS_67_v2.pdf. Accessed: 23/12/2006.

Helgason A, Yngvadóttir B, Hrafnkelsson B, Gulcher J, Stefánsson K. 2004. An Icelandic example of the impact of population structure on association studies. *Nature Genetics* **37**: 90 - 95

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

Hinton L, Jenkins CN, McPhee S, Wong C, Lai KQ, et al. 1998. A survey of depressive symptoms among Vietnamese-American men in three locales:

prevalence and correlates. *The Journal of Nervous and Mental Disease* **186**(11): 677-683

Hitler A. 1925. *Mein Kampf*. Munich: Secker and Warburg.

HM Government. 1976. Race Relations Act 1976. HMSO

Hofstetter CR, Hovell MF, Lee J, Zakarian J, Park H, et al. 2004. Tobacco use and acculturation among Californians of Korean descent: a behavioral epidemiological analysis. *Nicotine and Tobacco Research* **6**(3): 481-489

Honer D. 2004. *Identifying Ethnicity: A comparison of two computer programmes designed to identify names of South Asian ethnic origin.* UK Centre for Evidence in Ethnicity Health & Diversity. University of Warwick. Available at: http://www2.warwick.ac.uk/fac/med/research/csri/ethnicityhealth/aspects_diversity/identifying_ethnicity/. Accessed: 22/06/2006.

Howard D and Hopkins PE. 2005. Editorial: race, religion and the census. *Population, Space and Place* **11**(2): 69-74

IDESCAT. 2006. *Els noms de la població de Catalunya per nacionalitats.* Institut d' Estadistica de Catalunya. Available at: http://www.idescat.net/orpi/Orpi?TC=X. Accessed: 12/05/06.

Instituto de Estadística de la Comunidad de Madrid. 2006. *Guía de nombres y primer apellido de los residentes en la Comunidad de Madrid 1998-2005* Available at: http://www.madrid.org/iestadis/fijas/otros/anecdotas.htm. Accessed: 12/02/06.

International Organisation for Standardisation. 2007. *Codes for the representation of names of languages* Available at: http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=39534&ICS1=1&ICS2=140&ICS3=20. Accessed: 05/03/2007.

Jackson P and Smith SJ. 1981. *Social interaction and ethnic segregation* London Academic Press.

Jha A. 2006. How DNA may tell police the surname of the criminal. *The Guardian* 22 February. Available at: http://www.guardian.co.uk/science/story/0,,1715023,00.html. Accessed: 24/03/2006.

Jobling MA. 2001. In the name of the father: surnames and genetics. *Trends in Genetics* **17**(6): 353-357

Johnston R, Burgess S, Wilson D, Harris R. 2006. School and Residential Ethnic Segregation: An Analysis of Variations across England's Local Education Authorities. *Regional studies* **40**(9): 973

Johnston R, Forrest J, Poulsen M. 2002a. Are there Ethnic Enclaves/Ghettos in English Cities? *Urban Studies* **39**: 591

Johnston R, Forrest J, Poulsen M. 2002b. The ethnic geography of EthniCities: The 'American model' and residential concentration in London. *Ethnicities* **2**(2): 209-235

Johnston R, Gregory D, Pratt G, Watts M. 2000. *The Dictionary of Human Geography*. Oxford: Blackwell.

Johnston R, Poulsen M, Forrest J. 2003. Ethnic residential concentration and a 'new spatial order'?: exploratory analyses of four United States metropolitan areas, 1980-2000. *International Journal of Population Geography* **9**(1): 39-56

Johnston R, Poulsen M, Forrest J. 2005. On the measurement and meaning of residential segregation: A response to Simpson. *Urban Studies* **42**: 1221

Johnston R, Poulsen M, Forrest J. 2006. Blacks and Hispanics in urban America: similar patterns of residential segregation? *Population, Space and Place* **12**(5): 389-406

Johnston R, Voas D, Poulsen M. 2003. Measuring spatial concentration: the use of threshold profiles. *Environment and Planning B: Planning and Design* **30**(1): 3-14

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

Jorde LB and Morgan K. 1987. Genetic structure of the Utah Mormons: isonymy analysis. *American Journal of Physical Anthropology* **72**(3): 403-412

Kahn J. 2005. Misreading race and genomics after BiDil. *Nature Genetics* **37**(7): 655-656

Karlsen S, Nazroo JY, Stephenson R. 2002. Ethnicity, environment and health: putting ethnic inequalities in health in their place. *Social Science & Medicine* **55**(9): 1647-1661

Karn V, ed. 1997. *Ethnicity in the 1991 Census. Volume 4. Employment, education and housing among the ethnic minority populations of Britain*. Office for National Statistics, HMSO London

Karypis G. 2003. *CLUTO, A Clustering Toolkit. Release 2.1.1. Technical Report: #02-017.* University of Minnesota, Department of Computer Science. Minneapolis, MN

Kaufman L and Rousseeuw PJ. 2005. *Finding Groups in Data*. Hoboken, NJ, US: Wiley.

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

Kimmerle MM. 1942. Norwegian-American Surnames in Transition. *American Speech* **17**(3): 158-165

King T, Ballereau S, Schürer KE, Jobling MA. 2006. Genetic signatures of coancestry within surnames. *Current Biology* **16**(4): 384-388

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

Kleiweg P. 2001. *Extended Kohonen Maps software***.** Rijksuniversiteit Groningen. Groningen. Available at: http://www.let.rug.nl/~kleiweg/kohonen/kohonen.html. Accessed: 04/05/2006.

Kolehmainen JI. 1939. Finnish Surnames in America. *American Speech* **14**(1): 33-38

Kuhn TS. 1977. Objectivity, Value Judgment, and Theory Choice. In *The Essential Tension: Selected Studies in Scientific Tradition and Change*, Kuhn TS (eds.), University of Chicago Press: Chicago: 320-339

Lai DW. 2004. Impact of culture on depressive symptoms of elderly Chinese immigrants. *Canadian Journal of Psychiatry* **49**(12): 820-827

Large P and Ghosh K. 2006. A methodology for estimating the population by ethnic group for areas within England. *Population Trends* **123**: 21-31

Lasker G. 1997. Census versus sample data in isonymy studies: relationship at short distances. *Human Biology* **69**(5): 733-738

Lasker GW. 1985. *Surnames and genetic structure*. 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

Lauderdale DS and Kestenbaum B. 2002. Mortality rates of elderly Asian American populations based on Medicare and Social Security data. *Demography* **39**(3): 529-540

Leino A, Mannila H, Pitknen RL. 2003. Rule Discovery and Probabilistic Modeling for Onomastic Data. In *Knowledge Discovery in Databases*, Lavraˇc N (eds.), Springler: Berlin: 291-302

Lemanski CL. 2006. Desegregation and Integration as Linked or Distinct? Evidence from a Previously 'White' Suburb in Post-apartheid Cape Town. *International Journal of Urban and Regional Research* **30**(3): 564-586

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

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

Lieberson S. 1981. An asymmetrical approach to segregation. In *Ethnic segregation in cities*, Peach C, Robinson V, Smith S (eds.), University of Georgia Press: Athens, GA: 61-82

Linguistic Minorities Project. 1985. *The Other Languages of England.* London: Routledge & Kegan Paul.

Lipson JG, Reizian AE, Meleis AI. 1987. Arab-American patients: a medical record review. *Social Science and Medicine* **24**(2): 101-107

Llitjos AF. 2002. Automatic pronunciation of proper names using language origin classes and unsupervised clustering. Presented at *Proceedings of Association for Computer Linguistics*, Philadelphia, PA.

Logan JR and Zhang W. 2004. Identifying Ethnic Neighborhoods with Census Data. In *Spatially Integrated Social Science: Examples in Best Practice*, Goodchild MF, Janelle DG (eds.), Oxford University Press: New York: 113-126

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.

Longley P. 2003. Geographical Information Systems: developments in socio-economic data infrastructures. *Progress in Human Geography* **27**(1): 114–121

Longley P, Webber R, Lloyd D. 2007. The Quantitative Analysis of Family Names: Historic Migration and the Present Day Neighbourhood Structure of Middlesbrough, United Kingdom. *Annals of the Association of American Geographers* **97**(1): 31-48

Longley PA and Goodchild MF. 2008. The use of geodemographics to improve public service delivery. In *Managing to improve public services*, Hartley J DC, Skelcher C, Wallace M (eds.), Cambridge University Press: Cambridge: Chapter 6 in press

Longley PA, Maguire DJ, Goodchild MF, Rhind D. 2005. *Geographic Information Systems and Science*. Chichester: Wiley.

Loury GC, Modood T, Teles SM. 2005. *Ethnicity, Social Mobility and Public Policy*. Cambridge: Cambridge University Press.

Lyra F. 1966. Polish Surnames in the United States. *American Speech* **41**(1): 39-44

M'charek A. 2005. *The Human Genome Diversity Project*. Cambridge: Cambridge University Press.

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

Manni F and Barrai I. 2001. Genetic Structures and Linguistic Boundaries in Italy: A Microregional Approach. *Human Biology* **73**(3): 335-347

Manni F, Guérard E, Heyer E. 2004. Geographic Patterns of (Genetic, Morphologic, Linguistic) Variation: How Barriers Can Be Detected by Using Monmonier's Algorithm *Human Biology* **76**(2): 173-190

Manni F, Toupance B, Sabbagh A, Heyer E. 2005. New method for surname studies of ancient patrilineal population structures, and possible application to improvement of Y-chromosome sampling. *Am J Phys Anthropol* **126**(2): 214-228

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

Martin D. 2002. Geography of the 2001 Census in England and Wales. *Population Trends* **108**: 7-15

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. 2000. *Race and ethnicity in modern Britain*. 2nd ed. Oxford: Oxford University Press.

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

Massey DS and Denton NA. 1988. The dimensions of residential segregation. *Social Forces* **67**: 281-315

Mateos P. 2007. Segregación residencial de minorías étnicas y el análisis geográfico del origen de nombres y apellidos [Residential segregation of ethnic minorities and geographic analysis of name origins]. *Cuadernos Geograficos* **40**: (in press)

Mateos P and Tucker DK. 2008. Forenames and Surnames in Spain in 2004. *Names: A Journal of Onomastic* **in press**

Mateos P and Webber R. 2006. Age, gender, and 'ethnicity'? How to segment populations by a slippery dimension in European multicultural geographies. Presented at *International Population Geographies Conference*, University of Liverpool, UK, 19-21 June. Available at:

http://www.geog.leeds.ac.uk/groups/pgrg/docs/conf2006/TIPG_Presentations/Session_09/Mateos.ppt Accessed: 18/10/2006.

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

McDonald F and McDonald ES. 1980. The Ethnic Origins of the American People, 1790. *The William and Mary Quarterly* **37**(2): 179-199

McEvoy B and Bradley DG. 2006. Y-chromosomes and the extent of patrilineal ancestry in Irish surnames. *Human Genetics* **119**(1-2): 212-219

Mitchell R, Shaw M, Dorling D. 2000. *Inequalities in life and death: what if Britain were more equal?* Bristol: Policy Press.

Modood T. 2005. *Multicultural Politics: Racism, Ethnicity and Muslims in Britain*. Edimburg: Edimburg University Press.

Morning A. 2008. Ethnic Classification in Global Perspective: A Cross-National Survey of the 2000 Census Round. *Population Research and Policy Review* **27**: (in press)

Nanchahal K, Mangtani P, Alston M, dos Santos Silva I. 2001. Development and validation of a computerized South Asian Names and Group Recognition Algorithm (SANGRA) for use in British Health-related studies. *Journal of Public Health Medicine* **23**(4): 278-285

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

Nature Genetics. 2001. Editorial. Genes, drugs and race. *Nature Genetics* **29**: 265-269

Nazroo J. 1997. *The health of Britain's ethnic minorities.* London: Policy Studies Institute.

Nazroo J. 2003a. Patterns of and explanations for ethnic inequalities in health. In *Explaining ethnic differences: changing patterns of disadvantage in Britain*, Mason D (eds.), Policy Press: Bristol

Nazroo JY. 2003b. The Structuring of Ethnic Inequalities in Health: Economic Position, Racial Discrimination, and Racism. *Am J Public Health* **93**(2): 277-284

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

NHS Health and Social Care Information Centre. 2005. *Health Survey for England 2004: The Health of Minority Ethnic Groups - headline tables.* Public Health Statistics. London. Available at: http://www.ic.nhs.uk/pubs/hlthsvyeng2004ethnic/HSE2004Headlineresults.pdf/file.

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.

Nicoll A, Bassett K, Ulijaszek SJ. 1986. What's in a name? Accuracy of using surnames and forenames in ascribing Asian ethnic identity in English populations. *Journal of Epidemiology and Community Health* **40**(4): 364-368

Nobles M. 2000. *Shades of Citizenship: Race and the Census in Modern Politics*. Stanford: Stanford University Press.

O'Sullivan D and Wong DWS. 2007. A Surface-Based Approach to Measuring Spatial Segregation. *Geographical Analysis* **39**(2): 147-169

Office for Management and Budget. 1978. Directive No. 15: Racial and Ethnic Standards for Federal Statistics and Administrative Reporting. *Federal Register* **43**(May 4): 19269

Office for National Statistics. 2000. *The 2001 Census questions.* Census News 44, Annex 20-28. Available at: http://www.statistics.gov.uk/census2001/pdfs/cenews44.pdf. Accessed: 12/01/2007.

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.

Office for National Statistics. 2006a. *2011 Census test questionnaire for England and Wales.* Available at: http://www.statistics.gov.uk/census/pdfs/2007_test_H1_form.pdf. Accessed: 28/03/2007.

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

Olson S. 2002. *Mapping human history: genes, race, and our common origins*. New York: First Mariner Books.

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

Oppé TE. 1967. The Health of West Indian Children. *Proceedings of the Royal Society of Medicine* **57**: 321-323

Oppenheimer GM. 2001. Paradigm lost: race, ethnicity, and the search for a new population taxonomy. *Am J Public Health* **91**(7): 1049-1055

Owen D, ed. 1996. *Size, structure and growth of the ethnic minority populations*. Office for National Statistics, HMSO London

Owen D. 2006. Spatial analysis of segregation in the UK. Presented at *Royal Geographic Society with Institute of British Geographers Annual Conference*, London, 30/08/2006.

Parekh S and Parekh S. 2003. *Asian Babies' Names from the Hindu, Muslim and Sikh traditions*. Tadworth, Surrey, UK: Elliot Right Way Books.

Park R, Bugess E, McKenzie R. 1925. *The City: Suggestions for Investigation of Human Behaviour in the Urban Environment*. Chicago: University of Chicago Press.

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.

Passel JS and Word DL. 1980. Constructing the list of Spanish surnames for the 1980 Census an application of Bayes theorem. Presented at *Annual meeting of the Population Association of America*, Denver, Colorado, April 1980.

Passel JS, Word DL, McKenney ND, Kim Y. 1982. Postcensal estimates of the Asian population in the United States description of methods using surname and administrative records. Presented at *Annual meeting of the Population Association of America*, San Diego, California, April 1982.

Patman F and Thompson P. 2003. Names: A New Frontier in Text Mining. Presented at *Intelligence and Security Informatics (ISI), First NSF/NIJ Symposium*,Lecture Notes in Computer Science, Tucson, AZ, June 2-3. Available at: http://www.ecom.arizona.edu/ISI/long_paper_sample.pdf.

Peach C. 1980. Ethnic segregation and intermarriage. *Annals of the Association of American Geographers* **70**(3): 371

Peach C. 1981. Conflicting interpretations of segregation. In *Social interaction and ethnic segregation* Jackson P, Smith SJ (eds.), Academic Press: London

Peach C. 1996a. Does Britain have ghettos? *Transactions of the Institute of British Geographers* **21**: 216-235

Peach C. 1996b. *Ethnicity in the 1991 Census. Volume 2. The ethnic minorities of Great Britain*. London: Office for National Statistics, HMSO.

Peach C. 1996c. Good segregation, bad segregation. *Planning Perspectives* **11**: 379-398

Peach C. 1996d. The meaning of segregation. *Planning, practice and research* **11**(2): 137

Peach C. 1998. South Asian and Caribbean ethnic minority housing choice in Britain. *Urban studies* **35**(10): 1657-1680

Peach C. 1999a. London and New York: contrasts in British and American models of segregation with a comment by Nathan Glazer. *International Journal of Population Geography* **5**(5): 319-347

Peach C. 1999b. Social Geography. *Progress in Human Geography* **23**(2): 282-288

Peach C. 2000. Discovering white ethnicity and parachuting plurality. *Progress in Human Geography* **24**(4): 620-626

Peach C. 2006. Islam, ethnicity and South Asian religions in the London 2001 Census. *Transactions of the Institute of British Geographers* **31**(3): 353-370

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

Peach C, Robinson V, Smith S. 1981. *Ethnic segregation in cities*. Athens, GA: University of Georgia Press.

Pearce N, Foliaki S, Sporle A, Cunningham C. 2004. Genetics, race, ethnicity, and health. *British Medical Journal* **328**: 1070-1072

Perkins RC. 1993. *Evaluating the Passel-Word Spanish surname list 1990 decennial census post enumeration survey results.* Technical Working Paper 4. US Bureau of the Census, Population Division. Washington DC. Available at: http://www.census.gov/population/www/documentation/twps0004.html. Accessed: 29/05/05.

Petersen W. 2001. Surnames in US Population Records. *Population and Development Review* **27**(2): 315

Pfeffer N. 1998. Theories in health care and research: Theories of race, ethnicity and culture. *British Medical Journal* **317**(7169): 1381-1384

Phillips D. 1998. Black Minority Ethnic Concentration, Segregation and Dispersal in Britain. *Urban Studies* **35**(10): 1681-1703

Phillips T. 2005. *After 7/7: Sleepwalking to segregation.* Speech given to Manchester Council for Community Relations. Manchester. Available at: http://www.cre.gov.uk/Default.aspx.LocID-0hgnew07s.RefLocID-0hg00900c002.Lang-EN.htm. Accessed: 04/03/2006.

Philpott TL. 1978. *The Slum and the Ghetto: Neighborhood Deterioration and Middle Class Reform, Chicago, 1880-1930*. New York: Oxford University Press.

Piazza A, Rendine S, Zei G, Moroni A, Cavalli-Sforza LL. 1987. Migration rates of human populations from surname distribution. *Nature* **329**: 714 - 716

Pissanetzky S. 1984. *Sparse Matrix Technology*. New York, NY: Academic Press.

PLASC/NPD User Group. 2007. *National Pupil Database: Current content & structure.* University of Bristol. Available at: http://www.bris.ac.uk/depts/CMPO/PLUG/userguide/guide.htm. Accessed: 09/05/2007.

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

Polednak AP. 1993. Lung cancer rates in the Hispanic population of Connecticut, 1980-88. *Public Health Rep* **108**(4): 471-476

Poulain M, Foulon M, Degioanni A, Darlu P. 2000. Flemish immigration in Wallonia and in France: patronyms as data. *The History of the Family* **5**(2): 227

Poulsen M. 2005. The 'new geography' of ethnicity in Britain? Presented at *Royal Geographical Society with the Institute of British Geographers Annual Conference*, London, 31/08/2005.

Poulsen M, Johnson R, Forrest J. 2002. Plural Cities and Ethnic Enclaves: Introducing a Measurement Procedure for Comparative Study. *International Journal of Urban and Regional Research* **26**(2): 229-243

Poulsen M, Johnston R, Forrest J. 2001. Intraurban ethnic enclaves: introducing a knowledge-based classification method. *Environment and Planning A* **33**: 2071-2082

Purvis TL. 1984. The European Ancestry of the United States Population, 1790. *The William and Mary Quarterly* **41**(1): 85

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

Rahman MM, Luong NT, Divan HA, Jesser C, Golz SD, et al. 2005. Prevalence and predictors of smoking behavior among Vietnamese men living in California. *Nicotine and Tobacco Research* **7**(1): 103-109

Rankin J and Bhopal R. 1999. Current census categories are not a good match for identity. *British Medical Journal* **318**(7199): 1696

Ratcliffe P, ed. 1996. *Ethnicity in the 1991 Census. Volume 3. Social geography and ethnicity in Britain: geographical spread, spatial concentration and internal migration*. Office for National Statistics, HMSO London

Razum O, Zeeb H, Akgun S. 2001. How useful is a name-based algorithm in health research among Turkish migrants in Germany? *Tropical Medicine and International Health* **6**(8): 654-661

Reaney PH. 1958. *A Dictionary of British Surnames*. London: Routledge and Kegan Paul.

Rees P and Boden P. 2006. *Estimating London's new migrant population: Stage 1 - review of methodology.* Greater London Authority. London. Available at: http://www.london.gov.uk/mayor/refugees/docs/nm-pop.pdf Accessed: 12/10/2006.

Rees P and Butt F. 2004. Ethnic change and diversity in England, 1981-2001. *Area* **36**(2): 174-186

Renan E. 1990 [1882]. What is a nation. In *Nation and narration*, Bhabha H (eds.), Routledge: London: 7-19

Rigoutsos I and Floratos A. 1998. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm [published erratum appears in Bioinformatics 1998;14(2):229]. *Bioinformatics* **14**(1): 55-67

Rissel C, Ward JE, Jorm L. 1999. Estimates of smoking and related behaviour in an immigrant Lebanese community: does survey method matter? *Australian and New Zealand Journal of Public Health* **23**(5): 534-537

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

Rodriguez-Larralde A, Barrai I, Nesti C, Mamolini E, Scapoli C. 1998. Isonymy and isolation by distance in Germany. *Human Biology* **70**: 1041-1056

Rodriguez-Larralde A, Gonzales-Martin A, Scapoli C, Barrai I. 2003. The Names of Spain: A Study of the Isonymy Structure of Spain. *American Journal of Physical Anthropology* **121**: 280-292

Rodriguez-Larralde A, Morales J, Barrai I. 2000. Surname Frequency and the Isonymy Structure of Venezuela. *American Journal of Human Biology* **12**: 352-362

Rodriguez-Larralde A, Scapoli C, Beretta M, Nesti C, Mamolini E, et al. 1998. Isonymy and the genetic structure of Switzerland. II. Isolation by distance. *Annals of Human Biology* **25**: 533-540

Rogers C. 1995. *The Surname Detective. Investigating Surname Distributions in England 1086-Present Day*. Manchester: Manchester University Press.

Rossiter WS. 1909. *A Century of Population Growth, from the First Census of the United States to the Twelfth, 1790-1900.* US Bureau of the Census. Government Printing Office. Washington DC.

Ruhlen M. 1987. *A guide to the World's Languages*. Standford, CA: Stanford University Press.

Ruhlen M. 1994. *On the Origin of Languages. Studies in Linguistic Taxonomy*. Standford, CA: Stanford University Press.

Salt J. 2006. *Current Trends in International Migration in Europe.* European Committee on Migration (CDMG) 51st meeting, 15th March. Council of Europe. Strasbourg. Available at: http://www.coe.int/t/e/social_cohesion/migration/documentation/Publications_and_reports/2006_Salt_report_en.pdf.

SAS. 2006. Cary, North Carolina, US. Available at: http://www.sas.com/software/sas9/. Accessed: 30/11/2006.

Scapoli C, Goebl H, Sobota S, Mamolini E, Rodriguez-Larralde A, et al. 2005. Surnames and dialects in France: Population structure and cultural evolution. *Journal Of Theoretical Biology* **237**(1): 75-86

Scapoli C, Mamolini E, Carrieri A, Rodriguez-Larralde A, Barrai I. 2007. Surnames in Western Europe: A comparison of the subcontinental populations through isonymy. *Theoretical Population Biology* **71** 37-48

Schürer KE. 2004. Surnames and the search for regions. *Local Population Studies* **72**

Science. 2005. So much more to know... *Science* **309**(5731): 78 - 102

Seeman MV. 1980. Name and identity. *Can J Psychiatry* **25**(2): 129-137

Senior PA and Bhopal R. 1994. Ethnicity as a variable in epidemiological research. *British Medical Journal* **309**(6950): 327-330

Sheth T, Nair C, Nargundkar M, Anand S, Yusuf S. 1999. Cardiovascular and cancer mortality among Canadians of European, south Asian and Chinese origin from 1979 to 1993: an analysis of 1.2 million deaths. *Canadian Medical Association Journal* **161**(2): 132-138

Shevky E and Williams M. 1949. *The Social Areas of Los Angeles, Analysis and Typology*. Berkeley: University of California Press.

Shriver MD and Kittles RA. 2004. Genetic ancestry and the search for personalized genetic histories. *Nature Reviews Genetics* **5**(8): 611-618

Silva MJ, Martins B, Chaves M, Afonso AP, Cardoso N. 2006. Adding geographic scopes to web resources. *Computers, Environment and Urban Systems* **30**(4): 378-399

Simpson EH. 1949. Measurement of diversity. *Nature* **163**: 688

Simpson L. 2004. Statistics of racial segregation: measures, evidence and policy. *Urban Studies* **41**: 661-681

Simpson L. 2005a. Measuring residential segregation. Presented at *Census: present and future*,ESRC/JISC Census Programme, Leicester, 16 November 2005. Available at: http://www.ccsr.ac.uk/research/migseg.htm Accessed: 24/04/2006.

Simpson L. 2005b. On the measurement and meaning of residential segregation: a reply to Johnston, Poulsen and Forrest. *Urban Studies* **42**(7): 1229 - 1230

Simpson L. 2006. The numerical liberation of dark areas. *SAGE Race Relations Abstracts* **31**(2): 5-25

Simpson L. 2007. Ghettos of the mind: the empirical behaviour of indices of segregation and diversity. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **170**(2): 405-424

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

Smith-Bannister S. 1997. *Names and Naming Patterns in England 1538-1700*. Clarendon, PA, USA: Oxford University Press.

Smith G, Neaton J, Wentworth D, Stamler R, Stamler J. 1998. Mortality differences between black and white men in the USA: contribution of income and other risk factors among men screened for the MRFIT. *Lancet* **351**: 934-939

Sorenson Molecular Genealogy Foundation. 2007. *Why Molecular Genealogy?* Available at: http://www.smgf.org/. Accessed: 09/03/2007.

Sparks E. 2005. Experian Consumer Dynamics Characteristics. Personal Communication

Statbel. 2006. *Noms de famille les plus fréquents - Belgique et Régions.* Available at: http://statbel.fgov.be/figures/d21a_fr.asp. Accessed: 12/09/2006.

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

Stillwell J and Duke-Williams O. 2005. Ethnic population distribution, immigration and internal migration in Britain. What evidence of linkage at the district scale. Presented at *British Society for Population Studies Annual Conference*, University of Kent at Canterbury 12-14 September. Available at: http://www.lse.ac.uk/collections/BSPS/pdfs/Stillwell_ethnicpopdist_2005.pdf Accessed: 20/06/2006.

Stillwell J and Phillips D. 2006. Diversity and Change: Understanding the Ethnic Geographies of Leeds. *Journal of Ethnic and Migration Studies* **32**(7): 1131 - 1152

Surname Profiler. 2006. Available at: www.spatial-literacy.org.

Szczepura A. 2005. Access to health care for ethnic minority populations. *Postgraduate Medical Journal* **81**(953): 141-147

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

The Economist. 2007a. Immigration. *The Economist*, 19th May: 34

The Economist. 2007b. What's in a name? *The Economist*, Technology Quaterly Survey, 10th March: 27

The Guardian. 2004. Gateway to the past. 08 April. Available at: http://education.guardian.co.uk/higher/artsandhumanities/story/0,,1188368,00.html. Accessed: 18/05/2006.

Thiel H and Finezza AJ. 1971. A note on the measurement of racial integration of schools by means of informational concepts. *Journal of Mathematical Sociology* **1**: 187-194

Tibón G. 2001. *Diccionario etimológico comparado de los apellidos Españoles, Hispanoamericanos y Filipinos*. 3ª Edición. México D.F.: Fondo de Cultura Económica.

Tu SP, Yasui Y, Kuniyuki A, Schwartz SM, Jackson JC, et al. 2002. Breast cancer screening: stages of adoption among Cambodian American women. *Cancer Detection and Prevention* **26**(1): 33-41

Tucker DK. 2001. Distribution of forenames, surnames, and forename-surname pairs in the United States. *Names* **49**: 69-96

Tucker DK. 2002. Distribution of forenames, surnames, and forename-surname pairs in Canada. *Names* **50**(2): 105-132

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. 2004a. The forenames and surnames from the GB 1998 Electoral Roll compared with those from the UK 1881 Census. *Nomina* **27**: 5-40

Tucker DK. 2004b. What happened to the UK 1881 Census surnames by 1997. *Nomina* **27**: 91-118

Tucker DK. 2005. The cultural-ethnic-language group technique as used in the Dictionary of American Family Names (DAFN). *Onomastica Canadiana* **87**(2): 71-84

Tucker DK. 2006. A Comparison of Irish surnames in the United States with those of Eire. *Names* **54**: 55-75

Tucker DK. 2007a. Personal communication.

Tucker DK. 2007b. Surname distribution prints from the UK 1998 Electoral Roll compared with those from other distributions. *Nomina* **30** In Press

University of Cambridge. 2004. *English Dictionary*. Cambridge: Cambridge University Press.

US Bureau of the Census. 1953. *Persons of Spanish Surname.* US Census of Population:1950, Vol. IV, Special Report P-E, No. 3C, U.S. Department of Commerce. Washington D.C.: US Government Printing Office.

US Bureau of the Census. 1980. *1980 census of population and housing: Spanish list technical documentation*. Washington DC: Data User Services Division.

US Census Bureau. 2006. *US Census Bureau Geneaology Resources.* Available at: http://www.census.gov/genealogy/www/. Accessed: 12/05/2006.

US Senate. 1928. *Immigration quotas on the basis of national origin.* Rep. Miscellaneous Documents 8870 vol.1 nr 65, 70th Congress 1st Session. Washington, DC.

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

Vickers D and Rees P. 2007. Creating the UK National Statistics 2001 output area classification. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **170**(2): 379-403

Voas D and Williamson P. 2000. The Scale of Dissimilarity: Concepts, Measurement and an Application to Socio-Economic Variation Across England and Wales. *Transactions of the Institute of British Geographers* **25**(4): 465-481

Wade N. 2005. Firm matches Icelanders' genetic makeup with region they come from *New York Times* January 2. Available at: http://www.sfgate.com/cgi-bin/article.cgi?file=/c/a/2005/01/02/MNG47AJ74D1.DTL. Accessed: 04/06/2005.

Ward JH. 1963. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* **58**: 236-244

Wasserman S and Galaskiewicz J. 1995. *Advances in Social Network Analysis: Research in the Social and Behavioral Sciences*. Thousand Oaks, California: SAGE

Weber M. 1980 [1921]. *Wirtschaft und Gesellschaft (Eng. Tr. Economy and Society)*. Tübingen: Mohr.

Whitehead M. 1992. The Health Divide. In *Inequalities in health: the Black Report and the Health Divide*, Townsend P, Whitehead M, Davidson N (eds.), Penguin: London

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

Williams K. 2007. *US Patent Application: Name classifier algorithm.* Available at: www.uspto.gov (search for patent number '20070005597'). Accessed: 19/03/2007.

Williams K and Patman F. 2005. Personal Entity Extraction Filtering using Name Data Stores. Presented at *International Conference on Intelligence Analysis*, McLean, VA, 2-6 May. Available at: https://analysis.mitre.org/proceedings/Final_Papers_Files/33_Camera_Ready_Paper.pdf Accessed: 05/02/2006.

Winnie WW, Jr. 1960. The Spanish surname criterion for identifying Hispanos in the southwestern United States: A preliminary evaluation. *Social Forces* **38**(4): 363-366

Wong DWS. 2003. Spatial decomposition of segregation indices: a framework toward measuring segregation at multiple Levels *Geographical Analysis* **35**(3): 179-194

Wong DWS. 2004. Comparing traditional and spatial segregation measures: a spatial scale perspective. *Urban Geography* **25**(1): 66-82

Word DL, Passel JS, Causey BD, Fernandez EF. 1978. Determining a List of Spanish Surnames by Analysis of Geographical Distributions. Presented at *Annual Meeting of Southern Regional Demographic Group*, San Antonio, Texas, October.

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.

Yasuda N, Cavalli-Sforza LL, Skolnick M, Moroni A. 1974. The evolution of surnames: an analysis of their distribution and extinction. *Theor Popul Biol.* **1**: 123-142

Yavari P, Hislop TG, Abanto Z. 2005. Methodology to identify Iranian immigrants for epidemiological studies. *Asian Pacific Journal of Cancer Prevention* **6**(4): 455-457

Appendix 1: List of Published Outputs from PhD

The following articles and conference papers are a direct output of this PhD research.

Full copies of the papers and presentations are available on the author's research webpage at: www.casa.ucl.ac.uk/pablo/publications.htm

(a) Refereed Academic Journals

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 (The full paper is included at the end of the thesis as Appendix 7)

MATEOS, P. (2007) Segregación residencial de minorías étnicas y el análisis geográfico del origen de nombres y apellidos [Residential segregation of ethnic minorities and geographic analysis of name origins], *Cuadernos Geográficos*, 40 (in press) ISSN: 0210-5462

MATEOS, P. and TUCKER, DK. (2008) Forenames and Surnames in Spain in 2004. *Names: A Journal of Onomastic* (in press)

LONGLEY P.A. and MATEOS, P. (2005): Un nuevo y prominente papel de los SIG y el Geomarketing en la provisión de servicios públicos [A new and prominent role of GIS and Geodemographics in the delivery of public services], *GeoFocus*, 5, 1-5. ISSN: 1578-5157

(b) Working Papers and Technical Reports

MATEOS, P., WEBBER, R., and LONGLEY, PA. (2007) The Cultural, Ethnic and Linguistic Classification of Populations and Neighbourhoods using Personal Names, *CASA Working Paper 116*, Centre for Advanced Spatial Analysis, University College London . ISSN 1467-1298 (available at http://www.casa.ucl.ac.uk/working_papers/paper116.pdf)

(c) Conference Papers

MATEOS, P. and LONGLEY, PA. (2007) The Modifiable Ethnic Unit Problem (MEUP); Defining and measuring ethnicity in multicultural societies. Paper to be presented at the *British Society for Population Studies Annual Conference*, University of St.Andrews, 11-13 September.

MATEOS, P. and LONGLEY, PA. (2007) A name-based ethnicity classification to subdivide populations in groups of common origin. Paper presented at *International Conference on DNA Sampling* , Muséum National d' Historie Naturelle (MNHN) - Musée de l'Homme, Paris, France, 15-16 March.

MATEOS, P. (2006) Segregación urbana de minorías étnicas y el análisis del origen de nombres y apellidos [Residential segregation of ethnic minorities and

geographic analysis of name origins]. Paper presented at *XII Congreso Nacional de Tecnologías de Información Geográfica*, Granada, Spain, 19-22 September

MATEOS, P., LONGLEY, PA. and WEBBER, R. (2006) El Análisis Geodemográfico de Apellidos en México [Geodemographic Analysis of Surnames in Mexico . Paper presented at *Reunión Anual de la Sociedad Mexicana de Demografía (SOMEDE),* Guadalajara, Mexico, 5-9 September

MATEOS, P. (2006) You Name it! The Past and Future of Name Geography Research. Paper presented at *Royal Geographic Society (RGS-IBG) Annual Conference*, London, 1-2 September.

MATEOS, P. (2006) El estudio de migraciones y minorías étnicas a través del análisis geográfico de nombres y apellidos. Paper presented at *Congreso de Poblacion Española*, Pamplona, Spain, 29-30 June

MATEOS, P., WEBBER, R., and LONGLEY, PA. (2006) The World in names: new geographies of ethnic minorities revealed through family name and personal name analysis. Paper presented at the *European Population Conference*, Liverpool, 21-24 June

MATEOS, P., and WEBBER, R. (2006) Age, gender, and… ethnicity? How to segment populations by a slippery dimension in European multicultural geographies . Paper presented at the *International Population Geographies Conference*, Liverpool, 19-21 June. (Awarded Best Postgraduate Paper Prize)

MATEOS, P., WEBBER, R., and LONGLEY, PA. (2006) How segregated are name origins? A new method of measuring ethnic residential segregation . in Priestnall G. and Aplin P. (eds.) *Proceedings of the GIS Research UK 14th Annual Conference (GISRUK)*, University of Nottingham, UK 5-7 April, 285-291.

MATEOS, P., WEBBER, R., and LONGLEY, PA. (2006) Measuring spatial segregation of ethnic minorities in the UK at fine scales and individual level through family name and personal name analysis. Paper presented at the *Association of American Geographers (AAG) 2006 Annual Meeting*, Chicago 7-11 March.

MATEOS P., JONES CE., LONGLEY PA., and WEBBER R. (2005) Data Mining of ethnic information in patient records through birthplace & name analysis. An example in Camden PCT . Paper presented at the *Royal Geographic Society with Institute of British Geographers (RGS-IBG) Annual Conference*, London , 30 August

JONES CE., MATEOS P., LONGLEY PA., and WEBBER R. (2005) The discriminatory power of geodemographics to inform health promotion strategies: Application in breast screening programs . Paper presented at the fifth *International Interdisciplinary Conference on Geomedical Systems (Geomed)*, Fitzwilliam College , University of Cambridge , 16-17 September

JONES, CE., MATEOS, P. LONGLEY, PA., and WEBBER, R. (2005) Who does not eat their greens? Geodemographics, health promotion and neighbourhood health inequalities. Proceedings of the *GIS Research UK conference (GISRUK)*, University of Glasgow , Glasgow 6-8 April, p. 16

(d) Conference Posters

EVANS, J.T., SMITH E.G., ROBERTSON, R., MATEOS, P., WEBBER, R., HAWKEY, P.M. (2007) Molecular Epidemiology of Mycobacterium tuberculosis in the UK Midlands: Comparison of Predominantly Affected Patient Populations. *Health Protection Agency Conference*, Warwick University, 17-19 September

MATEOS, P., LONGLEY, PA. and WEBBER, R. (2006) El estudio de las migraciones en Latinoamerica a través del análisis geográfico de apellidos. Paper presented at *II Congreso de la Asociación Latinoamericana de Población (ALAP)*, Guadalajara, Mexico, September 3-6

JONES, CE., MATEOS, P., WEBBER, R., LONGLEY , PA., MURRAY, D. and DAVIS, M. (2005) The discriminatory power of Geodemographic classifications in targeting health promotion and disease prevention initiatives to tackle inequalities. Poster presented at the *Faculty of Public Health Annual Scientific Meeting,* Scarborough , UK , 7-9 June.

Appendix 2: Ethnicity Classifications

Ethnicity classification used in the Pupil Level Annual School Survey (PLASC)

(Source: PLASC/NPD User Group, 2007)

101= British
102 = English
103 = Scottish
104 = Welsh
105 = Irish
106 = Traveller of Irish Heritage
107 = Other White British
108 = Any Other White Background
109 = Albanian
110 = Bosnian-Herzegovinian
111 = Croation
112 = Greek/Greek Cypriot
113 = Greek
114 = Greek Cypriot
115 = Kosovan
116 = Italian
117 = Portuguese
118 = Serbian
119 = Turkish/Turkish Cypriot
120 =Turkish
121 = Turkish Cypriot
122 = White European
123 = White Eastern European
124 = White Western European
125 = Other White
126 = Gypsy/Roma
127 = White and Black Caribbean
128 = White and Black African
129 = White and Asian
130 = White and Pakistani
131 = White and Indian
132 = White and Any Other Asian Back
133 = Any Other Mixed Background
134 = Asian and Any Other Ethnic Group
135 = Asian and Black
136 = Asian and Chinese
137 = Black and Any Other Ethnic Group
138 = Black and Chinese
139 = Chinese and Any Other Ethnic Group
140 = White and Any Other Ethnic Group
141 = White and Chinese
142 = Other Mixed Background
143 = Indian
144 = Pakistani
145 = Bangladeshi
146 = Any Other Asian Background
147 = Mirpuri Pakistani
148 = Other Pakistani
149 = Kashmiri Pakistani
150 = African Asian
151 = Kashmiri Other
152 = Nepali
153 = Sinhalese
154 = Sri Lankan Tamil
155 = Other Asian
156 = Caribbean
157 = African
158 = Any Other Black Background
159 = Angolan
160 = Congolese
161 = Ghanaian
162 = Nigerian
163 = Sierra Leonian
164 = Somali
165 = Sudanese
166 = Other Black African
167 = Black European
168 = Black North American
169 = Other Black
170 = Chinese
171 = Hong Kong Chinese
172 = Malaysian Chinese
173 = Singaporean Chinese
174 = Taiwanese
175 = Other Chinese
176 = Any Other Ethnic Group
177 = Afghanistani
178 = Arab
179 = Egyptian
180 = Filipino
181 = Iranian
182 = Iraqi
183 = Japanese
184 = Korean
185 = Kurdish
186 = Latin American
187 = Lebanese
188 = Libyan
189 = Malay
190 = Moroccan
191 = Polynesian
192 = Thai
193 = Vietnamese
194 = Yemeni
195 = Other Ethnic Group
198 = Parent/pupil preferred not to say
199 = Ethnic group information not sought.
998 = Refused
999 = Information Not Obtained

Write-in categories reported in the 2001 Census ethnicity. Figures for London.

(Source: GLA Commissioned Tables C0183 - Ethnicity, Write-in by sex and age (200+ people) Greater London)

Write-in Ethnic Group	People
British, Mixed British	4,097,664
Irish	220,187
English	154,203
Scottish	7,020
Welsh	6,895
Cornish	468
Northern Irish	387
Cypriot (part not stated)	7,360
Greek	17,888
Greek Cypriot	23,340
Turkish	37,827
Turkish Cypriot	14,074
Italian	35,252
Irish Traveller	497
Gypsy/Romany	490
Polish	15,928
Baltic States (Estonian, Latvian, Lithuanian)	2,248
Commonwealth of (Russian) Independent States	11,606
Kosovan	6,896
Albanian	3,226
Bosnian	1,695
Croatian	1,954
Serbian	1,349
Other republics which made up the former Yugoslavia	2,674
Mixed: Irish and other white	7,071
Other white European, European Mixed	185,690
Other mixed white	19,239
Other white, white unspecified	171,744
White and Black Caribbean	70,093
White and Black African	33,282
White and Asian	57,561
Black and Asian	3,946
Black and Chinese	590
Black and White	4,226
Chinese and White	4,871
Asian and Chinese	660
Other Mixed, Mixed unspecified	35,027
Indian or British Indian	429,877
Pakistani or British Pakistani	140,888
Bangladeshi or British Bangladeshi	153,021

Write-in Ethnic Group	People
Punjabi	1,149
Kashmiri	734
East African Asian	5,328
Sri Lankan	53,307
Tamil	4,758
Sinhalese	565
Caribbean Asian	4,070
British Asian	14,625
Mixed Asian	2,786
Other Asian, Asian unspecified	18,334
Black Caribbean	337,260
Black African	366,626
Somali	6,172
Nigerian	1,844
Black British	46,348
Mixed Black	9,001
Other Black, Black unspecified	8,344
Chinese	79,579
Vietnamese	11,719
Japanese	19,415
Filipino	19,669
Malaysian	3,384
Hindu	778
Jewish	8,912
Muslim	707
Sikh	2,814
Arab	20,256
North African	11,218
Middle Eastern (excluding Israeli, Iranian and 'Arab')	20,537
Israeli	2,304
Iranian	16,494
Kurdish	9,659
Moroccan	4,133
Latin American	9,188
South and Central American	15,607
Multi-ethnic islands	15,952
Any other group	29,469

Total	**7,171,959**

Appendix 3: CEL Taxonomy

List of 185 CEL Types and their characteristics and proposed aggregations, including the final 66 CEL Subgroups

CEL Type Code	CEL Group	CEL Subgroup	CEL Type	People GB 2004	Geographical Area	Religion	Major Language	Major Language SIL Code	Major Language Family Tree	1991 Census Ethnic Group	2001 Census Ethnic Group	2001 Census Religion	2001 Census COB
AF110	AFRICAN	AFRICAN	AFRICA	3316	AFRICA	CHRISTIAN: PROTESTANT	Not Applicable	Not Applicable	Not Applicable	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF429	AFRICAN	AFRICAN	BENIN	7	AFRICA	MUSLIM	French	FRN	Indo-European;Italic;Romance;Italo-Western	2- Black - African	N) Black or Black British - African	MUSLIM	COB_Elsewhere
AF212	AFRICAN	AFRICAN	BOTSWANA	8	AFRICA	CHRISTIAN: PROTESTANT	Tswana	TSW	Niger-Congo;Atlantic-Congo;Volta-Congo;Benue-Congo	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF322	AFRICAN	AFRICAN	BURUNDI	18	AFRICA	CHRISTIAN	Rundi	RUD	Niger-Congo;Atlantic-Congo;Volta-Congo;Benue-Congo	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF430	AFRICAN	AFRICAN	CAMEROON	72	AFRICA	CHRISTIAN	Fulfulde	FUB	Niger-Congo;Atlantic-Congo;Atlantic;Northern	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF431	AFRICAN	AFRICAN	GAMBIA	11	AFRICA	MUSLIM	Wolof	WOF	Niger-Congo;Atlantic-Congo;Atlantic;Northern	2- Black - African	N) Black or Black British - African	MUSLIM	COB_Elsewhere
AF433	AFRICAN	AFRICAN	GUINEA	15	AFRICA	MUSLIM	French	FRN	Indo-European;Italic;Romance;Italo-Western	2- Black - African	N) Black or Black British - African	MUSLIM	COB_Elsewhere
AF434	AFRICAN	AFRICAN	IVORY COAST	122	AFRICA	MUSLIM	BaoulŽ	BCI	Niger-Congo;Atlantic-Congo;Volta-Congo;Kwa	2- Black - African	N) Black or Black British - African	MUSLIM	COB_Elsewhere
AF324	AFRICAN	AFRICAN	KENYAN AFRICAN	1197	AFRICA	CHRISTIAN: PROTESTANT	Gikuyu	KIU	Niger-Congo;Atlantic-Congo;Volta-Congo;Benue-Congo	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF435	AFRICAN	AFRICAN	LIBERIA	5	AFRICA	MUSLIM	Kpelle	KPE	Niger-Congo;Mande;Western;Central-Southwestern	2- Black - African	N) Black or Black British - African	MUSLIM	COB_Elsewhere
AF214	AFRICAN	AFRICAN	MADAGASCAR	2	AFRICA	CHRISTIAN: CATHOLIC	Malagasy	PLT	Austronesian;Malayo-Polynesian;Barito;East	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF215	AFRICAN	AFRICAN	MALAWI	23	AFRICA	CHRISTIAN: PROTESTANT	Nyanja	NYJ	Niger-Congo;Atlantic-Congo;Volta-Congo;Benue-Congo	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF216	AFRICAN	AFRICAN	NAMIBIA	1	AFRICA	CHRISTIAN: CATHOLIC	Afrikaans	AFK	Indo-European;Germanic;West;Low Saxon-Low Franconian	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
AF217	AFRICAN	AFRICAN	OTHER AFRICAN	1575	AFRICA	CHRISTIAN	Not Applicable	Not Applicable	Not Applicable	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF325	AFRICAN	AFRICAN	RWANDA	18	AFRICA	CHRISTIAN	Rwanda	RUA	Niger-Congo;Atlantic-Congo;Volta-Congo;Benue-Congo	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF437	AFRICAN	AFRICAN	SENEGAL	37	AFRICA	MUSLIM	Wolof	WOL	Niger-Congo;Atlantic-Congo;Atlantic;Northern	2- Black - African	N) Black or Black British - African	MUSLIM	COB_Elsewhere
AF218	AFRICAN	AFRICAN	SWAZILAND	5	AFRICA	CHRISTIAN: PROTESTANT	Swati	SWZ	Niger-Congo;Atlantic-Congo;Volta-Congo;Benue-Congo	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF327	AFRICAN	AFRICAN	TANZANIA	104	AFRICA	CHRISTIAN: PROTESTANT	English	ENG	Indo-European;Germanic;West;English	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF219	AFRICAN	AFRICAN	ZAIRE	41	AFRICA	CHRISTIAN: PROTESTANT	Luba-Kasai	LUB	Niger-Congo;Atlantic-Congo;Volta-Congo;Benue-Congo	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF220	AFRICAN	AFRICAN	ZAMBIA	274	AFRICA	CHRISTIAN: PROTESTANT	Bemba	BEM	Niger-Congo;Atlantic-Congo;Volta-Congo;Benue-Congo	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF221	AFRICAN	AFRICAN	ZIMBABWE	991	AFRICA	CHRISTIAN: PROTESTANT	Shona	SHD	Creole;French based	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF211	AFRICAN	BLACK SOUTHERN AFRICA	BLACK SOUTHERN AFRICA	5198	AFRICA	CHRISTIAN: PROTESTANT	Zulu	ZUU	Niger-Congo;Atlantic-Congo;Volta-Congo;Benue-Congo	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF213	AFRICAN	CONGOLESE	CONGO	1164	AFRICA	CHRISTIAN	Luba-Kasai	LUB	Niger-Congo;Atlantic-Congo;Volta-Congo;Benue-Congo	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF323	AFRICAN	ETHIOPIAN	ETHIOPIA	1238	AFRICA	CHRISTIAN: OTHER	Amharic	AMH	Afro-Asiatic;Semitic;South;Ethiopian	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF432	AFRICAN	GHANAIAN	GHANA	46095	AFRICA	CHRISTIAN	Akan	TWS	Niger-Congo;Atlantic-Congo;Volta-Congo;Kwa	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF436	AFRICAN	NIGERIAN	NIGERIA	88243	AFRICA	CHRISTIAN	Yoruba	YOR	Niger-Congo;Atlantic-Congo;Volta-Congo;Benue-Congo	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
AF438	AFRICAN	SIERRA LEONIAN	SIERRA LEONE	6155	AFRICA	MUSLIM	English	ENG	Indo-European;Germanic;West;English	2- Black - African	N) Black or Black British - African	MUSLIM	COB_Elsewhere
AF328	AFRICAN	UGANDAN	UGANDA	1018	AFRICA	CHRISTIAN: PROTESTANT	Ganda	LAP	Niger-Congo;Atlantic-Congo;Volta-Congo;Benue-Congo	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
CL110	CELTIC	IRISH	CELTIC	45653	BRITISH ISLES	CHRISTIAN	English	ENG	Indo-European;Germanic;West;English	0- White	A) White - British	CHRISTIAN	COB_Republic of Ireland
CL211	CELTIC	IRISH	IRELAND	3172876	BRITISH ISLES	CHRISTIAN: CATHOLIC	English	ENG	Indo-European;Germanic;West;English	0- White	B) White - Irish	CHRISTIAN	COB_Republic of Ireland
CL212	CELTIC	IRISH	NORTHERN IRELAND	223988	BRITISH ISLES	CHRISTIAN	English	ENG	Indo-European;Germanic;West;English	0- White	A) White - British	CHRISTIAN	COB_Northern Ireland
CL213	CELTIC	SCOTTISH	SCOTLAND	4749864	BRITISH ISLES	CHRISTIAN: PROTESTANT	English	ENG	Indo-European;Germanic;West;English	0- White	A) White - British	CHRISTIAN	COB_Scotland
CL314	CELTIC	WELSH	WALES	3065041	BRITISH ISLES	CHRISTIAN: PROTESTANT	Welsh	WLS	Indo-European;Celtic;Insular;Brythonic	0- White	A) White - British	CHRISTIAN	COB_Wales
EA212	EAST ASIAN	CHINESE	CHINA	21185	EAST ASIA	BHUDDIST	Chinese, Mandarin	CHN	Sino-Tibetan;Chinese	7- Chinese	R) Other Ethnic Groups - Chinese	BHUDDIST	COB_Elsewhere
EA218	EAST ASIAN	CHINESE	MALAYSIAN CHINESE	3238	EAST ASIA	BHUDDIST	Chinese, Min Nan	CFR	Sino-Tibetan;Chinese	7- Chinese	R) Other Ethnic Groups - Chinese	BHUDDIST	COB_Elsewhere
EA327	EAST ASIAN	EAST ASIAN	CAMBODIA	59	EAST ASIA	BHUDDIST	Khmer, Central	KHM	Austro-Asiatic; Mon-Khmer; Eastern Mon-Khmer; Khmer	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	BHUDDIST	COB_Elsewhere
EA110	EAST ASIAN	EAST ASIAN	EAST ASIA	627	EAST ASIA	BHUDDIST	Chinese, Mandarin	CHN	Sino-Tibetan;Chinese	7- Chinese	R) Other Ethnic Groups - Chinese	BHUDDIST	COB_Elsewhere
EA213	EAST ASIAN	EAST ASIAN	EAST ASIAN CARIBBEAN	2	AMERICAS	BHUDDIST	Chinese, Mandarin	CHN	Sino-Tibetan;Chinese	7- Chinese	R) Other Ethnic Groups - Chinese	BHUDDIST	COB_Elsewhere
EA414	EAST ASIAN	EAST ASIAN	FIJI	10	EAST ASIA	HINDU	Hindustani	HIF	Indo-European;Indo-Iranian;Indo-Aryan;East Central zone	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	HINDU	COB_Elsewhere
EA429	EAST ASIAN	EAST ASIAN	HAWAII	2	EAST ASIA	CHRISTIAN	Hawaii Creole English	HWC	Creole; English based; Pacific	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	CHRISTIAN	COB_Elsewhere
EA316	EAST ASIAN	EAST ASIAN	INDONESIA	116	EAST ASIA	MUSLIM	Javanese	JAN	Austronesian;Malayo-Polynesian;Western Malayo-Polynesian;Sundic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
EA328	EAST ASIAN	EAST ASIAN	LAOS	120	EAST ASIA	BHUDDIST	Lao	LAO	Tai-Kadai; Kam-Tai; Be-Tai; Tai-Sek; Tai; Southwestern; Lao-Phutai	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	BHUDDIST	COB_Elsewhere
EA430	EAST ASIAN	EAST ASIAN	MAORI	1	EAST ASIA	CHRISTIAN	Maori	MRI	Austronesian; Malayo-Polynesian; Central-Eastern; Eastern Malayo-Polynesian	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	CHRISTIAN	COB_Elsewhere
EA431	EAST ASIAN	EAST ASIAN	MAURITIUS	2	EAST ASIA	HINDU	Hindi	HND	Indo-European;Indo-Iranian;Indo-Aryan;Central zone	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	HINDU	COB_Elsewhere
EA319	EAST ASIAN	EAST ASIAN	MYANMAR	1601	EAST ASIA	BHUDDIST	Burmese	BMS	Sino-Tibetan;Tibeto-Burman;Lolo-Burmese;Burmish	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	BHUDDIST	COB_Elsewhere
EA420	EAST ASIAN	EAST ASIAN	POLYNESIA	54	EAST ASIA	CHRISTIAN	Tahitian	TAH	Austronesian; Malayo-Polynesian; Central-Eastern; Eastern Malayo-Polynesian; Oceanic; Central-Eastern Oceanic; Remote Oceanic; Central Pacific; East Fijian-Polynesian; Polynesian; Nuclear; East; Central; Tahitic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	CHRISTIAN	COB_Elsewhere
EA432	EAST ASIAN	EAST ASIAN	SAMOA	10	EAST ASIA	CHRISTIAN	Samoan	SMO	Austronesian; Malayo-Polynesian; Central-Eastern; Eastern Malayo-Polynesian	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	CHRISTIAN	COB_Elsewhere
EA221	EAST ASIAN	EAST ASIAN	SINGAPORE	583	EAST ASIA	BHUDDIST	Chinese, Min Nan	CFR	Sino-Tibetan;Chinese	7- Chinese	R) Other Ethnic Groups - Chinese	BHUDDIST	COB_Elsewhere
EA422	EAST ASIAN	EAST ASIAN	SOLOMON ISLANDS	8	EAST ASIA	CHRISTIAN	English	ENG	Indo-European;Germanic;West;English	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	CHRISTIAN	COB_Elsewhere
EA211	EAST ASIAN	EAST ASIAN	SOUTH EAST ASIA	371	EAST ASIA	BHUDDIST	Chinese, Min Nan	CFR	Sino-Tibetan;Chinese	7- Chinese	R) Other Ethnic Groups - Chinese	BHUDDIST	COB_Elsewhere
EA324	EAST ASIAN	EAST ASIAN	THAILAND	407	EAST ASIA	BHUDDIST	Thai	THJ	Tai-Kadai;Kam-Tai;Be-Tai;Tai-Sek	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	BHUDDIST	COB_Elsewhere
EA225	EAST ASIAN	EAST ASIAN	TIBET	13	EAST ASIA	BHUDDIST	Tibetan	BOD	Sino-Tibetan;Tibeto-Burman;Himalayish;Tibeto-Kanauri	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	BHUDDIST	COB_Elsewhere
EA433	EAST ASIAN	EAST ASIAN	TONGA	9	EAST ASIA	CHRISTIAN	Tongan	TON	Austronesian; Malayo-Polynesian; Central-Eastern; Eastern Malayo-Polynesian	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	CHRISTIAN	COB_Elsewhere

CEL Type Code	CEL Group	CEL Subgroup	CEL Type	People GB 2004	Geographical Area	Religion	Major Language	Major Language SIL Code	Major Language Family Tree	1991 Census Ethnic Group	2001 Census Ethnic Group	2001 Census Religion	2001 Census COB
EA434	EAST ASIAN	EAST ASIAN	TUVALU	2	EAST ASIA	CHRISTIAN	Tuvaluan	TVL	Austronesian; Malayo-Polynesian; Central-Eastern; Eastern Malayo-Polynesian	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	CHRISTIAN	COB_Elsewhere
EA215	EAST ASIAN	HONG KONGESE	HONG KONG	119566	EAST ASIA	BHUDDIST	Chinese, Cantonese	YUH	Sino-Tibetan;Chinese	7- Chinese	R) Other Ethnic Groups - Chinese	BHUDDIST	COB_Elsewhere
EA323	EAST ASIAN	KOREAN	SOUTH KOREA	2315	EAST ASIA	BHUDDIST	Korean	KKN	Language Isolate;	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	BHUDDIST	COB_Elsewhere
EA317	EAST ASIAN	MALAYSIA	MALAYSIA	2092	EAST ASIA	MUSLIM	Malay	MLI	Austronesian;Malayo-Polynesian;Western Malayo-Polynesian;Sundic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
EA326	EAST ASIAN	VIETNAM	VIETNAM	15723	EAST ASIA	BHUDDIST	Vietnamese	VIE	Austro-Asiatic;Mon-Khmer;Viet-Muong;Vietnamese	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	BHUDDIST	COB_Elsewhere
EN315	ENGLISH	BLACK CARIBBEAN	BLACK CARIBBEAN	23665	AMERICAS	CHRISTIAN: PROTESTANT	English	ENG	Indo-European;Germanic;West;English	1- Black - Caribbean	M) Black or Black British - Caribbean	CHRISTIAN	COB_Elsewhere
EN314	ENGLISH	ENGLISH	BRITISH SOUTH AFRICA	45	AFRICA	CHRISTIAN: PROTESTANT	English	ENG	Indo-European;Germanic;West;English	0- White	A) White - British	CHRISTIAN	
EN213	ENGLISH	ENGLISH	CHANNEL ISLANDS	23995	BRITISH ISLES	CHRISTIAN: PROTESTANT	English	ENG	Indo-European;Germanic;West;English	0- White	A) White - British	CHRISTIAN	COB_England
EN211	ENGLISH	ENGLISH	CORNWALL	107068	BRITISH ISLES	CHRISTIAN: PROTESTANT	English	ENG	Indo-European;Celtic;Insular;Brythonic	0- White	A) White - British	CHRISTIAN	COB_England
EN110	ENGLISH	ENGLISH	ENGLAND	31118965	BRITISH ISLES	CHRISTIAN: PROTESTANT	English	ENG	Indo-European;Germanic;West;English	0- White	A) White - British	CHRISTIAN	COB_England
EU215	EUROPEAN	AFRIKAANS	AFRIKAANS	7805	AFRICA	CHRISTIAN: PROTESTANT	Afrikaans	AFK	Indo-European;Germanic;West;Low Saxon-Low Franconian	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU734	EUROPEAN	ALBANIA	ALBANIA	3440	EASTERN EUROPE	CHRISTIAN: GREEK ORTHODOX	Albanian	ALS	Indo-European;Albanian;Tosk;	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU727	EUROPEAN	BALKAN	BALKAN	16274	EASTERN EUROPE	CHRISTIAN	Serbian	SDD	Indo-European;Slavic;South;Western	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU839	EUROPEAN	BALKAN	BULGARIA	109	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Bulgarian	BLG	Indo-European;Slavic;South;Eastern	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU733	EUROPEAN	BALKAN	CROATIA	1362	EASTERN EUROPE	CHRISTIAN: CATHOLIC	Croatian	CRX	Indo-European;Slavic;South;Western	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU731	EUROPEAN	BALKAN	MACEDONIA	371	EASTERN EUROPE	CHRISTIAN: GREEK ORTHODOX	Macedonian	MKJ	Indo-European;Slavic;South;Eastern	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU730	EUROPEAN	BALKAN	MONTENEGRO	44	EASTERN EUROPE	MUSLIM	Serbian	SDD	Indo-European;Slavic;South;Western	0- White	C) White - Any other White background	MUSLIM	COB_Elsewhere
EU728	EUROPEAN	BALKAN	SERBIA	5279	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Serbian	SDD	Indo-European;Slavic;South;Western	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU732	EUROPEAN	BALKAN	SLOVENIA	1282	EASTERN EUROPE	CHRISTIAN: CATHOLIC	Slovenian	SLV	Indo-European;Slavic;South;Western	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU624	EUROPEAN	BALTIC	ESTONIA	778	EASTERN EUROPE	CHRISTIAN: PROTESTANT	Estonian	EST	Uralic;Finno-Ugric;Finno-Permic;Finno-Cheremisic	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU625	EUROPEAN	BALTIC	LATVIA	1559	EASTERN EUROPE	CHRISTIAN: CATHOLIC	Latvian	LAV	Indo-European; Baltic; Eastern	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU626	EUROPEAN	BALTIC	LITHUANIA	1790	EASTERN EUROPE	CHRISTIAN: CATHOLIC	Lithuanian	LIT	Indo-European;Baltic;Eastern	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU836	EUROPEAN	CZECH & SLOVAKIAN	CZECH REPUBLIC	4357	EASTERN EUROPE	CHRISTIAN: CATHOLIC	Czech	CZC	Indo-European;Slavic;West;Czech-Slovak	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU837	EUROPEAN	CZECH & SLOVAKIAN	SLOVAKIA	524	EASTERN EUROPE	CHRISTIAN	Slovak	SLO	Indo-European;Slavic;West;Czech-Slovak	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU212	EUROPEAN	DUTCH	BELGIUM (FLEMISH)	4417	CENTRAL EUROPE	CHRISTIAN: PROTESTANT	Vlaams	VLS	Indo-European;Germanic;West;Low Saxon-Low Franconian	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
EU214	EUROPEAN	DUTCH	NETHERLANDS	20495	CENTRAL EUROPE	CHRISTIAN	Dutch	DUT	Indo-European;Germanic;West;Low Saxon-Low Franconian	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
EU523	EUROPEAN	ENGLISH	MALTA	8027	SOUTHERN EUROPE	CHRISTIAN: CATHOLIC	Maltese	MLS	Afro-Asiatic;Semitic;Central;South	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU110	EUROPEAN	EUROPEAN OTHER	EUROPEAN	31341	CENTRAL EUROPE	CHRISTIAN	German	GER	Indo-European;Germanic;West;High German	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
EU211	EUROPEAN	FRENCH	BELGIUM	815	CENTRAL EUROPE	CHRISTIAN	French	FRN	Indo-European;Italic;Romance;Italo-Western	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
EU213	EUROPEAN	FRENCH	BELGIUM (WALLOON)	618	CENTRAL EUROPE	CHRISTIAN: CATHOLIC	French	WLZ	Indo-European;Italic;Romance;Italo-Western	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
EU317	EUROPEAN	FRENCH	BRETON	640	CENTRAL EUROPE	CHRISTIAN: CATHOLIC	French	FRN	Indo-European;Italic;Romance;Italo-Western	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
EU318	EUROPEAN	FRENCH	CANADA	299	AMERICAS	CHRISTIAN: PROTESTANT	English	ENG	Indo-European;Germanic;West;English	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU316	EUROPEAN	FRENCH	FRANCE	125754	CENTRAL EUROPE	CHRISTIAN: CATHOLIC	French	FRN	Indo-European;Italic;Romance;Italo-Western	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
EU319	EUROPEAN	FRENCH	FRENCH CARIBBEAN	3	AMERICAS	CHRISTIAN: CATHOLIC	French	FRN	Indo-European;Italic;Romance;Italo-Western	1- Black - Caribbean	M) Black or Black British - Caribbean	CHRISTIAN	COB_Elsewhere
EU420	EUROPEAN	GERMAN	GERMANY	129190	CENTRAL EUROPE	CHRISTIAN	German	GER	Indo-European;Germanic;West;High German	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
EU421	EUROPEAN	GERMAN	SWITZERLAND	128	CENTRAL EUROPE	CHRISTIAN	SchwyzerdŸtsch	GSW	Indo-European;Germanic;West;High German	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU838	EUROPEAN	HUNGARIAN	HUNGARY	11768	EASTERN EUROPE	CHRISTIAN: CATHOLIC	Hungarian	HNG	Uralic;Finno-Ugric;Ugric;Hungarian	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU522	EUROPEAN	ITALIAN	ITALY	229931	SOUTHERN EUROPE	CHRISTIAN: CATHOLIC	Italian	ITN	Indo-European;Italic;Romance;Italo-Western	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
EU729	EUROPEAN	MUSLIM	BOSNIA AND HERZEGOVINA	1034	EASTERN EUROPE	MUSLIM	Bosnian	BWF	Indo-European;Slavic;South;Western	0- White	C) White - Any other White background	MUSLIM	COB_Elsewhere
EU835	EUROPEAN	POLISH	POLAND	155743	EASTERN EUROPE	CHRISTIAN: CATHOLIC	Polish	PQL	Indo-European;Slavic;West;Lechitic	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU840	EUROPEAN	ROMANIAN	ROMANIA	744	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Romanian	RUM	Indo-European;Italic;Romance;Eastern	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU841	EUROPEAN	ROMANIAN	ROMANIA BANAT	29	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Romanian	RUM	Indo-European;Italic;Romance;Eastern	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU842	EUROPEAN	ROMANIAN	ROMANIA DOBREGA	28	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Romanian	RUM	Indo-European;Italic;Romance;Eastern	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU843	EUROPEAN	ROMANIAN	ROMANIA MANAMURESCRIANA	331	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Romanian	RUM	Indo-European;Italic;Romance;Eastern	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere

CEL Type Code	CEL Group	CEL Subgroup	CEL Type	People GB 2004	Geographical Area	Religion	Major Language	Major Language SIL Code	Major Language Family Tree	1991 Census Ethnic Group	2001 Census Ethnic Group	2001 Census Religion	2001 Census COB
EU844	EUROPEAN	ROMANIAN	ROMANIA MOLDOVA	200	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Romanian	RUM	Indo-European;Italic;Romance;Eastern	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU845	EUROPEAN	ROMANIAN	ROMANIA MUNTENIA	364	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Romanian	RUM	Indo-European;Italic;Romance;Eastern	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU846	EUROPEAN	ROMANIAN	ROMANIA TRANSILVANIA	835	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Romanian	RUM	Indo-European;Italic;Romance;Eastern	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU950	EUROPEAN	RUSSIAN	AZERBAIJAN	12	CENTRAL ASIA	MUSLIM	Azerbaijani, North	AZE	Altaic;Turkic;Southern;Azerbaijani	0- White	C) White - Any other White background	MUSLIM	COB_Elsewhere
EU948	EUROPEAN	RUSSIAN	BELARUS	27	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Belarusan	RUW	Indo-European;Slavic;East;	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU951	EUROPEAN	RUSSIAN	GEORGIA	185	CENTRAL ASIA	CHRISTIAN: RUSSIAN ORTHODOX	Georgian	GEO	South Caucasian;Georgian	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU947	EUROPEAN	RUSSIAN	RUSSIA	11118	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Russian	RUS	Indo-European;Slavic;East;	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
EU949	EUROPEAN	UKRANIAN	UKRAINE	3948	EASTERN EUROPE	CHRISTIAN: RUSSIAN ORTHODOX	Ukrainian	UKR	Indo-European;Slavic;East;	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
GR110	GREEK	GREEK	GREECE	29134	SOUTHERN EUROPE	CHRISTIAN: GREEK ORTHODOX	Greek	GRK	Indo-European;Greek;Attic;	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
GR211	GREEK	GREEK	GREEK CYPRUS	79304	SOUTHERN EUROPE	CHRISTIAN: GREEK ORTHODOX	Greek	GRK	Indo-European;Greek;Attic;	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
GR212	GREEK	GREEK	GREEK ORTHODOX	932	SOUTHERN EUROPE	CHRISTIAN: GREEK ORTHODOX	Greek	GRK	Indo-European;Greek;Attic;	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
HI213	HISPANIC	PORTUGUESE	ANGOLA	458	AFRICA	CHRISTIAN: CATHOLIC	Portuguese	POR	Indo-European;Italic;Romance;Italo-Western	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
HI212	HISPANIC	PORTUGUESE	BRAZIL	1949	AMERICAS	CHRISTIAN: CATHOLIC	Portuguese	POR	Indo-European;Italic;Romance;Italo-Western	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	CHRISTIAN	COB_Elsewhere
HI214	HISPANIC	PORTUGUESE	GOA	990	SOUTH ASIA	CHRISTIAN: CATHOLIC	Portuguese	POR	Indo-European;Italic;Romance;Italo-Western	4- Indian	H) Asian or Asian British - Indian	CHRISTIAN	COB_Elsewhere
HI211	HISPANIC	PORTUGUESE	PORTUGAL	86930	SOUTHERN EUROPE	CHRISTIAN: CATHOLIC	Portuguese	POR	Indo-European;Italic;Romance;Italo-Western	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
HI419	HISPANIC	SPANISH	BASQUE	1568	SOUTHERN EUROPE	CHRISTIAN: CATHOLIC	Basque	BSQ	Basque;	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
HI316	HISPANIC	SPANISH	CASTILLIAN	10775	SOUTHERN EUROPE	CHRISTIAN: CATHOLIC	Spanish	SPN	Indo-European;Italic;Romance;Italo-Western	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
HI520	HISPANIC	SPANISH	CATALAN	3105	SOUTHERN EUROPE	CHRISTIAN: CATHOLIC	Catalan	CLN	Indo-European;Italic;Romance;Italo-Western	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
HI621	HISPANIC	SPANISH	GALICIAN	511	SOUTHERN EUROPE	CHRISTIAN: CATHOLIC	Galician	GLN	Indo-European;Italic;Romance;Italo-Western	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
HI110	HISPANIC	SPANISH	HISPANIC	6084	SOUTHERN EUROPE	CHRISTIAN: CATHOLIC	Spanish	SPN	Indo-European;Italic;Romance;Italo-Western	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
HI317	HISPANIC	SPANISH	LATIN AMERICA	3644	AMERICAS	CHRISTIAN: CATHOLIC	Spanish	SPN	Indo-European;Italic;Romance;Italo-Western	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	CHRISTIAN	COB_Elsewhere
HI318	HISPANIC	SPANISH	PHILIPPINES	1976	EAST ASIA	CHRISTIAN: CATHOLIC	Filipino	FIL	Austronesian;Malayo-Polynesian;Meso Philippine;Central Philippine	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	CHRISTIAN	COB_Elsewhere
HI315	HISPANIC	SPANISH	SPAIN	80180	SOUTHERN EUROPE	CHRISTIAN: CATHOLIC	Spanish	SPN	Indo-European;Italic;Romance;Italo-Western	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
IN110	INTERNATIONAL	INTERNATIONAL	INTERNATIONAL	15799	UNCLASSIFIED	Not Applicable	Not Applicable	Not Applicable	Not Applicable	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	NOT APPLICABLE	COB_Elsewhere
JP110	JAPANESE	JAPANESE	JAPAN	6335	EAST ASIA	BHUDDIST	Japanese	JPN	Japanese;Japanese	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	BHUDDIST	COB_Elsewhere
JA313	JEWISH AND ARMENIAN	ARMENIAN	ARMENIAN	4353	CENTRAL ASIA	CHRISTIAN: ORTHODOX_CALCEDONIAN	Armenian	ARM	Indo-European;Armenian	0- White	C) White - Any other White background	CHRISTIAN	COB_Elsewhere
JA211	JEWISH AND ARMENIAN	JEWISH	JEWISH	80522	DIASPORIC	JEWISH	Hebrew	HBR	Afro-Asiatic;Semitic;Central;South	0- White	C) White - Any other White background	JEWISH	COB_Elsewhere
JA110	JEWISH AND ARMENIAN	JEWISH	JEWISH AND ARMENIAN	72	DIASPORIC	Not Applicable	Not Applicable	Not Applicable	Not Applicable	0- White	C) White - Any other White background	NOT APPLICABLE	COB_Elsewhere
JA212	JEWISH AND ARMENIAN	JEWISH	SEPHARDIC JEWISH	821	DIASPORIC	JEWISH	Ladino	LAD	Indo-European;Italic;Romance;Italo-Western	0- White	C) White - Any other White background	JEWISH	COB_Elsewhere
ML427	MUSLIM	BANGLADESHI	BANGLADESH	179401	SOUTH ASIA	MUSLIM	Bengali	BNG	Indo-European;Indo-Iranian;Indo-Aryan;Eastern zone	6- Bangladeshi	K) Asian or Asian British - Bangladeshi	MUSLIM	COB_Elsewhere
ML640	MUSLIM	ERITREAN	ERITREA	1397	AFRICA	CHRISTIAN: OTHER	TigrŽ	TIE	Afro-Asiatic;Semitic;South;Ethiopian	2- Black - African	N) Black or Black British - African	CHRISTIAN	COB_Elsewhere
ML212	MUSLIM	IRANIAN	IRAN	10312	MIDDLE EAST	MUSLIM	Farsi	PES	Indo-European;Indo-Iranian;Iranian;Western	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML216	MUSLIM	LEBANESE	LEBANON	3107	MIDDLE EAST	MUSLIM	Arabic	ARB	Afro-Asiatic;Semitic;Central;South	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML743	MUSLIM	MUSLIM	BALKAN MUSLIM	10	EASTERN EUROPE	MUSLIM	Bosnian	BWF	Indo-European;Slavic;South;Western	0- White	C) White - Any other White background	MUSLIM	COB_Elsewhere
ML431	MUSLIM	MUSLIM	MALAYSIAN MUSLIM	220	EAST ASIA	MUSLIM	Malay	MLI	Austronesian;Malayo-Polynesian;Western Malayo-Polynesian;Sundic	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML639	MUSLIM	MUSLIM	SUDAN	468	AFRICA	MUSLIM	Arabic	ARB	Afro-Asiatic;Semitic;Central;South	2- Black - African	N) Black or Black British - African	MUSLIM	COB_Elsewhere
ML637	MUSLIM	MUSLIM	WEST AFRICAN MUSLIM	2399	AFRICA	MUSLIM	Arabic	ARB	Afro-Asiatic;Semitic;Central;South	2- Black - African	N) Black or Black British - African	MUSLIM	COB_Elsewhere
ML213	MUSLIM	MUSLIM MIDDLE EAST	IRAQ	262	MIDDLE EAST	MUSLIM	Arabic	ARB	Afro-Asiatic;Semitic;Central;South	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML214	MUSLIM	MUSLIM MIDDLE EAST	JORDAN	55	MIDDLE EAST	MUSLIM	Arabic	ARB	Afro-Asiatic;Semitic;Central;South	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML215	MUSLIM	MUSLIM MIDDLE EAST	KUWAIT	3	MIDDLE EAST	MUSLIM	Arabic	ARB	Afro-Asiatic;Semitic;Central;South	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML211	MUSLIM	MUSLIM MIDDLE EAST	MIDDLE EAST	672	MIDDLE EAST	MUSLIM	Arabic	ARB	Afro-Asiatic;Semitic;Central;South	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML110	MUSLIM	MUSLIM MIDDLE EAST	MUSLIM	103514	MIDDLE EAST	MUSLIM	Arabic	ARB	Afro-Asiatic;Semitic;Central;South	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML217	MUSLIM	MUSLIM MIDDLE EAST	OMAN	5	MIDDLE EAST	MUSLIM	Arabic	ARB	Afro-Asiatic;Semitic;Central;South	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML218	MUSLIM	MUSLIM MIDDLE EAST	SAUDI ARABIA	186	MIDDLE EAST	MUSLIM	Arabic	ARB	Afro-Asiatic;Semitic;Central;South	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML219	MUSLIM	MUSLIM MIDDLE EAST	SYRIA	142	MIDDLE EAST	MUSLIM	Arabic	ARB	Afro-Asiatic;Semitic;Central;South	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML220	MUSLIM	MUSLIM MIDDLE EAST	UNITED ARAB EMIRATES	14	MIDDLE EAST	MUSLIM	Arabic	ARB	Afro-Asiatic;Semitic;Central;South	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere

CEL Type Code	CEL Group	CEL Subgroup	CEL Type	People GB 2004	Geographical Area	Religion	Major Language	Major Language SIL Code	Major Language Family Tree	1991 Census Ethnic Group	2001 Census Ethnic Group	2001 Census Religion	2001 Census COB
ML221	MUSLIM	MUSLIM MIDDLE EAST	YEMEN	6	MIDDLE EAST	MUSLIM	Arabic	ARB	Afro-Asiatic;Semitic;Central;South	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML532	MUSLIM	MUSLIM NORTHAFRICAN	ALGERIA	2585	AFRICA	MUSLIM	Arabic	ARB	Afro-Asiatic;Semitic;Central;South	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML533	MUSLIM	MUSLIM NORTHAFRICAN	EGYPT	479	AFRICA	MUSLIM	Arabic	ARB	Afro-Asiatic;Semitic;Central;South	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML535	MUSLIM	MUSLIM NORTHAFRICAN	LIBYA	38	AFRICA	MUSLIM	Arabic	ARB	Afro-Asiatic;Semitic;Central;South	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML536	MUSLIM	MUSLIM NORTHAFRICAN	MOROCCO	572	AFRICA	MUSLIM	Arabic	ARB	Afro-Asiatic;Semitic;Central;South	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML534	MUSLIM	MUSLIM NORTHAFRICAN	TUNISIA	39	AFRICA	MUSLIM	Arabic	ARB	Afro-Asiatic;Semitic;Central;South	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML428	MUSLIM	MUSLIM SOUTH ASIAN	MUSLIM INDIA	25704	SOUTH ASIA	MUSLIM	Punjabi	PNJ	Indo-European;Indo-Iranian;Indo-Aryan;Central zone	4- Indian	H) Asian or Asian British - Indian	MUSLIM	COB_Elsewhere
ML326	MUSLIM	MUSLIM STANS	AFGHANISTAN	3687	CENTRAL ASIA	MUSLIM	Farsi	PRS	Indo-European;Indo-Iranian;Iranian;Western	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML322	MUSLIM	MUSLIM STANS	KAZAKHSTAN	11	CENTRAL ASIA	MUSLIM	Kazakh	KAZ	Altaic;Turkic;Western;Aralo-Caspian	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML323	MUSLIM	MUSLIM STANS	KYRGYZSTAN	2	CENTRAL ASIA	MUSLIM	Kirghiz	KDO	Altaic;Turkic;Western;Aralo-Caspian	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML324	MUSLIM	MUSLIM STANS	TURKMENISTAN	8	CENTRAL ASIA	MUSLIM	Turkmen	TUR	Afro-Asiatic;Chadic;Biu-Mandara;A	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML325	MUSLIM	MUSLIM STANS	UZBEKISTAN	3	CENTRAL ASIA	MUSLIM	Uzbek	UZB	Altaic;Turkic;Eastern;	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML429	MUSLIM	PAKISTANI	PAKISTAN	508699	SOUTH ASIA	MUSLIM	Punjabi	PNB	Indo-European;Indo-Iranian;Indo-Aryan;Northwestern zone	5- Pakistani	J) Asian or Asian British - Pakistani	MUSLIM	COB_Elsewhere
ML430	MUSLIM	PAKISTANI KASHMIR	PAKISTANI KASHMIR	91472	SOUTH ASIA	MUSLIM	Kashmiri	KSH	Indo-European;Indo-Iranian;Indo-Aryan;Northwestern zone	5- Pakistani	J) Asian or Asian British - Pakistani	MUSLIM	COB_Elsewhere
ML638	MUSLIM	SOMALIAN	SOMALIA	33260	AFRICA	MUSLIM	Somali	SOM	Afro-Asiatic;Cushitic;East;Somali	2- Black - African	N) Black or Black British - African	MUSLIM	COB_Elsewhere
ML741	MUSLIM	TURKISH	TURKEY	50706	MIDDLE EAST	MUSLIM	Turkish	TRK	Altaic;Turkic;Southern;Turkish	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ML742	MUSLIM	TURKISH	TURKISH CYPRUS	1205	MIDDLE EAST	MUSLIM	Turkish	TRK	Altaic;Turkic;Southern;Turkish	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	MUSLIM	COB_Elsewhere
ND211	NORDIC	DANISH	DENMARK	20561	NORTHERN EUROPE	CHRISTIAN: PROTESTANT	Danish	DNS	Indo-European;Germanic;North;East Scandinavian	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
ND315	NORDIC	FINNISH	FINLAND	5685	NORTHERN EUROPE	CHRISTIAN: PROTESTANT	Finnish	FIN	Uralic;Finnic	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
ND212	NORDIC	NORDIC	ICELAND	115	NORTHERN EUROPE	CHRISTIAN: PROTESTANT	Icelandic	ICE	Indo-European;Germanic;North;West Scandinavian	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
ND110	NORDIC	NORDIC	NORDIC	6377	NORTHERN EUROPE	CHRISTIAN: PROTESTANT	Not Applicable	Not Applicable	Not Applicable	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
ND214	NORDIC	NORWEGIAN	NORWAY	186375	NORTHERN EUROPE	CHRISTIAN: PROTESTANT	Norwegian	NNO	Indo-European;Germanic;North;West Scandinavian	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
ND213	NORDIC	SWEDISH	SWEDEN	19090	NORTHERN EUROPE	CHRISTIAN: PROTESTANT	Swedish	SWD	Indo-European;Germanic;North;East Scandinavian	0- White	C) White - Any other White background	CHRISTIAN	COB_Other EU countries
SK110	SIKH	SIKH	INDIA SIKH	283657	SOUTH ASIA	SIKH	Punjabi	PNJ	Indo-European;Indo-Iranian;Indo-Aryan;Central zone	4- Indian	H) Asian or Asian British - Indian	SIKH	COB_Elsewhere
SA211	SOUTH ASIAN	HINDI INDIAN	INDIA HINDI	319677	SOUTH ASIA	HINDU	Hindi	HND	Indo-European;Indo-Iranian;Indo-Aryan;Central zone	4- Indian	H) Asian or Asian British - Indian	HINDU	COB_Elsewhere
SA213	SOUTH ASIAN	HINDI INDIAN	INDIA SOUTH	302	SOUTH ASIA	BHUDDIST	Hindi	HND	Indo-European;Indo-Iranian;Indo-Aryan;Central zone	4- Indian	H) Asian or Asian British - Indian	BHUDDIST	COB_Elsewhere
SA316	SOUTH ASIAN	HINDI NOT INDIAN	BANGLADESH HINDU	2974	SOUTH ASIA	HINDU	Bengali	BNG	Indo-European;Indo-Iranian;Indo-Aryan;Eastern zone	6- Bangladeshi	K) Asian or Asian British - Bangladeshi	HINDU	COB_Elsewhere
SA214	SOUTH ASIAN	HINDI NOT INDIAN	HINDU NOT INDIAN	22106	SOUTH ASIA	HINDU	Hindi	HND	Indo-European;Indo-Iranian;Indo-Aryan;Central zone	4- Indian	H) Asian or Asian British - Indian	HINDU	COB_Elsewhere
SA212	SOUTH ASIAN	INDIA NORTH	INDIA NORTH	75282	SOUTH ASIA	HINDU	Hindi	HND	Indo-European;Indo-Iranian;Indo-Aryan;Central zone	4- Indian	H) Asian or Asian British - Indian	HINDU	COB_Elsewhere
SA522	SOUTH ASIAN	SOUTH ASIAN OTHER	ASIAN CARIBBEAN	581	AMERICAS	HINDU	Hindi	HND	Indo-European;Indo-Iranian;Indo-Aryan;Central zone	4- Indian	L) Asian or Asian British - Any other Asian background	HINDU	COB_Elsewhere
SA317	SOUTH ASIAN	SOUTH ASIAN OTHER	BHUTAN	3	SOUTH ASIA	BHUDDIST	Dzongkha	DZO	Sino-Tibetan;Tibeto-Burman;Himalayish;Tibeto-Kanauri	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	BHUDDIST	COB_Elsewhere
SA523	SOUTH ASIAN	SOUTH ASIAN OTHER	GUYANA	911	AMERICAS	HINDU	Hindi	HND	Indo-European;Indo-Iranian;Indo-Aryan;Central zone	4- Indian	L) Asian or Asian British - Any other Asian background	HINDU	COB_Elsewhere
SA421	SOUTH ASIAN	SOUTH ASIAN OTHER	KENYAN ASIAN	1121	AFRICA	HINDU	Hindi	HND	Indo-European;Indo-Iranian;Indo-Aryan;Central zone	4- Indian	L) Asian or Asian British - Any other Asian background	HINDU	COB_Elsewhere
SA318	SOUTH ASIAN	SOUTH ASIAN OTHER	NEPAL	150	SOUTH ASIA	HINDU	Nepali	NEP	Indo-European;Indo-Iranian;Indo-Aryan;Northern zone	8- Any other ethnic group	L) Asian or Asian British - Any other Asian background	HINDU	COB_Elsewhere
SA420	SOUTH ASIAN	SOUTH ASIAN OTHER	SEYCHELLES	71	SOUTH ASIA	CHRISTIAN: CATHOLIC	Seselwa Creole French	CRS	Creole;French based	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	CHRISTIAN	COB_Elsewhere
SA110	SOUTH ASIAN	SOUTH ASIAN OTHER	SOUTH ASIA	12699	SOUTH ASIA	BHUDDIST	Hindi	HND	Indo-European;Indo-Iranian;Indo-Aryan;Central zone	8- Any other ethnic group	S) Other Ethnic Groups - Any other ethnic group	BHUDDIST	COB_Elsewhere
SA315	SOUTH ASIAN	SRI LANKAN	SRI LANKA	53919	SOUTH ASIA	BHUDDIST	Sinhala	SNH	Indo-European;Indo-Iranian;Indo-Aryan;Sinhalese-Maldivian	4- Indian	L) Asian or Asian British - Any other Asian background	BHUDDIST	COB_Elsewhere
ZU110	UNCLASSIFIED	VOID	UNCLASSIFIED	21826	UNCLASSIFIED	Not Applicable	Not Applicable	Not Applicable	Not Applicable	9- Unknown	Y) Unclassified	NOT APPLICABLE	COB_Elsewhere
ZZ316	VOID	VOID	NOT FOUND	110049		Not Applicable	Not Applicable	Not Applicable	Not Applicable	9- Unknown	Y) Unclassified	NOT APPLICABLE	COB_Elsewhere
ZZ110	VOID	VOID	VOID	12118	UNCLASSIFIED	Not Applicable	Not Applicable	Not Applicable	Not Applicable	9- Unknown	Y) Unclassified	NOT APPLICABLE	COB_Elsewhere
ZZ211	VOID	VOID	VOID SURNAME	819	UNCLASSIFIED	Not Applicable	Not Applicable	Not Applicable	Not Applicable	9- Unknown	Y) Unclassified	NOT APPLICABLE	COB_Elsewhere
ZZ212	VOID	VOID	VOID INITIAL	94621	UNCLASSIFIED	Not Applicable	Not Applicable	Not Applicable	Not Applicable	9- Unknown	Y) Unclassified	NOT APPLICABLE	COB_Elsewhere
ZZ213	VOID	VOID	VOID OTHER	56	UNCLASSIFIED	Not Applicable	Not Applicable	Not Applicable	Not Applicable	9- Unknown	Y) Unclassified	NOT APPLICABLE	COB_Elsewhere
ZZ214	VOID	VOID	VOID PERSONAL NAME	5464	UNCLASSIFIED	Not Applicable	Not Applicable	Not Applicable	Not Applicable	9- Unknown	Y) Unclassified	NOT APPLICABLE	COB_Elsewhere
ZZ215	VOID	VOID	VOID TITLE	1858	UNCLASSIFIED	Not Applicable	Not Applicable	Not Applicable	Not Applicable	9- Unknown	Y) Unclassified	NOT APPLICABLE	COB_Elsewhere

Appendix 4: Automated Classification Algorithms

Structured Query Language (SQL) queries developed to build the automated classification of names.

See chapter 6 for detailed explanation of the functionality of each of these queries that relate to the steps described in the chapter

SQL colour and fonts used here follow the convention used by the editor PL/SQL Developer v.7 for Oracle 10.

- Comments to the SQL code are preceded by '`--` ' and appear in *red* and *italics*.
- SQL statements appear in **black** and highlighted in **bold**
- Table names, field names or any other variable name appear in normal black not highlighted
- User imputed values appear in blue
- Individual queries are separated by a black line

---

1) Sample dataset of Non British forenames is created

```
-- Create a table of Non British Forenames freq >10 and freq<4000

create table tbl_cel_forename_sample as

--select count(*) from
(
select t.id_nr,
       t.forename,
       t.cel,
       lu.cel_subgroup,
       lu.cel_group,
       fq.freq,
       fq.gender
  from tbl_cel_forename t
  left join tbl_cel_lookup lu on t.cel = lu.cel_type
  left join tbl_uk_er_firstname_freq2 fq on t.forename = fq.firstname
 where t.cel_group <> 'ENGLISH'
   and t.cel_group <> 'CELTIC'
   and t.cel_group <> 'IRISH'
   and fq.freq > 10
)
```

---

2) Query to create a Surname-Forename Matrix for Non-British Forenames

```
-- This query creates a new table of surname-firstname frequencies.
-- It is based on the 2004 Electoral Roll individual records
-- It is restricted to consider only those records with Forenames considered of
```

```
a'Non-British' CELs
-- Non-British FCEL is reduced to forenames with frequency between 10 and 4000,
-- contained in table tbl_cel_forename_sample


create table tbl_uk_er_aa_sur_fore_matrix as

select er.surname, er.firstname, count(distinct(er.id)) freq
from tbl_cel_forename_sample f, tbl_uk_elec_roll er
where er.firstname = f.forename
group by er.surname, er.firstname
order by er.surname, er.firstname
```

3) Calculations to cleanse the previous table

```
3.1) Remove 'VOID' Surnames in Master CEL (Personal Names, Invalid Entries, etc.

delete from tbl_uk_er_aa_sur_fore_matrix t

 where exists (select *
          from tbl_cel_surname cel
         where t.surname = cel.surname
           and cel.cel_group = 'VOID')
--8,500 rows removed
```

```
3.2) Remove Surnames with numeric characters
delete from tbl_uk_er_aa_sur_fore_matrix t
where REGEXP_LIKE(t.surname, '[[:digit:]]')
--55 rows deleted
```

```
3.3) Remove Surnames with strange characters

delete from tbl_uk_er_aa_sur_fore_matrix t
where REGEXP_LIKE(t.surname, '&')
--9 rows deleted

select * from tbl_uk_er_aa_sur_fore_matrix t
where REGEXP_LIKE(t.surname, '''')
--878 rows deleted

delete from tbl_uk_er_aa_sur_fore_matrix t
where REGEXP_LIKE(t.surname, '#')
--1 row deleted
```

4) Operations to filter the tbl_uk_er_aa_sur_fore_matrix (MT) table

```
4.1) Remove surnames in MT with national UK frequency of 1 or 2

delete from tbl_uk_er_aa_sur_fore_matrix mt
 where exists (select *
          from tbl_uk_er_surname_freq fq
         where mt.surname = fq.surname
           and fq.uk_frequency < 3)

179136 rows deleted
```

```
4.2) Delete Surnames and Forenames with short Names <4 characters

delete from tbl_uk_er_aa_sur_fore_matrix t
where LENGTH (t.firstname) <4

delete from tbl_uk_er_aa_sur_fore_matrix t
where LENGTH (t.surname) <4
```

```
4.3) Remove surnames in MT with Forename types per Surname = 1 AND Surname-Forename
Freq = 1

delete from tbl_uk_er_aa_sur_fore_matrix mt
where exists
(
select * from tbl_uk_er_aa_sur_forefreq ff
where mt.surname = ff.surname
and ff.fore_typ = 1
and ff.surfore_freq = 1
)
```

5) Selection of the top CEL subgroup for each forename

```
--This query assigns a CEL SUBGROUP to each FORENAME in the sample
-- selecting the Top CEL SUBGROUP other than English of their surnames in the Total
ER
create or replace view vw_uk_er_full_fore_subgrp as

select distinct oq.forename, oq.cel_subgroup, oq.sbgr_sur_tok, oq.sbgr_sur_typ,
oq.perc_subgrp
  from -- This first query reports 1 row per FORENAME-SUBGROUP pair,
       --calculating the share of each SUBGROUP in the forename

        (select distinct fs.firstname forename,
                         lu.cel_subgroup,
                         sum(mt.pair_tok) sbgr_sur_tok,
                         count(mt.surname) sbgr_sur_typ,
                         (sum(mt.pair_tok) / min(fs.surfore_freq)) perc_subgrp
           from tbl_uk_er_aa_for_surfreq fs
           left join tbl_uk_er_full_fore_sur_matrix mt on fs.firstname =
                                                          mt.firstname
           left join tbl_cel_surname st on mt.surname = st.surname
           left join tbl_cel_lookup lu on lu.cel_type = st.cel
          where lu.cel_type is not null
            and lu.cel_subgroup <> 'ENGLISH'
          group by fs.firstname, lu.cel_subgroup) oq,

       -- This second nested queries report the max FORENAME-SUBGROUP pair,
       --based on the MAX the share of each SUBGROUP in the forename

       (select distinct ir.forename, max(ir.perc_subgrp) max_sbgrp
          from (select distinct fs.firstname forename,
                                lu.cel_subgroup,
                                sum(mt.pair_tok) sbgr_sur_tok,
                                count(mt.surname) sbgr_sur_typ,
                                (sum(mt.pair_tok) / min(fs.surfore_freq)) perc_subgrp
                  from tbl_uk_er_aa_for_surfreq fs
                  left join tbl_uk_er_full_fore_sur_matrix mt on fs.firstname =
                                                                 mt.firstname
                  left join tbl_cel_surname st on mt.surname = st.surname
                  left join tbl_cel_lookup lu on lu.cel_type = st.cel
                 where lu.cel_type is not null
                   and lu.cel_subgroup <> 'ENGLISH'
                 group by fs.firstname, lu.cel_subgroup) ir
         group by ir.forename) mx

 -- This links the two queries

 where oq.forename = mx.forename
   and oq.perc_subgrp = mx.max_sbgrp

--But it creates a problem of forenames with several CEL Subgroups
--Which gets removed by the following query
```

```
-- Creates a table with a unique Subgroup per FORENAME
create table tbl_uk_er_aa_fore_subgrp_freq as
(
select distinct * from
-- The view that creates the MAX Subgroups
```

```
vw_uk_er_full_fore_subgrp vw
        -- This query makes sure that only forenames with ONE UNIQUE MAX Subgroup are
used
where exists (select * from
              (
              select tf.forename, count(tf.cel_subgroup) cnt
               from vw_uk_er_full_fore_subgrp tf
               group by tf.forename
               ) q
               where cnt =1
               and  vw.forename= q.forename)
)
```

```
-- This query joins vw_uk_er_aa_fore_subgrp_diag with tbl_uk_er_aa_fore_subgrp_freq
and creates a new summary table


create table tbl_uk_er_aa_fore_subgrp_all as
(
select sa.forename, sa.samp_for_tok, sa.tot_fore_tok, sa.subgrp_fore_tok,
sa.subgrp_sur_typ, sa.samp_sur_typ, sa.tot_sur_typ,
sa.sur_cel_subgroup sam_subgroup,
tt.cel_subgroup tot_subgroup, tt.sbgr_sur_tok, tt.sbgr_sur_typ
from vw_uk_er_aa_fore_subgrp_diag sa
left join tbl_uk_er_aa_fore_subgrp_freq tt on tt.forename = sa.forename
)
```

---

6) Standardisation of percentages by z-scores, averaging z_typ and z_tok and truncating them

```
-- This set of queries calculates the final SCORE for each FORENAME and its
association with a SUBGROUP
create table tbl_uk_er_aa_fore_z_scores as
(
-- This query averages the Z-SCORES and adds a value of one to create the final score
for a FORENAME
-- at the end there is a filter to select only positive values of the final_score
(>0)
select z.*, (((z.z_tok+z.z_typ)/2)+1) final_score
from
(
-- This query links two subqueries of stats and calculates Z-SCORES
select pr.*,
       case
         when st.std_tok = 0 then
          0
         else
          ((pr.subgrp_tok_perc - st.avg_tok) / st.std_tok)
       end z_tok,
       case
         when st.std_typ = 0 then
          0
         else
          ((pr.subgrp_typ_perc - st.avg_typ) / st.std_typ)
       end z_typ
from
-- This query calculates percentages of tok and typ by FORENAME
--based on tbl_uk_er_aa_fore_subgrp_all table
(
select t.forename,
       t.tot_subgroup,
       (t.sbgr_sur_tok / t.tot_fore_tok) subgrp_tok_perc,
       (t.sbgr_sur_typ / t.tot_sur_typ) subgrp_typ_perc
  from tbl_uk_er_aa_fore_subgrp_all t
 where t.tot_subgroup is not null
) pr,

-- This query calculates the average and stddev of the above query
```

```
(select q.tot_subgroup,
       avg(q.subgrp_tok_perc) avg_tok,
       stddev(q.subgrp_tok_perc) std_tok,
       avg(q.subgrp_typ_perc) avg_typ,
       stddev(q.subgrp_typ_perc) std_typ
  from -- This is the same query as above
       (select t.forename,
               t.tot_subgroup,
               (t.sbgr_sur_tok / t.tot_fore_tok) subgrp_tok_perc,
               (t.sbgr_sur_typ / t.tot_sur_typ) subgrp_typ_perc
          from tbl_uk_er_aa_fore_subgrp_all t
         where t.tot_subgroup is not null) q
 group by q.tot_subgroup
 ) st
 --This joins the two subqueries
 where pr.tot_subgroup = st.tot_subgroup
) z
-- This is a filter to select only positive values of the final_score (>0)
where (((z.z_tok+z.z_typ)/2)+1)> 0
)
```

7) Cycle 1: Forename-Surname-Clustering (FSC)

```
This query reports all the possible CEL subgroups per surname.
It links the main Forename-Surname Matrix (tbl_uk_er_full_fore_sur_matrix) with the
Forename seed list (tbl_uk_er_aa_fore_z_scores) and the gender weighting table (
tbl_uk_er_gender_weight)

create table tbl_uk_er_sur_subgrp_score
as
(
select distinct surname,
                tot_subgroup,
                sum(pair_tok) fore_tok,
                count(pair_tok) fore_typ,
                sum(fs_pair_score) subgrp_cumm_score

  from (select distinct mt.surname,
                        zs.tot_subgroup,
                        zs.final_score z_f_score,
                        sx.gender,
                        mt.pair_tok,
                       (gd.weight* zs.final_score * mt.pair_tok) fs_pair_score
          from tbl_uk_er_full_fore_sur_matrix mt
          left join tbl_uk_er_aa_fore_z_scores zs on mt.firstname =
                                                     zs.forename
          left join tbl_uk_er_firstname_freq2 sx on mt.firstname =
                                                    sx.firstname
         left join tbl_uk_er_gender_weight gd on sx.gender = gd.gender
         )
group by surname, tot_subgroup
)
```

```
--This is the second query that selects the SUBGROUP with MAX value of subgroup SCORE
from the previous table (tbl_uk_er_sur_subgrp_score)
--and calculates final SCORE

create table tbl_uk_er_sur_subgrp_lookup as
(
select distinct
       sc.surname,
       sc.tot_subgroup,
       (sc.subgrp_cumm_score/sc.fore_tok) subgrp_score_pp,
       (sc.fore_tok/stok.uk_frequency) subgrp_tok_tot_perc

from
tbl_uk_er_sur_subgrp_score sc,
tbl_uk_er_surname_freq stok,

--this query selects Non-British surnames
```

```
(select distinct heu.surname from
tbl_cel_surname heu, tbl_cel_lookup lu
where heu.cel = lu.cel_type
and lu.cel_subgroup <> 'ENGLISH'
and lu.cel_subgroup <>  'IRISH'
and lu.cel_subgroup <>  'WELSH'
and lu.cel_subgroup <>  'SCOTTISH') nuk
,
-- This query gets the MAX score
(
select sc.surname, max(sc.subgrp_cumm_score) max_score
  from tbl_uk_er_sur_subgrp_score sc
 where sc.tot_subgroup is not null
   and sc.tot_subgroup <> 'ENGLISH'
   and sc.tot_subgroup <> 'IRISH'
   and sc.tot_subgroup <> 'SCOTTISH'
   and sc.tot_subgroup <> 'WELSH'
   group by sc.surname) mx
-- This links the two queries with two tables together
 where sc.surname = mx.surname
   and sc.subgrp_cumm_score = mx.max_score
   and sc.surname = nuk.surname
   and sc.surname = stok.surname
);

   -- Create/Recreate indexes
create index IDX_UK_ER_SUR_SBGR_LU_SBGR on TBL_UK_ER_SUR_SUBGRP_LOOKUP (TOT_SUBGROUP)
  tablespace USERS
  pctfree 10
  initrans 2
  maxtrans 255
  storage
  (
    initial 64K
    minextents 1
    maxextents unlimited
  );
create index IDX_UK_ER_SUR_SBGR_LU_SU on TBL_UK_ER_SUR_SUBGRP_LOOKUP (SURNAME)
  tablespace USERS
  pctfree 10
  initrans 2
  maxtrans 255
  storage
  (
    initial 64K
    minextents 1
    maxextents unlimited
  );
```

```
-- Creates a table summarising stats for the relationship between F_CEL and S_CEL
subgroup pairs
create table tbl_uk_er_full_fore_sur_pair as
select q.f_subgrp,
       q.s_subgrp,
       count(*) pair_typ,
       avg(q.f_score) avg_f_score,
       avg(q.s_score) avg_s_score,
       avg(q.subgrp_tok_tot_perc) avg_tok_perc

  from (select distinct mt.firstname,
                        fs.tot_subgroup f_subgrp,
                        fs.final_score f_score,
                        mt.surname,
                        ss.tot_subgroup s_subgrp,
                        ss.subgrp_score_pp s_score,
                        ss.subgrp_tok_tot_perc
          from tbl_uk_er_full_fore_sur_matrix mt
         inner join tbl_uk_er_aa_fore_z_scores fs on mt.firstname =
                                                     fs.forename
         inner join tbl_uk_er_sur_subgrp_lookup ss on mt.surname =
                                                      ss.surname) q
 group by q.f_subgrp, q.s_subgrp
 order by q.f_subgrp, q.s_subgrp
```

8) Second round of Forename-Surname-Clustering (FSC)

```
-- Creates a table with new FORENAME_SUBGROUP pairs and new score
-- This is the second round of the F_S_C tecnique going from 90k surnames back to all
forenames
create table tbl_uk_er_aa_fore_subgr_dtld as

select q.forename, q.s_subgrp, sum(q.pair_tok) fore_tok , sum(q.pair_tok*q.s_score)
cum_score
from
(
select distinct mt.firstname forename,
                mt.pair_tok,
                fs.tot_subgroup f_subgrp,
                fs.final_score f_score,
                mt.surname,
                ss.tot_subgroup s_subgrp,
                ss.subgrp_score_pp s_score,
                ss.subgrp_tok_tot_perc
  from tbl_uk_er_full_fore_sur_matrix mt
  left join tbl_uk_er_aa_fore_z_scores fs on mt.firstname = fs.forename
 inner join tbl_uk_er_sur_subgrp_lookup ss on mt.surname = ss.surname) q
 group by q.forename, q.s_subgrp
 order by q.forename, q.s_subgrp
;
 -- Create/Recreate indexes
create index IDX_UK_ER_AA_FOR_SBG_DT_FOR on TBL_UK_ER_AA_FORE_SUBGR_DTLD (FORENAME)
  tablespace USERS
  pctfree 10
  initrans 2
  maxtrans 255
  storage
  (
    initial 64K
    minextents 1
    maxextents unlimited
  );
create index IDX_UK_ER_AA_FOR_SBG_DT_SG on TBL_UK_ER_AA_FORE_SUBGR_DTLD (S_SUBGRP)
  tablespace USERS
  pctfree 10
  initrans 2
  maxtrans 255
  storage
  (
    initial 64K
    minextents 1
    maxextents unlimited
  );
```

```
-- This query selects the MAX SCORE and SUBGROUP for each NON-BRIT FORENAME
-- with a match in the previous 90k surname-subgrp file

create table tbl_uk_er_fore_subgrp_lookup as
select qm.*,
       (qm.cum_score / qm.fore_tok) avg_score_pp,
       (qm.fore_tok / fq.freq) sbgrp_tok_tot_perc
  from
  tbl_uk_er_firstname_freq2 fq,
  (select distinct *
          from tbl_uk_er_aa_fore_subgr_dtld dt
         where dt.forename <> ' ') qm, --removes blank forenames
       -- this query selects the record with max score non Brit
       (select d2.forename, max(d2.cum_score) max_score
          from tbl_uk_er_aa_fore_subgr_dtld d2
         where d2.s_subgrp <> 'ENGLISH'
           and d2.s_subgrp <> 'IRISH'
           and d2.s_subgrp <> 'WELSH'
           and d2.s_subgrp <> 'SCOTTISH'
         group by d2.forename) mx,
       --this query selects Non-British forenames only
```

```
       (select distinct heu.forename
          from tbl_cel_forename heu, tbl_cel_lookup lu
         where heu.cel = lu.cel_type
           and lu.cel_subgroup <> 'ENGLISH'
           and lu.cel_subgroup <> 'IRISH'
           and lu.cel_subgroup <> 'WELSH'
           and lu.cel_subgroup <> 'SCOTTISH') nuk
-- Links the table and three queries
 where qm.forename = mx.forename
   and qm.cum_score = mx.max_score
   and qm.forename = nuk.forename
   and qm.forename = fq.firstname
   -- Limits the results by removing weakest assignments
   -- in terms of low scores (>0.8) and low perc over total freq (<0.2)
   and (qm.cum_score / qm.fore_tok)>= 0.8
   and (qm.fore_tok / fq.freq)>= 0.2

--End of queries
```

Appendix 5: Sample of CEL Classified Names in the Automated Approach

Sample of forenames seed list (see section 6.3 for full details)

FORENAME	CEL SUBGROUP	SCORE
ABUDUL	PAKISTANI	1.24
ABUL	BANGLADESHI	0.36
ACACIO	PORTUGUESE	1.80
ACHAL	HINDI_INDIAN	1.78
ACHALA	HINDI_INDIAN	0.79
ACHHAR	SIKH	1.08
ACHILLE	ITALIAN	1.61
ACHILLEAS	GREEK	1.72
ACHILLES	GREEK	2.03
ACHIM	GERMAN	2.03
ACHLA	HINDI_INDIAN	1.26
ADAEZE	NIGERIAN	0.10
ADAIR	SCOTTISH	1.87
ADAKU	NIGERIAN	0.34
ADAL	PAKISTANI	0.84
ADALAT	PAKISTANI	2.52
ADALBERTO	ITALIAN	0.55
ADALET	TURKISH	0.95
ADALGISA	ITALIAN	1.12
ADAMA	SIERRA LEONIAN	0.43
ADAMANTIA	GREEK	1.08
ADAMANTIOS	GREEK	1.04
ADAMINA	SCOTTISH	4.58
ADAMO	ITALIAN	2.31
ADAMOS	GREEK	2.37
ADANNA	NIGERIAN	0.95
ADAOBI	NIGERIAN	0.44
ADAORA	NIGERIAN	1.30
ADARSH	HINDI_INDIAN	0.54
ADBUL	PAKISTANI	0.58
ADDO	GHANAIAN	2.95
ADDOLORATA	ITALIAN	1.46
ADDUL	PAKISTANI	1.09
ADEBAMBO	NIGERIAN	1.61
ADEBANKE	NIGERIAN	2.48
ADEBAYO	NIGERIAN	1.76
ADHAM	MUSLIM_MIDDLE EAST	1.80
ADIBA	PAKISTANI	1.36
ADIJAT	NIGERIAN	1.16
ADIL	PAKISTANI	0.53
ADILA	PAKISTANI	1.35
ADILIA	PORTUGUESE	2.75
ADILSON	PORTUGUESE	0.85
ADINA	SCOTTISH	0.72
ADIO	NIGERIAN	0.34
ADISA	NIGERIAN	0.49
ADITI	HINDI_INDIAN	0.28
ADJEI	GHANAIAN	2.30
ADJOA	GHANAIAN	2.26
ADLIN	SCOTTISH	1.35
ADMAN	WELSH	0.45
ADNAAN	PAKISTANI	2.10
ADNAN	PAKISTANI	0.53
ADOLF	POLISH	0.16
ADOLFO	ITALIAN	0.51
ADOLPHINE	WELSH	1.36
ADONIS	GREEK	0.27
ADORACION	SPANISH	0.63
ADREES	PAKISTANI	2.68
ADRI	DUTCH	1.93
ADRIANO	ITALIAN	1.30
ADRIANUS	DUTCH	1.41
ADRIE	WELSH	0.80
ADRIS	PAKISTANI	1.63
ADUA	ITALIAN	0.78
ADUKE	NIGERIAN	0.94
ADUL	PAKISTANI	0.85
ADUNNI	NIGERIAN	1.54
ADUNOLA	NIGERIAN	2.08
ADWOA	GHANAIAN	1.73
AEDAN	IRISH	2.33
AEJAZ	PAKISTANI	2.68
AESHA	PAKISTANI	2.49
AEYSHA	PAKISTANI	0.67
AFAF	MUSLIM_MIDDLE EAST	2.51
AFAM	NIGERIAN	1.58
AFAN	WELSH	3.66
AFAQ	PAKISTANI	2.33

Appendix 6: NHS Research Governance and Ethical Approval Documents

Letters from NHS North Central London Research Consortium confirming research governance and ethnical approval

NHS

North Central London Research Consortium

Research Operations Unit
3rd Floor West Wing
St Pancras Hospital
London
NW1 0PE

February 16th 2006

Mr Pablo Mateos
R&D Associates – Public Health Intelligence Team
Camden Primary Care Trust
Room 24, 4th Floor – West Wing
St Pancras Hospital
London
NWI 0PE

Dear Mr Mateos

REC Ref: 05/Q0511/130
Title: Knowledge Transfer Partnership - Geodemographics in Public Health

I am pleased to note that the Local Research Ethics Committee has informed us that there are no ethical reasons why your study should not proceed.

Projects are registered with the North Central London Research Consortium if they utilise patients, staff, records, facilities or other resources of Camden Primary Care Trust, Islington Primary Care Trust, the Camden & Islington Mental Health and Social Care Trust, Barnet Primary Care Trust, Enfield Primary Care Trust or Haringey Teaching Primary Care Trust.

The Camden Primary Care Trust therefore grants approval to begin research based on the proposal reviewed by the ethics committee and subject to any conditions set out in their letter of 29th December 2005. Should you fail to adhere to these conditions or deviate from the protocol reviewed by the ethics committee, then this approval would become void. The approval is also subject to your consent for information to be extracted from your project registration form for inclusion in NHS project registration/management databases and, where appropriate, the National Research Register and the UCL Clinical Research Network register.

Permission to conduct research is also conditional on the research being conducted in accordance with the Department of Health Research Governance Framework for Health and Social Care:

- Appendix A to this letter outlines responsibilities of principal investigators;

- Appendix B details the research governance responsibilities for other researchers. It also outlines the duties of all researchers under the Health and Safety at Work Act 1974. Principal investigators should disseminate the contents of Appendix B to all those in their research teams.

Further information on the research governance framework for health and social care can be found on the DH web pages at http://www.doh.gov.uk/research/
Staff working within trusts covered by the research consortium can also find the information on the Trust Intranet.

Researchers are also reminded that personally identifiable information on living persons must be collected, stored, processed and disclosed in accordance with the Data Protection Act 1998. Such data may be in the form of electronic files, paper files, voice recordings or photographs/scans/X-rays. Further information on the Data Protection Act is available from your organisations Data Protection Officer or from the Consortium R&D Unit. The Medical Research Council also publishes the guidance booklet 'Personal Information in Medical Research' which is available from http://www.mrc.ac.uk/pdf-pimr.pdf

Except in the case of commercially funded research projects, the following acknowledgement and disclaimer MUST appear on all publications arising from your work.
*"This work was undertaken with the support of [***Insert Trust***] Trust, who received [***insert "funding" or a "proportion of funding" ***]* from the NHS Executive; the views expressed in this publication are those of the authors and not necessarily those of the NHS Executive".*

** "a proportion of funding" where the research is also supported by an external funding body; "funding" where no external funding has been obtained.*

This is a requirement of the contract between the Trust and the NHS Executive in which the Trust receives funding to cover the infrastructure costs associated with performing non-commercial research.

Please make all members of the research team aware of the contents of this approval. I wish you every success with your research.

Yours sincerely,

Research Governance Manager

COPY NHS

Camden & Islington Community Local Research Ethics Committee
Room 3/14
Third Floor, West Wing
St Pancras Hospital
4 St Pancras Way
London
NW1 0PE

29 December 2005

Professor Paul Longley
Professor (Chair) of Geographic Information Science
University College London
Centre for Advanced Spatial Analysis (CASA)
1-19 Torrington Place
London
WC1E 6BT

Dear Professor Longley

Full title of study: **Knowledge Transfer Partnership- The application of Geodemographics and Geographic Information systems in Public Health**
REC reference number: **05/Q0511/130**

The Research Ethics Committee reviewed the above application at the meeting held on 19 December 2005.

Ethical opinion

The Committee considered this project an interesting and informative one with concise aims and objectives.

The members of the Committee present gave a favourable ethical opinion of the above research on the basis described in the application form, protocol and supporting documentation.

Conditions of approval

It must be stressed that the Committee was disappointed to learn that this project had already commenced prior to seeking and gaining ethical approval from the Local Ethics Committee. Please note for future applications that the ethical approval process is a very stringent directive that cannot be compromised.

At the meeting of the 19 December 2005 the researchers' representative (Principal Investigator; Pablo Mateos) was informed by the Chair to initiate the immediate termination of any further work on this project until full ethical approval had been received.

The Committee noted that the main ethical concern regarding this study is the fact that the 'participants' data sets are to be accessed and reviewed without seeking the participants consent. Such a dilemma is however typical of this type of research.

In regards to the discussion about application to the Patient Information Advisory Group (PIAG) this has since found to be an unnecessary action. The Committee previously considered it may be appropriate for this study to seek approval from PIAG but further review of the honorary employment contracts between the researchers and the PCT have clarified that this will not be necessary.

The favourable opinion is given provided that you comply with the conditions set out in the enclosed document. You are advised to study the conditions carefully.

Approved documents

The documents reviewed and approved at the meeting were:

Document	*Version*	*Date*
Application	1	30 November 2005
Investigator CV	Paul Longley	01 December 2005
Protocol	1	
Compensation Arrangements	Knowledge Transfer Partnerships	30 January 2004
Supervisors CV	Richard Webber	01 December 2005

Research governance approval

The study should not commence at any NHS site until the local Principal Investigator has obtained final research governance approval from the R&D Department for the relevant NHS care organisation.

Membership of the Committee

The members of the Ethics Committee who were present at the meeting are listed on the attached sheet.

Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees (July 2001) and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

05/Q0511/130	**Please quote this number on all correspondence**

With the Committee's best wishes for the success of this project

Yours sincerely

Enclosures: *List of names and professions of members who were present at the meeting and those who submitted written comments*

Standard approval conditions

Copy to:

University College London (UCL)
Knowledge Transfer Partnership Office
Gower St,
London
WC1E 6BT

Pablo Mateos, Principal Investigator

R&D Department for NHS care organisation at lead site

Islington NHS
Primary Care Trust

[redacted]
Information Analyst
1st Floor, Dpt Public Health
Islington PCT
338-346 Goswell Road
London EC1V 7LQ

[redacted]

28/11/2005

Pablo Mateos
R&D Associate -Public Health Intelligence Team
Directorate of Public Health, Camden PCT
Room 24, 4th Floor - West Wing
St Pancras Hospital, 4 St Pancras Way
London NW1 0PE

Re: Caldicott Request Approval

Dear Pablo,
I am writing on behalf of Dr. [redacted] Islington PCT's Caldicott guardian, to confirm that she has approved your request dated 9/9/2005 regarding using Islington's data to improve ethnicity coding using patient's names. She has signed your form on 22/11/2005.

Kind Regards,

[redacted]

Appendix 7: Peer Reviewed Publication

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